PREMIER REFERENCE SOURCE

WEB TECHNOLOGIES

Concepts, Methodologies, Tools, and Applications

ARTHUR TATNALL

VOLUME I

Web Technologies:
Concepts, Methodologies, Tools, and Applications

Arthur Tatnall
Victoria University, Australia

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content:	Kristin Klinger
Development Editor	Julia Mosemann
Senior Managing Editor:	Jamie Snavely
Managing Editor:	Michael Brehm
Assistant Managing Editor:	Carole Coulson
Typesetters:	Michael Brehm, Carole Coulson, Elizabeth Duke, Christopher Hrobak, Jamie Snavely, Sean Woznicki
Cover Design:	Lisa Tosheff
Printed at:	Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Web technologies : concepts, methodologies, tools and applications / Arthur Tatnall, editor.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-60566-982-3 (hbk.) -- ISBN 978-1-60566-983-0 (ebook) 1. World Wide Web. 2. Internet. 3. Information technology. I. Tatnall, Arthur.
TK5105.888.W377 2010
004.6'7--dc22
2009037778

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

List of Contributors

Contents

Volume I

Section I. Fundamental Concepts and Theories

This section serves as the foundation for this exhaustive reference tool by addressing crucial theories essential to the understanding of Web technologies. Chapters found within these pages provide an excellent framework in which to position Web technologies within the field of information science and technology. Individual contributions provide overviews of the mobile Web, semantic Web, and Web 2.0, while also exploring critical stumbling blocks of this field. Within this introductory section, the reader can learn and choose from a compendium of expert research on the elemental theories underscoring the research and application of Web technologies.

Section II. Development and Design Methodologies

This section provides in-depth coverage of conceptual architectures, frameworks and methodologies related to the design and implementation of Web technologies. Throughout these contributions, research fundamentals in the discipline are presented and discussed. From broad examinations to specific discussions on particular frameworks and infrastructures, the research found within this section spans the discipline while also offering detailed, specific discussions. Basic designs, as well as abstract developments, are explained within these chapters, and frameworks for designing successful Web sites, Web-based applications, and Web portals are provided.

Volume II

Section III. Tools and Technologies

This section presents extensive coverage of the technology that informs and impacts Web technologies. These chapters provide an in-depth analysis of the use and development of innumerable devices and tools, while also providing insight into new and upcoming technologies, theories, and instruments that will soon be commonplace. Within these rigorously researched chapters, readers are presented with examples of the tools that facilitate and support the emergence and advancement of Web technologies. In addition, the successful implementation and resulting impact of these various tools and technologies are discussed within this collection of chapters.

Section IV. Utilization and Application

This section introduces and discusses the utilization and application of Web technologies. These particular selections highlight, among other topics, the application of semantic Web technologies to e-tourism, e-banking, and in car repairs as well as the adoption of Web services in digital libraries. Contributions included in this section provide excellent coverage of today's online environment and insight into how Web technologies impact the fabric of our present-day global village.

Section V. Organizational and Social Implications

This section includes a wide range of research pertaining to the social and organizational impact of Web technologies around the world. Chapters included in this section analyze social marketing, e-government, Web vendors, and Web tourism. The inquiries and methods presented in this section offer insight into the implications of Web technologies at both a personal and organizational level, while also emphasizing potential areas of study within the discipline.

Volume III

Section VI. Managerial Impact

This section presents contemporary coverage of the managerial implications of Web technologies. Particular contributions address Web software engineering and Web-enabled employee life-cycle process management. The managerial research provided in this section allows executives, practitioners, and researchers to gain a better sense of how Web technologies can inform their practices and behavior.

Section VII. Critical Issues

This section addresses conceptual and theoretical issues related to the field of Web technologies, which include issues related to usage, as well as failures and successes in Web implementation. Within these chapters, the reader is presented with analysis of the most current and relevant conceptual inquires within this growing field of study. Particular chapters address privacy concerns in Web logging, Web information extraction, and Web rules. Overall, contributions within this section ask unique, often theoretical questions related to the study of Web technologies and, more often than not, conclude that solutions are both numerous and contradictory.

Volume IV

Section VIII. Emerging Trends

This section highlights research potential within the field of Web technologies while exploring uncharted areas of study for the advancement of the discipline. Chapters within this section highlight emerging semantic Web applications, Web personalization, and learning on the Web. These contributions, which conclude this exhaustive, multi-volume set, provide emerging trends and suggestions for future research within this rapidly expanding discipline.

Preface

Since its development just two decades ago, the World Wide Web has grown to become the infrastructure that supports innumerable applications essential to everyday life. It's not an exaggeration to claim that if you can think it, you can create a Web page about it. We use Web sites and the information they contain to create and connect with a seemingly unlimited amount of information. As such, it is important to understand the tools and technologies that support the continued growth of the Web and contribute to its role as an increasingly-pervasive aspect of our lives.

With the constant changes in the landscape of Web technologies, it is a challenge for researchers and experts to take in the volume of innovative advances and up-to-the-moment research in this diverse field. Information Science Reference is pleased to offer a four-volume reference collection on this rapidly growing discipline, in order to empower students, researchers, academicians, and practitioners with a wide-ranging understanding of the most critical areas within this field of study. This collection provides the most comprehensive, in-depth, and recent coverage of all issues related to the development of cutting-edge Web technologies, as well as a single reference source on all conceptual, methodological, technical and managerial issues, and the opportunities, future challenges and emerging trends related to the development, application, and implications of Web technologies.

This collection entitled, **"Web Technologies: Concepts, Methodologies, Tools, and Applications"** is organized in eight (8) distinct sections, providing the most wide-ranging coverage of topics such as: 1) Fundamental Concepts and Theories; 2) Development and Design Methodologies; 3) Tools and Technologies; 4) Utilization and Application; 5) Organizational and Social Implications; 6) Managerial Impact; 7) Critical Issues; and 8) Emerging Trends. The following provides a summary of what is covered in each section of this multi-volume reference collection:

Section 1, ***Fundamental Concepts and Theories***, serves as a foundation for this extensive reference tool by addressing crucial theories essential to the understanding of Web technologies. Chapters such as "Tips for Tracking Web Information Seeking Behavior" by Brian Detlor, Maureen Hupfer, and Umar Ruhi and "A Proposed Template for the Evaluation of Web Design Strategies" by Dimitrios Xanthidis, David Nicholas, and Paris Argyrides provide analyses of user behavior and Web design. "Mobile Social Web; Opportunities and Drawbacks," by Thorsten Caus, Stefan Christmann, and Svenja Hagenhoff presents an overview of recent trends in mobile Web usage, which is becoming an increasingly important area of study as more and more people obtain Internet access for their wireless devices. Later selections, such as "Web 2.0 and E-Discovery" by Bryan Kimes and "The Power and Promise of Web 2.0 Tools" by G. Andrew Page and Radwan Ali explore the application of Web 2.0 as well as the issues companies must address as a result. These and several other foundational chapters provide a wealth of expert research on the elemental concepts and ideas which surround Web design and access.

Section 2, ***Development and Design Methodologies***, presents in-depth coverage of the conceptual design and architecture of Web sites, services, and systems. "Paralingual Web Design and Trust in E-

Government," by Roy H. Segovia, Murray E. Jennex, and James Beatty and "Designing Medical Research Web Sites" by Jonathan Grady, Michael B. Spring, and Armando J. Rotondi discuss context-specific Web design projects, highlighting the importance of recognizing the specific needs and requirements of different development initiatives. The latter half of this section introduces concepts that relate to the development of Semantic Web services. Chapters such as "A Semantic Web-Based Approach for Building Personalized News Services" by Flavius Frasincar, Jethro Borsje, and Leonard Levering and "Building Semantic Web Portals with a Model-Driven Design Approach" by Marco Brambilla and Federico M. Facca offer specific considerations for the creation of Semantic Web services, while later selections such as "Rule Markup Languages and Semantic Web Rule Languages" by Adrian Paschke and Harold Boley and "Semantic Web Rule Languages for Geospatial Ontologies" by Philip D. Smart, Alia I. Abdelmoty, Baher A. El-Geresy, and Christopher B. Jones present more technical considerations relating to the use and communication of rule languages in the Semantic Web. With 20 contributions from leading international researchers, this section offers copious developmental approaches and methodologies for Web services and technologies.

Section 3, ***Tools and Technologies***, presents extensive coverage of the various tools and technologies used in the development and implementation of Web services and applications. This comprehensive section opens with the chapters "New Paradigms: A Collaborative Web Based Research Tool," by Hamish Holewa, and "Adaptability and Adaptivity in The Generation of Web Applications," by Raoudha Ben Djemaa, Ikram Amous, and Abdelmajid Ben Hamadou, which describe new tools that support the development of Web applications and the challenges faced in the management and creation of new technology. "Migrating Web Services in Mobile and Wireless Environments," by Myung-Woo Park, Yeon-Seok Kim, and Kyong-Ho Lee revisits Web use on wireless devices, specifically exploring the mitigation and replication of Web services among mobile devices. Later selections such as "Web 2.0 Technologies: Social Software Applied to Higher Education and Adult Learning" by Teresa Torres-Coronas, M. Arántzazu Vidal-Blasco, Ricard Monclús-Guitart, M. José Simón-Olmos, and Araceli Rodríguez-Merayo and "Interactive Whiteboards in the Web 2.0 Classroom" by David Miller and Derek Glover provide insight into the use of specific Web tools (namely social software and interactive whiteboards) in educational settings. In all, this section provides coverage of a variety of Web tools and technologies under development and in use.

Section 4, ***Utilization and Application***, describes the implementation and use of an assortment of Web technologies. Including chapters such as "Semantic Web Take-Off in a European Industry Perspective" by Alain Léger, Jean Charlet, Johannes Heinecke, Paola Hobson, Lyndon J.B. Nixon, François Goasdoué, and Pavel Shvaiko and "Semantic Web for Media Convergence: A Newspaper Case" by Ferran Perdrix, Juan Manuel Gimeno, Rosa Gil, Marta Oliva, and Roberto García provide specific insight into the application of Web tools and technologies in both the professional and private sector. "Mailing Lists and Social Semantic Web" by Sergio Fernández, Jose E. Labra, Diego Berrueta, Patricia Ordóñez de Pablos, and Lian Shi describes the use of mailing lists and presents a method for extracting data from these lists. Later selections, such as "A Context-Based Approach to Web 2.0 and Language Education" by Gary Motteram and Susan Brown and "Exploring the Effects of Web-Enabled Self-Regulated Learning and Online Class Frequency on Students' Computing Skills in Blended Learning Courses" by Pei-Di Shen and Chia-Wen Tsai suggest approaches and consider the impact of Web-based learning on student performance. Contributions found in this section provide comprehensive coverage of the practicality and current use of Web technologies.

Section 5, ***Organizational and Social Implications***, includes chapters discussing the impact of Web technology on social and organizational practices. Chapters such as "Building Trust in E-Commerce through Web Interface," by Muneesh Kumar and Mamta Sareen and and "Swift Trust in Web Vendors:

The Role of Appearance and Functionality," by Xin Li, Guang Rong, and Jason B. Thatcher discuss the growth and influence of e-commerce and the important role trust plays in impacting e-marketplaces. Specific Web implementation and resulting implications of such initiatives are explored in selections such as "Assessing the Performance of Airline Web Sites: The ARTFLY Case" by Elad Harison and Albert Boonstra and "Aviation-Related Expertise and Usability: Implications for the Design of an FAA E-Government Web Site" by Ferne Friedman-Berg, Kenneth Allendoerfer, and Shantanu Pai. This section continues with discussions of Web accessibility and customization, concluding with a discussion of educational implications of Web technology. Overall, these chapters present a detailed investigation of how Web technology is implemented and how this implementation impacts the individual and society as a whole.

Section 6, ***Managerial Impact***, presents focused coverage of Web services and technology as it relates to improvements and considerations in the workplace. "Employee Life-Cycle Process Management Improvement with Web-Enabled Workflow Systems" by Leon Welicki, Javier Piqueres Juan, Fernando Llorente Martin, and Victor de Vega Hernandez presents a real-world case of constructing a Web-enabled worklflow for managing employee-life cycle processes, which include hiring and dismissing of employees. "Web Engineering in Small Jordanian Web Development Firms: An XP Based Process Model" by Haroon Altarawneh and Asim El-Shiekh describes a model for small Web project development and explains, from a managerial perspective, how this differs from the more large-scale implementation projects adopted by larger firms. In all, the chapters in this section offer specific perspectives on how work and Web technologies interact and inform each other to create more meaningful user experiences.

Section 7, ***Critical Issues***, addresses vital issues related to Web technology, which include privacy and quality, among other topics. Chapters such as "Privacy Concerns for Web Logging Data" by Kirstie Hawkey explore the issues that must be considered when collecting user data and offer recommendations for enhancing privacy. Later selections, such as "Search Engine-Based Web Information Extraction" by Gijs Geleijnse and Jan Korst, continue the discussion of information gathering and extraction which, in this chapter, is discussed in terms of approaches to expressing and sharing structured information in Semantic Web languages. This section continues by asking unique questions about information literacy, as well as presenting new solutions to questions about the social Web and Web services profiling.

The concluding section of this authoritative reference tool, ***Emerging Trends***, highlights areas for future research within the field of Web technology, while exploring new avenues for the advancement of the discipline. Beginning this section is "The Social Semantic Desktop: A New Paradigm Towards Deploying the Semantic Web on the Desktop" by Ansgar Bernardi, Mehdi Jazayeri, Stefan Decker, Cédric Mesnage, Ludger van Elst, Knud Möller, Gunnar Aastrand Grimnes, Michael Sintek, Tudor Groza, Leo Sauermann, and Siegfried Handschuh. This selection presents the Social Semantic Desktop project, addressing design considerations of a project whose aim is to blur the lines between individual applications and users' physical workspace. Trends in marketing are explored in "Social Media Marketing; Web X.0 of Opportunities" by Lemi Baruh with the aim of introducing new techniques for advertisers whose aim is to reach consumers through social media. These and several other emerging trends and suggestions for future research can be found within the final section of this exhaustive multi-volume set.

Although the primary organization of the contents in this multi-volume work is based on its eight sections, offering a progression of coverage of the important concepts, methodologies, technologies, applications, social issues, and emerging trends, the reader can also identify specific contents by utilizing the extensive indexing system listed at the end of each volume. Furthermore to ensure that the scholar, researcher and educator have access to the entire contents of this multi volume set as well as additional coverage that could not be included in the print version of this publication, the publisher will provide unlimited multi-user electronic access to the online aggregated database of this collection for the life of the edition, free of charge when a library purchases a print copy. This aggregated database provides far more contents than what can be included in the print version in addition to continual updates. This unlimited access, coupled with the continuous updates to the database ensures that the most current research is accessible to knowledge seekers.

The diverse and comprehensive coverage of Web technologies presented in this four-volume authoritative publication will contribute to a better understanding of all topics, research, and discoveries in this developing, significant field of study. Furthermore, the contributions included in this multi-volume collection series will be instrumental in the expansion of the body of knowledge in this enormous field, resulting in a greater understanding of the fundamental concepts and technologies while fueling the research initiatives in emerging fields. We at Information Science Reference, along with the editor of this collection, hope that this multi-volume collection will become instrumental in the expansion of the discipline and will promote the continued growth of all aspects of Web technology.

Web Technologies:
Concepts, Applications, Trends and Research Issues

Arthur Tatnall
Victoria University, Australia

INTRODUCTION

Web-based systems and technologies are now used for a vast number of applications, and this chapter aims to provide an overview of the technologies themselves and also the uses to which the Web is now put, as well as the social and political impact of this use. A number of important concepts underlie the Web as well as a good deal of jargon, and some of the main concepts and terms are explained here. Design and development of web-based systems is an important topic and this is briefly discussed along with some of the tools and issues involved in this development. It is impossible to do justice to the huge range of applications of the Web in an incredibly diverse range of areas, but this chapter attempts to do just this by examining some of the most important applications. Not all aspects of the Web can be considered to be either worthwhile or healthy for society as a whole and issues like identify theft and the distribution of pornography, both of which have been made easier by access to the Web, are also discussed. No technological innovation can be useful, however, until it has been adopted and the factors leading to adoption of some Web-based systems and not others, and to adoption of some aspects of these systems by one organisation and other aspects by another organisation are also discussed. Finally this chapter also tackles the issue of researching the Web; what this involves and what sort of methodologies might be appropriate in this socio-technical area.

THE WEB: TECHNOLOGY, APPLICATIONS AND PEOPLE

Today everyone knows of the World Wide Web and very many people around the world make daily use of its facilities. It is hard to imagine what it must have been like before the Web became such an important part of our lives, but it was only in 1989 that Tim Berners-Lee, based at the European Laboratory for Particle Physics (CERN[1]), in looking for a solution to the management and sharing of the large amounts of scientific information his colleagues created, wrote a proposal for a large online hypertext database that by 1991 had become what we now call the World Wide Web (Lawrence, Newton, Corbitt, Braithwaite, & Parker, 2002; Sklar, 2009). Thus the Web began as a means of improving information sharing and document handling between the research scientists at CERN and throughout the world. It was designed to allow pages containing hypertext to be stored in a way that allowed other computers access to these pages. It was probably not until about the mid 1990s, however, that the Web began to really gain in popularity. At that time, few could have foreseen the multitude of uses it would be put to by 2010, and the number of people who would make use of it. It is no exaggeration to say that the Web has now become quite ubiquitous.

The Internet has been around much longer of course, tracing its ancestry back to the ARPANET (Advanced Research Projects Agency Network), developed to link US Defense Department researchers with those in several universities in the USA. It became operational in late 1969 (Lawrence et al., 2002; Davison, Burgess, & Tatnall, 2008). The first appearance of the term ‘Internet’ was in 1974 as an abbreviation for ‘Internetworking’ (Network Working Group, 1974), and things developed from there, with electronic mail soon becoming an important form of communication within the research community that used this technology. At this time, however, making use of the Internet was not something that the average person or business could easily do or find much value in, and it was not until the later advent of the Web that use of the Internet became common, and a general topic of conversation in many communities.

No one knows exactly the size of the Internet, but an article in New Scientist (Barras, 2009) notes that back in 2005, Google estimated that the Internet contained 5 million terabytes of data. In July 2008 when the new search engine Cuil.com commenced operation claiming to be the world’s largest search engine, Google announced that it had registered a trillion unique pages, but in reality the Internet is probably even bigger than this as some estimates suggest that the pages indexed by Google and Cuil may represent only a hundredth of the information on the Internet (Barras, 2009). This same article suggests that 210 billion e-mails were sent every day in 2008.

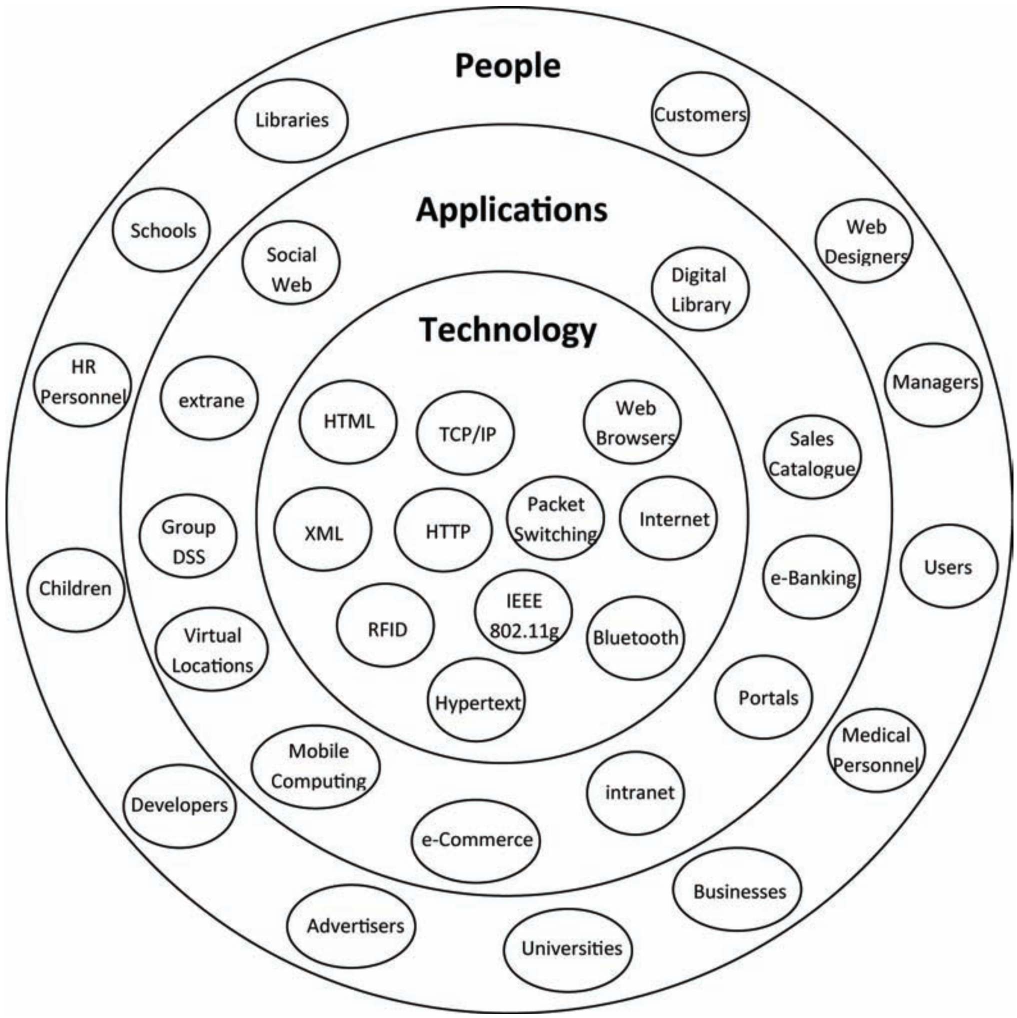
Today a great deal has been written about the Web, its concepts, its technologies, its design tools, its applications and the social and political effects that have gone with its growth. The Web is not, of course, just technology and the socio-technical nature of Web systems is an important consideration. For the purposes of this chapter, I will consider the study of Web-based systems and Web technology as consisting of three areas: technology, applications and people. Figure 1 (below) shows these as three concentric circles (with technology in the middle). Each of these circles then contains a number of entities related to that circle. (Space does not permit all relevant entities to be shown here.)

This introductory chapter will present an overview of these topics. Of necessity it cannot cover every topic fully and must be somewhat selective in its coverage with some topics only touched on and others omitted completely.

FUNDAMENTAL CONCEPTS

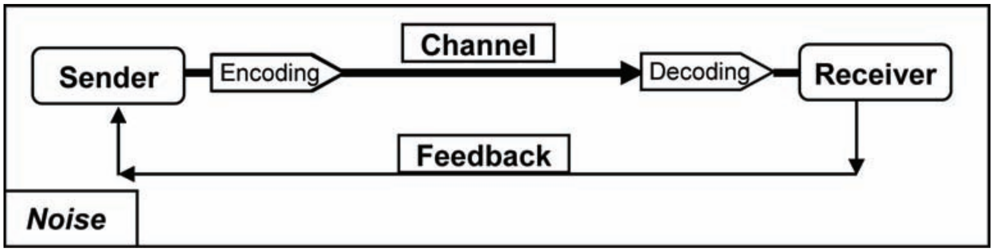
Both the Internet and the World Wide Web come with a large amount of jargon. Some understanding of this jargon and also of the concepts underlying web technology is necessary to fully appreciate the complexity and power of the Web. For those readers who may be new to some of these ideas, some of the main terms and concepts, along with a brief description of their meaning and use, will now be presented.

Figure 1. Components of web-based systems



- **Hypertext** allows words (or other objects) in one document to be linked to other documents. It provides a dynamic means of organising and accessing information where pages of information are connected together by hypertext links. A **Hyperlink** (Hypertext Link) can be text, or a picture that is associated with the location (path and filename) of another document and used to link to this other document. These documents are called **web pages** and can contain text, graphics, video and audio as well as hyperlinks. A **website** is a collection of related pages stored on a **web server**. A web server is made up of computer hardware and software, based on a PC or other larger computer.
- A **URL** (Uniform Resource Locator) gives the address or location of any specific website. Each URL defines the path that will transmit the document, the Internet protocol being used, and the server on which the website is located. Each Internet address is translated into a series of numbers called an **IP address**. A **domain name** is used by an organisational entity to identify its website and is based on the **Domain Name System** (**DNS**) hierarchy.

Figure 2. A Communication Model



- A **web browser** is a software application used to access and view web pages. Web browsers include: Internet Explorer, Firefox, Safari, Opera and Netscape Navigator.
- **Communications Model:** Communication involves a sender dispatching some form of message to a receiver. It occurs within a particular context or setting, and involves the transfer of some form of information from the sender to the receiver over some type of communications channel (Davison et al., 2008). The message is coded into an appropriate form by the sender before transmission, and later decoded by the receiver. Feedback lets the sender know how the message was received. Noise, or interference, can upset transmission and lead to the message received differing from the one that was sent.
- **Packet switching**. Data to be transmitted is broken into discrete packets consisting of groups of characters, which are then sent independently through whichever path between sender and receiver is most convenient at the time. The connection is virtual and so the data may follow different paths. Each packet is labelled electronically with codes to indicate its origin and destination, and may follow a different path through the network to that of other packets. When they reach their destination the packets are re-assembled to produce the original message.
- **Internet connection protocols**: instructions for connecting a computer to the Internet. Two important protocols are PPP (Point to Point) and SLIP (Serial Line Internet Protocol).
- **HTTP** (HyperText Transfer Protocol) is the set of rules that locate and move files of all types, including multimedia files, across the Internet. It does not process the packages of data it transmits but simply ensures they reach their destination and so defines computer links to the Web. **HTTPS** (HyperText Transfer Protocol (Secure)) indicates that HTTP is to be used, but with a different default port and an additional encryption/ authentication layer between HTTP and TCP.
- **Open Systems Interconnection (OSI)**. As networks operate in somewhat different ways, there was a need to define standards for data communication to provide a means for computers from different manufacturers to communicate with each other. Data sent over a network must arrive at its destination in a timely, correct, and recognisable format and to facilitate this, the OSI model consists of seven layers, each of which is selected to perform a well-defined function. While OSI is an international standard it is really more important as a concept and TCP/IP is the standard that is actually used.
- **TCP/IP** (Transmission Control Protocol, Internet Protocol). These are two protocols that support the basic operation of the Internet and include rules that computers on a network use to establish and break connections. TCP/IP controls the assembly of a message into small packets before transmission, controls the reassembly of packets once they reach their destination, has rules for routing individual data packages from their source to their destination and allows users to send and receive messages, find information, exchange data and download software.

- **HTML** (HyperText Markup Language) is used to describe how a web browser should display a text file retrieved from a web server. HTML defines how a page is displayed and is a derivative of **SGML** – Standard Generalised Markup Language which is a standard system (an ISO standard) used to specify document structure. HTML allows the developer of a web page to define hyperlinks between that document and any others that are required.
- **XML** (Extendible Markup Language) defines the page's actual content and also interprets the meaning of the data**.** It defines which data are displayed, whereas HTML only defines how a page is displayed. Elements of structured information include: content (pictures, graphics, text etc) and the role the content plays in the document; where the content is located in the document influences the meaning given to it.
- **XHTML** (Extensible Hypertext Markup Language) is a newer version of HTML, based on XML and designed to support newer devices.
- **Java** is a programming language used to create many web-based applications. Much of Java's syntax derived from C++ and its compiled version can run on any computer architecture in a Java virtual machine.
- An **intranet** is a private (internal) network accessible only by selected individuals within an organisation. It is based on web technology, but access is restricted so as to exclude the outside world. An intranet uses web browsers and hypertext links in the same way as the World Wide Web, the only real differences being where the web pages are located, and who can access them. (Applications of intranets are discussed further in a later section.)
- An **extranet** is a private network (based on web technologies) that links selected parts of a company's intranet with its customers, suppliers, or other authorised business partners. (Applications of extranets are discussed in a later section.)
- **EDI** (Electronic Document Exchange) is the exchange of data or 'business documents' directly between computer systems in business partner companies using a standardised data format. It has been used for more than 30 years predominantly for purchasing and inventory. This direct computer-to-computer form of EDI has, to a large extent, now been superseded by web-based purchasing systems.
- **RFID** (Radio Frequency Identification) is a technique for storing and retrieving data remotely using devices called RFID tags that can be attached to, or inserted into a product. An RFID tag incorporates a silicon chip and antennae that can broadcast a unique identification code when prompted by a reader device.
- **Bluetooth** is an open standard for short-range wireless communication (up to 10 metres) between digital devices. It comprises hardware, software and interoperability standards, protocols and requirements and is implemented through a small, low-cost radio chip that can be incorporated in mobile phones, PDAs, pocket PCs, computers, printers and similar devices. (It was initially developed by Ericsson and is named after Harald Bluetooth, a Viking and the King of Denmark between 940 and 981.)
- **The IEEE 802.11 Protocols**. The Standards Association of the Institute of Electrical and Electronic Engineers (IEEE) has developed a series of standards for wireless local area network (WLAN) devices operating in close proximity (up to 100m) under the generic title **IEEE 802.11**. There are a number of variants of this protocol, such as **IEEE 802.11a**, **IEEE 802.11b** and **IEEE 802.11g** (Wave Repo2001; O'Hara and & Petrick, 2005).
- **PDA** (Personal Digital Assistant). Sometimes now called a **Smart Phone**, these devices offer many of the communication and information characteristics of a laptop computer and mobile phone incorporated into a single device. Often a PDA includes a mobile (cell) phone along with Windows

Mobile operating system with mobile versions of Word, Excel and Outlook, Bluetooth and IEEE 802.11 connectivity and Internet access.

- **Personal Area Networks** (PAN) are ad hoc networks of personal digital devices such as laptop computers, personal digital assistants (PDA) and mobile phones able to transfer data and linked to other personal digital devices by Bluetooth, or some other wireless technology.
- The **Semantic Web** offers a common framework for data to be shared and reused across application, enterprise, and community boundaries. The World Wide Web Consortium (W3C), who are developing the Semantic Web, note that it is concerned with common formats for the integration and combination of data from diverse sources and the language needed to show how this data relates to the real world (http://www.w3.org/2001/sw/). It is an extension of Web in which the semantics of information and services on the web are defined. This then makes it possible to understand requests of people and machines who want to use web content.
- **Web 2.0** websites build on the interactive facilities of the Web and so allow users to do more than just retrieve information. This has led to the development of web-based communities and social-networking sites.
- **Internet service provider** (ISP). This is a company that offers its customers access to the Internet
- **Web hosting service provider**. These organisations provide space on their server for an organisation or individual's web page. They also offer connectivity of their servers to the Internet.
- **Social networking** sites (such as Facebook, MySpace, Twitter and LinkedIn) aim to build online communities of people who share interests. (Social networking is discussed more fully in a later section.)
- A **Wiki** is a website that allows the easy creation and editing of a number of interlinked collaborative pages. The Wikipedia encyclopedia is one of the best-known of these wikis (http://en.wikipedia.org/wiki/Wiki).
- A **Blog** (short for weblog) is a website set up by an individual to describe events and materials of interest to them and relevant to their life.

DEVELOPMENT AND DESIGN METHODOLOGIES AND TOOLS

Web development and design has much in common with the design of other information systems and a starting point should be in process of analysing and creating information systems (Tatnall, Davey, Burgess, Davison & Wenn, 2002). Creating an information system and getting it working can be considered to comprise three separate aspects:

- **Systems analysis:** the description of a logical model of the present or proposed system in non-technical terms, independent of the actual physical implementation.
- **Systems design:** the business of converting this logical model into a useable system.
- **Systems implementation:** involves the construction of the physical system to the level of hardware acquisition, programming and change management.

Web Development Issues

As many books describe the development of information systems I will say no more about this here. Web development, apart from following the logical process of information systems development, involves a number of other special considerations and I will briefly discuss some of these now (Sklar, 2009).

- Connection speed considerations. Will all your users have high speed broadband connections? If some are likely to be still using modems then the amount of data to be downloaded needs to be limited or else it will take too long to load and the potential user will give up.
- Different screen resolutions. Not all users will have the latest high resolution monitors and there perhaps need to be different version of the site for different monitor resolutions.
- Different operating systems (– Windows, Mac, Linux etc) and different web browsers may make your site appear differently. This needs to be taken into consideration and portability of design is a worthwhile goal.
- Design with the user in mind. This seems obvious, but is not always done. Consider the path taken by the eye in looking around the web page. Consider use of language appropriate to the intended user. Apart from whether the language is English, French, Spanish, German or Chinese, there is the issue of complexity – will the intended user understand? Another issue here is ensuring that users can find their way around inside the website hierarchy. The basic point here is that the website must be easily usable.
- As different cultures around the world see and understand things somewhat differently, creating different pages for different countries, as do many commercial websites, is worthwhile both from the language to be used and because different countries many need slightly different content.
- Accessibility is important. Consideration should be given to people with disabilities such as poor eyesight, colour blindness, poor hearing etc.
- Programming issues have much in common with those in information systems, except that the appropriate languages are often different.
- File and data structures – again, approaching this issue has much in common with similar issues in building an information system.
- Building a website needs graphical design capabilities as well as abilities in information systems design.
- Security. It is important that appropriate security is built into any website, relevant to the use that will be made of this site.
- Privacy of users of the website should be respected by not collecting unnecessary personal data and being careful not to pass on any data that is collected.
- Trust will be built up with website users if the site shows that it deserves such trust.
-

Web Development Tools

A multitude of different tools and commercial products exist to help with website creation which can be considered at two levels: design of the underlying code, data and file structures and design of the appearance of the screen itself. One level needs an information systems professional, the other someone with graphic design training and abilities. To create the code there are programming languages such as Java, C++ and Visual Basic and to create or edit graphics there are packages such as Photoshop or Dreamweaver.

One web development product is Microsoft Visual Studio with Visual Web Developer which includes a number of relevant different programming languages such as C++ and Visual Basic, and other tools useful for website creation. At another level, Microsoft Office Front Page and Expression Web allow website creation without the need for coding.

WEB-BASED APPLICATIONS

Since the start of its rapid growth in popularity in the mid 1990s, a huge number of web-based applications have been developed. It is, of course, impossible to describe even a small fraction of these but I will discuss just a few of the most interesting in this section.

E-Business (E-Commerce)

E-Business or electronic business (sometimes called e-Commerce) can be defined as the use of electronic means on the Web to conduct the business of buying and selling using websites with shopping carts and e-mail. This can involve many facets of a firm's operation being conducted electronically between the business and its partners, suppliers and customers (Davison et al., 2008). This is a very broad topic containing many sub-areas and represents a major use of the Web.

E-Commerce Business Models

There are several commonly described models that relate the electronic commerce aspects of a business to other entities of importance:

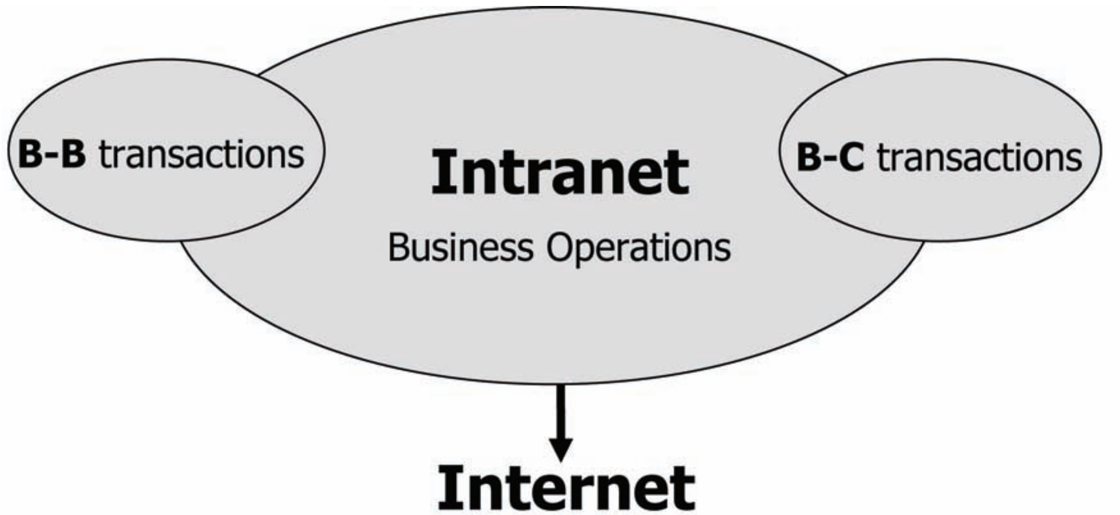
- Business-to-Business (B-B), for transactions conducted between businesses.
- Business-to-Consumer (B-C), for consumer shopping on the Web.
- Business-to-Government (B-G) is one of several other models used to connect to other entities.
- Business-to-Employee (B-E). Business operations within an organisation (such as B-E) often make use of the organisation's intranet.

E-Marketplace

A traditional approach to advertising using well known advertising media such as radio, newspapers and television is know as the 'Mass Media Model'. The key to this approach is that marketing campaigns were typically saturation campaigns that were designed for exposure to a large group of people, with the hope that a number of them would then purchase the product (Davison et al., 2008). On the other hand a 'Hypermedia Computer-Mediated Environment' represents a many-to-many communication model for an environment such as the Web where interaction between businesses can communicate with customers via a website or automated e-mail system without necessarily conversing with them directly (Hoffmad & Novak, 1996).

At the business-to-consumer level it is possible to purchase almost anything on the Web. One of the first major businesses to make use of the web was Amazon.com, launched in 1995, but since that time many other businesses have looked carefully at this business model and made use of it. The website of the UK government organisation Businesslink.gov.uk suggests that there are many different types of e-

Figure 3. E-Business and Web Technologies



marketplace. They note that these are based on a range of different business models and can be broadly divided into categories as follows:

- **Independent e-marketplace** – usually a B-B online platform operated by a third party that is open to buyers or sellers in a particular industry. To participate, some form of payment is usually required. It allows the business to access requests for quotations or advertisements in their relevant industry sector.
- **Buyer-oriented e-marketplace** – normally run by a consortium of buyers to establish an efficient purchasing environment. This can help the buyers lower their administrative costs and obtain better prices from suppliers. Suppliers can use this sort of website to place their catalogues for relevant customers.
- **Supplier-oriented e-marketplace** – set up and operated by a group of suppliers seeking to create an efficient sales channel to a large number of buyers. This often provides buyers with information about suppliers, particularly those they may not be familiar with.
- **Vertical e-marketplaces** provide online access to businesses of a particular industry sector. These include the chemical, construction, automotive and textiles industries. The advantage for suppliers in using a vertical e-marketplace is that this can increase operating efficiency and help to decrease supply chain costs, inventories and cycle time.
- **Horizontal e-marketplaces** are able to connect buyers and sellers across different industries or regions.

Online auctions where people and businesses can buy and sell a variety of goods and services are another form of e-Marketplace and eBay is a popular example of such a site. On eBay most sales take place through a set-time auction format, but other approaches are also used.

Many large retail stores also make use of the Web to both advertise and also sell their products. Before the advent of the Web telephone shopping was quite popular and allowed a number of businesses that were too small to afford their own retail premises to do business from a private home or other premises. The e-marketplace of the Web has meant a decrease in shopping of this sort but this has been replaced by a huge increase in the use of the Web by these small businesses. For a small business, the costs of having an online operation can be much lower than setting up a shop and many new 'virtual enterprises' are now operating in this area. Paying for purchase or goods or services on the Web can be facilitated by systems such as PayPal (and Paymate in Australia) that can act as an intermediary providing secure financial transactions to collect money from the buyer and pass this on to the supplier.

Travel Industry Information and Booking Services

In times past it was necessary to go to a travel agent for booking of hotels and flights, particularly if they involved international travel. The Web has changed all this with booking and information services readily available for many hotels and all but the smallest airlines. These services could be considered under the heading of e-Marketplace, but as they involve more than just the purchase of services and also provide information I have put them into a category of their own.

E-Banking

Daniel (1999) describes Internet banking as the provision of banking services to customers through Internet technology. Karjaluoto, Mattila and Pento (2002b) indicate that this involves the provision of banking services such as accessing accounts, transferring funds between accounts, and offering an online financial service. Use of web technology in the banking industry can thus be defined as the accessibility

Figure 4. IEEE Computer Society Digital Library Portal

and exchange of banking services and/or banking communication technologies on websites available through the World Wide Web by utilising Internet technology (Al-Hajri, 2005).

Karjaluoto et al. (2002a) then identify two significant benefits that may be gained from Internet banking:

- Benefits for the banks include a reduction in transactional costs along with the ability to reach customers anywhere.
- Benefits for customers include being able to take full advantage of various banking services available online.

Although e-Banking is now common in the developed world, this is not so much the case in developing countries. A recent exploratory study (Al-Hajri, 2005) of the uptake of Internet banking in Oman (a developing country) found that in Oman, the bank managers' perceptions of four issues: relative advantage, organisational performance, customer/organisational relationship and ease of use jointly provided a broader understanding of Internet technology adoption in the banking industry.

Digital Libraries

A digital library could be described as: "an information system that supports the access, retrieval, selection, organisation and management of a focused collection of multimedia objects" (Goh, Theng, Lim, Zhang, Chang, & Chatterjea, 2007). A digital library offers access to collections of selected and organised digital resources. These resources include books, journal articles and photographs (McCarthy, 2007) and a digital library's main advantage over physical libraries lies in their ease of access. Many cultural and historical associations as well as professional societies such as the IEEE Computer Society (http://www2.computer.org/portal/web/csdl) and Association for Computing Machinery (ACM) (http://portal.acm.org/portal.cfm) and the Project Management Institute (http://www.pmi.org/Resources/Pages/) have their own digital libraries.

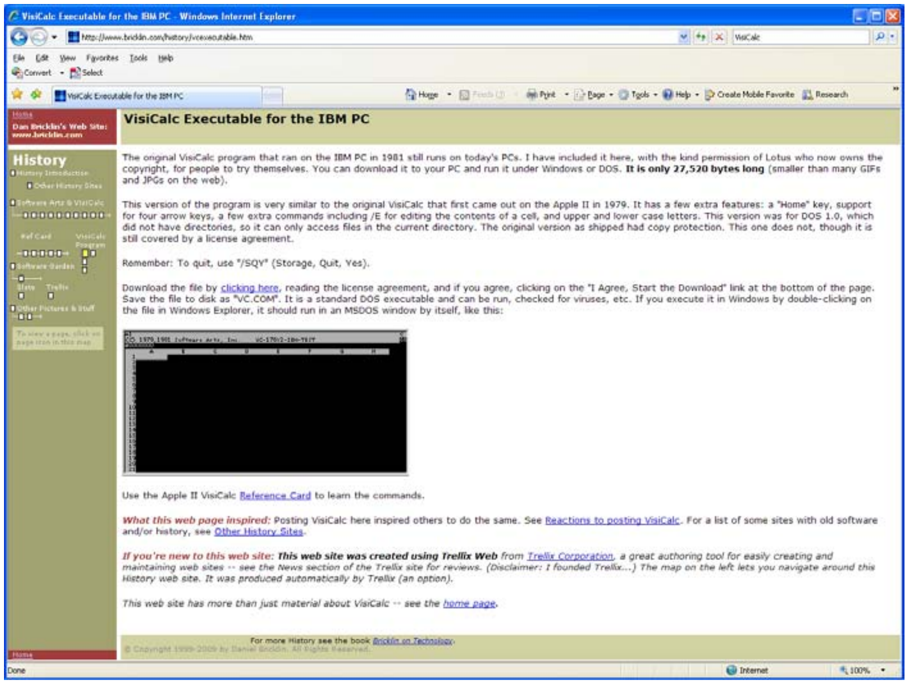
Virtual Machines and Locations

Colossus was, arguably, the world's first electronic digital computer and was used in Bletchley Park in the United Kingdom to break German military codes towards the end of World War II. A simulation of using Colossus to break the German Tunny code (Sale, 2006) is available from the Web at: http://www.codesandciphers.org.uk/anoraks /lorenz/tools/index.htm This is just one of the examples of virtual machines or operations, another is a simulation of VisiCalc, the first electronic spreadsheet which can be downloaded from: http://www.bricklin.com/history/vcexecutable.htm

The Web also offers access to a number of virtual tours of interesting locations. One such virtual tour is of the Italian city, Pompeii: http://www.italyguides.it/us/napoli/ancient_roman_city/virtual_tour_of_pompeii.htm. Another example is a tour of the Louvre Museum in Paris. This is available at: http://www.louvre.fr/llv/musee/visite_virtuelle.jsp?bmLocale=en

Each of these examples could have been accessed from files on a purchased CD, but being able to access them on the Web is yet another reason for the Web's popularity.

Figure 5. VisiCalc executable on a PC



Web Search Engines

The ability to obtain information has in the past involved going to a library and searching for this in a number of books. While that is still a worthwhile activity, it takes a good deal of time and search engines like Google, Cuil and Bing offer a much faster alternative. The convenience of using a Web search engine to find information or the answer to a question is such that people talk about 'Googling' the question and the verb 'to Google' has entered our vocabulary. There are also related specialist search engines such as Scholar Google, Google Images and Google Video. Not a lot needs to be said about search engines as they are so well known, except to remark on their power and on the huge impact they have had on our lives.

Web Encyclopaedias

An encyclopaedia in book form was (and still is) an extremely expensive purchase that many families felt obliged to make to help their children with their school work. While the book version is still useful and worth having, the online encyclopaedia has made accessing information ever so much easier. Products such as Wikipedia (http://en.wikipedia.org) have made a fundamental difference to accessing information by putting this in everyone's reach. Other Web encyclopaedia products include Encarta (http://encarta.msn.com/), Britannica online (http://www.britannica.com/) and Webopedia (http://www.webopedia.com/).

Web Portals

Web portals are now ubiquitous and researching their use in organisations and by individuals is important (Tatnall, 2007b; Tatnall, 2009d). To illustrate the wide range of sometimes quite specific applications now being filled by portals, the following list of topic categories (Tatnall, 2005) is taken from articles by the large number of academics and practitioners from around the world who contributed to the *Encyclopaedia of Portal Technology and Applications* (Tatnall, 2007a).

Portal Concepts, Design and Technology

As one might expect, portal concepts are an area of particular interest with topics such as: what is a portal? benefits and limitations of portals, comparing portals and web pages, evolution of portals, factors affecting the adoption of portals using activity theory, information visualisation, the ubiquitous portal, and portals of the mind (Tatnall, 2009d).

Research on portal design and technology also features prominently with topics such as: collaborative real-time information services via portals, digital interactive channel systems and portals, designing spatio-temporal portals for continuously changing network nodes, dynamic taxonomies and intelligent user-centric access to complex portal information, factors affecting portal design, developing semantic portals, an evolutionary approach to developing on-line learning portals in low bandwidth communities, the role of ontologies in portal design, Java portals and Java portlet specification API, large-scale ASP replication of database-driven portals, WSRP specification and alignment with JSR 168 portlet specification and user-centric knowledge representation for personalisation of Web portals.

Portal Implementation

As well as the technology itself, there needs to be some consideration of issues involved in the organisation implementing its portal application (Tatnall, 2009b). Whether the portal is based on a commercial product or if it is programmed from scratch is itself an important implementation issue. Research issues involved in the implementation of portals in specific organisations are another point of interest with topics like the following: evaluation of Web portals, portal quality issues, economical aspects when deploying enterprise portals, e-management portals and organisational behaviour, user acceptance affecting the adoption of enterprise portals, enabling technology and functionalities of shopping portals. Articles dealing with implementation issues often involve a case study of the organisation implementing the portal and discuss the goals and intentions of using a portal, how the technology was chosen, the implementation approach used and the problems incurred during the implementation. Not all implementation issues, of course, are technical ones as implementations all involve people.

Portal Uses and Applications

The largest area of research interest is in how portals are applied and used, and most of this research refers to quite specific applications such as: the Bizewest portal, the Bluegem portal, the European quality observatory portal, the future of portals in e-science, hosting portals on an e-marketplace, how corporate portals support innovation, how the Internet is modifying the news industry, industry portals for small businesses, portals for business intelligence, strategic planning portals, study of a wine industry portal, supplier portals in the automotive industry, supply chain management and portal technology, portal economics and business models, portals for integrated competence management, cultivating memories through the Beijing Olympics (2008) Advertainment portal, portals for workflow and business process management, project management Web portals, provision of product support through enterprise portals, employee self-service portals, a generic model of an enterprise portal, portal technologies and executive information systems implementation, the role of portals in consumer search behaviour and product customisation, guided product selection and comparison of e-commerce portals, business challenges of online banking portals, Web museums, Web portals as an exemplar for tourist destinations and a Web portal for the remote monitoring of nuclear power plants (Tatnall, 2009d). Within this area of applications it is possible to identify a number of major topics of interest. These are:

- **Education Portals**. Specific example include: academic management portals, large scale integrated academic portals, mobile education portals, artificial intelligence and education portals, high school portals, primary school portals, corporate e-learning portals, Weblogs, knowledge portals in education, and subject teaching portals.
- **Health and Medical Portals** examples include: empowerment and health portals, bioinformatics portals, biotechnology portals, nursing knowledge portals, network-centric healthcare and the entry point into the network and genomic and epidemiologic medical data portals.
- **Community Portals**. Topics researched in this area included: how to promote community portals, a community geographic domain names portal, designing a portal and community-community generator, local community Web portals and small businesses and the paradox of social portals.
- **E-Government Portals**. This is an area related to Community Portals and there was much research interest in government portals around the world. There were topics such as: portals in the public sector, e-government portals, e-value creation in a government Web portal in South Africa, government portals as a gateway for enhancing electronic governance in Singapore, interoperability integrating e-government portals, modelling public administration portals, service quality in the case of e-government portals, and state portals as a framework to standardise e-government services.
- **National Portals**. There is also important research into portals related to national issues: African Web portals, business module differentiation and a study of the top three Chinese portals, cross-cultural dimensions of national Web portals, growth of e-portals in Dubai, how portals help Chinese enterprises operate successfully in global markets, impacts and revenue models from Brazilian portals, Web museums and a case study of the French population.
- **Personal and Mobile Portals** is an area of growing interest as mobile technology continues to mature. Examples of this area are: accessible personalised portals, mobile portal technologies and business models, mobile portals as innovations, mobile portals for knowledge management, the MP3 player as a mobile digital music collection portal, widgets as personalised mini-portals, wireless local communities in mobile commerce and portals supporting a mobile learning environment.
- **Knowledge Management, Libraries and Professional Societies**. Knowledge Management, especially relating to libraries and professional societies is another area which attracts a number of researchers. They were interested in topics such as: designing portals for knowledge work, mobile portals for knowledge management, knowledge servers, the portal as information broker, portal strategy for managing organisational knowledge, a prototype portal for use as a knowledge management tool to identify knowledge assets in an organisation, library portals and an evolving information legacy, open access to scholarly publications and Web portals, the IFIP portal, and the portal features of major digital libraries.

Intranets

The term Enterprise Information Portal (or sometimes ‘corporate portal’) can be applied to the gateways to corporate intranets that are used to manage the knowledge within an organisation (Davison et al., 2008). These are often designed for business-to-employee (B-E) processes that offer employees a means to access and share data and information within the enterprise. An intranet offers a number of useful applications:

- **Access to company documents** such as minutes of meetings, forms, policies, procedures, phone lists, information documents, documentation and archived information.
- **Software access and download**. Access to common software applications (on a central server), software downloads, software updates (including anti-virus software).
- **Services** such as information and communications technology (ICT) technical support and diagnosis, training materials and on-line training, human resource data (e.g. leave entitlements), search facilities to search the site for a specific item.
- **Communication** – broadcast company information or policy, bulletin boards, chat, electronic mail, newsgroups, on-line meetings, video conferencing.
- **Consistent front-end to software applications** – single point of access to other systems, consistent interface to other systems, complexity of databases and applications is hidden, front-ends to legacy systems.
- **Links to outside systems** – the Internet, the company extranet (and possibly to business partner’s systems).

Extranets

Extranets have a number of uses including replacement of direct computer-to-computer Electronic Data Interchange (EDI). These extranets can use Internet technology to perform the same transactions, although sometimes rather more slowly. Extranets can also be used to assist with Supply Chain Management. A supply chain links manufacturers, retailers, customers and suppliers and involves the co-ordination of order generation, order taking, order fulfilment, order distribution and information sharing. Access by business partners to parts of the organisation’s intranet can be used to facilitate this process.

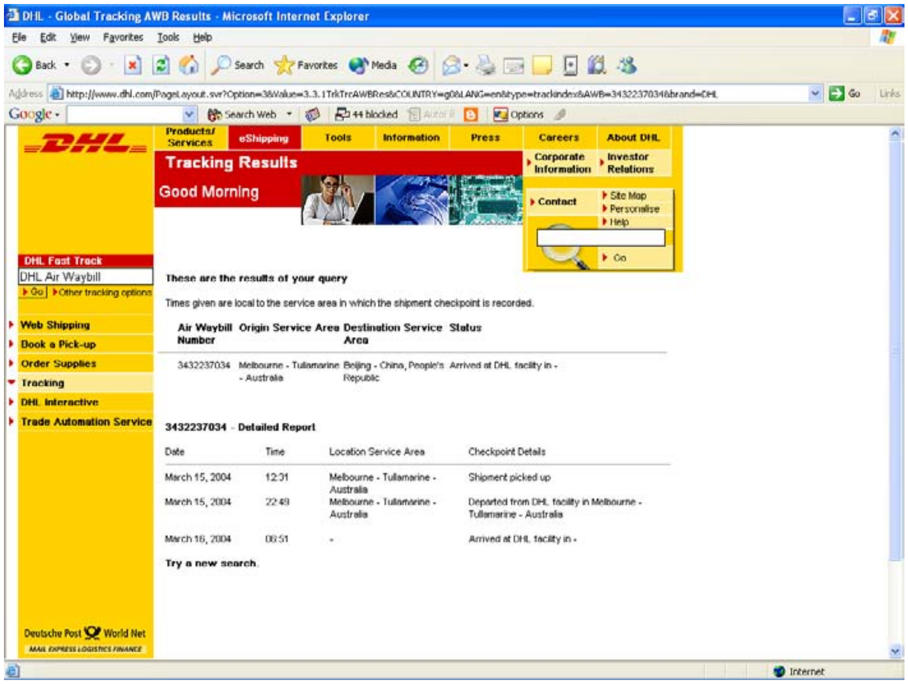
An extranet can also be used to facilitate ‘value added marketing’ as the organisation’s Web site will have basic information of interest to its customers. There may, however, be situations where it is useful to allow a customer access to its core systems to obtain, for example, account information. Another example is that a computer supplier may provide (password) extranet access to its own library for important customers.

Before the Internet, freight delivery companies like DHL (www.dhl.com.au) needed to spend a lot of time on the phone telling customers the location of their deliveries. It then occurred to someone that it would be better to let their customers see this for themselves by offering (limited) access to their own internal systems via an extranet.

Communication

The best know form of communication offered by the Web is, of course, e-mail. This is so well know to most people that I will say no more about it here except that it was available on the Internet before the Web came into existence, but was then much more difficult to use.

Figure 6. DHL's parcel tracking extranet



Another form of Web enabled communication is Skype (http://www.skype.com/intl/en/) which is a software application that enables its users to use the Internet to make almost free phone calls. Skype uses a proprietary form of Voice over Internet Protocol (VoIP) that allows the call to make use of the Internet rather than the normal phone system. Skype also offers other features including video conferencing and instant messaging.

Chat rooms are another form of communication and are offered by a number of companies. These include: Yahoo (http://messenger.yahoo.com/features/chatrooms), Match.com (http://www.match.com.au), Google Talk (http://www.google.com/talk/) and Gmail (http://mail.google.com/mail/help/chat.html). Chat simulates face-to-face communication in that the topic can be anything that the chatters want it to be and people can come and go at will. One big difference, however, is that Web chat is anonymous in many cases. In this respect it is a form of communication that is particularly related to the Web.

Group Decision Support Systems

Applications like Google Groups (http://groups.google.com/) and Yahoo Groups (http://groups.yahoo.com/) offer the possibility for groups of people to easily communicate with each other by e-mail. While this does not directly offer decision support, it does achieve part of this in allowing a group of people to communicate easily and discuss the topic under consideration.

A number of researchers have investigated Web-based group decision support systems (GDSS). Abdelhakim and Shirmohammadi (2007), for example, discuss a Web-based GDSS that can be used to select and evaluate educational multimedia and Ma, Quan and Zhiping (2009) describe a Web-based GDSS for Research and Development project outcome assessment in government funding agencies. The company DSS Resources (http://dssresources.com/) offers more product advice in this area.

The Social Web

The Web has also opened up a wide range of possibilities for people to socialise remotely in various ways. Whether socialising via the Web is a good thing or not rather depends on your point of view and I will come back to discuss this later.

Facebook (http://www.facebook.com/) is perhaps the most commonly known social networking Web site. It is designed so that people can join a network organised by their school, workplace, region or other similar entity and then interact with other people in this same network. They can then create their own personal profile, notify other people about themselves and send and receive messages from these people.

MySpace (http://www.myspace.com/) is another social networking Web site that offers similar facilities to Facebook.

LinkedIn (http://www.linkedin.com/) is a business-oriented social networking Web site that is mainly used for professional, rather than personal, networking.

Twitter (http://twitter.com/) is another social networking Web site that is used to send and receive short (140 character) text messages called tweets. These are displayed on their author's profile page and sent to other subscribers (know as followers). Twitter users can send and receive their tweets on their mobile phones using SMS (Short Message Service).

Another more specific example is **Academia.edu** (http://www.academia.edu/) which is designed, as its name suggest, to link academics and researchers.

YouTube (http://www.youtube.com/) is a video sharing Web site where users can upload and share videos they have made. These can be viewed on a PC in MPEG-4 format.

e-Learning

Electronic learning (e-Learning) involves support of learning using Web technology as the medium for teaching. In some cases, especially where the students are at a location that is remote to the teaching institution, it is used to replace face-to-face teaching entirely. In other cases it is used to supplement classroom learning. This form of teaching is used in universities and some schools, but also for training in many large companies, and particularly in companies whose operations span the globe. The largest provider of e-Learning applications is Blackboard Inc (http://www.blackboard.com/) whose products include WebCT and Blackboard. These products typically offer the facility for discussion boards, e-mail systems, live chat, Web pages and downloading content including documents. An alternative open source e-learning package is available from Moodle (http://moodle.org/).

For the student, one big advantage of e-learning is that they do not need to attend the teaching institution in person but can undertake their studies from home or their place of employment. When combined with mobile computing the possibilities for undertaking learning at any time and any location are great.

Mobile Computing and the Web

While mobile computing relates not only to the Web but also to other aspects of computing, it is particularly relevant to discuss this here as it offers the possibility of viewing Web content at any location. Few people would now question the idea that computers and communications devices can be small enough and mobile enough to use anywhere, but this has not always been the case. The first computers were huge machines filling whole rooms and weighed several tonnes and telephones were anything but portable until quite recently. Laptop and net computers, Personal Digital Assistants (PDA) and smart phones are now so small as to be extremely portable, but it is not just the hardware here that is significant but also the protocols for short-range radio communications that they require. There are two main existing standards for short-range radio communications (Davison et al., 2008):

- **IEEE 802.11** (WiFi) is a set of standards for wireless local area networks, developed by IEEE, that operates in the 5 GHz and 2.4 GHz public spectrum bands over a 50-100 metre radius at a maximum data transfer rate of up to about 50 Mbps.
- **Bluetooth** offers a 10 metre range with a peak data transfer speed of 1-3 Mbps. It is possible to operate at least ten Bluetooth devices within a 10 metre radius. Bluetooth provide a means of linking digital devices into Personal Area Networks (PAN)

Each of these protocols offers the possibility for a portable device to link to other portable devices or to access the Internet at almost any location. In city areas a number of retail premises (including McDonalds) offer free Internet WiFi 'hot spots' where customers (and those nearby) can gain access to the Internet for no cost using an IEEE 802.11 protocol in their laptop. Other locations, including most hotels, also offer wireless Internet access but at a cost.

The important point here is that these mobile devices, in conjunction with a WiFi protocol, make it possible to access the Web from almost any populated location.

Web-Based Gambling

While not permitted in some parts of the world, in others online gambling is perhaps being used at times to circumvent local gambling laws or at least to make it much easier for gamblers to indulge in this activity. Sites such as Web Gamble (http://www.webgamble.org/) and Wager Web (http://www.wagerweb.com/) are just a few of many sites that offer this facility. The Casino Gambling Web site (http://www.casinogamblingweb.com/) provides news reports and information about online gambling. For those with a gambling problem Gamblers Anonymous also has its own Web site (http://www.gamblersanonymous.org/).

ORGANISATIONAL, MANAGERIAL, SOCIAL AND POLITICAL IMPACT OF WEB TECHNOLOGIES

Organisational and Managerial Impacts

Are organisations different to manage when they make extensive use of the Web? The answer is in most ways that they are not, but there are some issues of which a manager must take account. I will mention just a few. The first is the ease of access to information by both employees and customers. An organisation can make good use of the ease with which it and its employees can gain information, but sitting an employee in front of a computer to search can also be a very time wasting exercise and it is important that proper goals and time limits are set for this type of activity. That customers can also easily obtain information about competitors and other products can be a problem which must be recognised by a business. Steps need to be taken to become aware of this problem and to seek means to overcome it

At a different level another managerial issue is in the use of an organisational intranet. It is quite easy to set up an intranet that can be used to store all sorts of company documents such as: policies, forms, procedures, minutes of meetings, internal phone lists and so on. While in the past it was necessary to distribute these documents in paper form they can now be put on the intranet and made easily available to all staff. There will now be no excuse for someone claiming ignorance of a company procedure or policy because they could not get hold of a copy of it. The intranet can also be used to broadcast messages from the CEO or other senior managers to all employees. If not abused this can be a very useful facility. Updates and patches for computer software can also be delivered in this way, making it possible for the organisation to keep much more up to date.

It is sometimes claimed that if some information, such as a catalogue or price list, is on the organisation's Web site then it must be up to date. This is, of course, absolute rubbish and a better claim is that if it is on the organisation's Web site then there is the *possibility* that it is always kept up to date. This possibility will only eventuate however, if it is *someone's job* to keep information on the company Web site up to date.

Another exciting management possibility with use of the Web is the creation of a virtual enterprise. Such an organisation does not need a shop front to interact with its customers but can order and supply its products remotely. Managing such an enterprise, where some employees operate from one location and others operate from different locations is quite different to running a business located under one roof and needs quite a different management style.

Social and Political Impacts

In common with the introduction of most new technologies, the advent and growth of the Web comes with the dual possibilities of good or ill ey and & Tatnall, 2007). Along with freedom in the transmission of knowledge comes a loss of control for authorities and the possibility of chaos. The introduction of Guttenberg's printing press in circa 1455 allowed mass production of the Bible and eventually changed the nature of Christianity in Europe. For many years printing was the technology used to distribute information. The came the telephone, and this technology allowed people to keep in contact remotely and to distribute news and information. The Web furthered this democratisation of information and allowed people to spread news and to keep in touch without the need to go through any official or governmental channels. Censorship and control of the Internet in some countries is another hot issue that is unlikely to go away. George (2005) suggests that the relationship between new technological media and political factors is far too dynamic and interdependent to be reduced to simple causal statements. He suggests that the less democratic the society, the more attractive the Internet looks as an emancipatory medium, but the more likely radical Internet use will be blocked or punished.

The role of the Internet in mobilising and assisting various protest movements is an interesting topic. Stohl and Ganesh (2009) note the importance of the Internet for spreading information, organising and constructing networks and forming individual and group identities. They also suggest that it can even act as a form of protest itself. An article in the Wall Street Journal (Qiang, 2005) discussed the role of the Internet in organising protests in China in 2005. The article noted that "the most fervently pro-nationalist segment of society overlaps demographically in China with those who spend the most time on the Internet, since both are primarily composed of young, educated, urban males" and goes on to describe how recent protests were almost exclusively organised using the Web.

The Internet & Democracy Blog (http://blogs.law.harvard.edu/idblog/2009/04/07/moldovan-youth-organize-protests-with-twitter/) describes how Moldovan youth organised protests with the aid of Twitter. It describes how Twitter not only helped rally protesters but also provided a real glimpse of what was happening on the ground. Debate also has been progressing on the role of technology in Thailand's recent protests (2009) http://neteffect.foreignpolicy.com/posts/2009/04/17/technologys_dubious_role_in_thailands_protests.

How use of the Web affects democracy and political activities is an interesting and evolving study. One example is that in recent elections in Australia both the Prime Minister (at the time) and the then Leader of the Opposition made considerable use of YouTube to spread their message. Similar use of technology also occurred in other countries and no doubt we will see an increase in the use of Web technologies of all types in the political process. A detailed study in Korea undertaken by Woo-Young (2005) found that citizen e-participation was characterised by: convenient access to detailed information, free expression and exchange of opinions, online activism led by politicised agenda and active formation of cyber groups. Woo-Young notes that Internet political portals are not connected with existing political power and capital, and that they facilitate communication between citizens.

Hutchins and Lester (2006), however, suggest that the use of the Web by environmental groups is an example of mob rule. The existence of a Web portal does not indicate the number of people subscribing to the philosophy of its owners. The Web site allows coordination of efforts, such as protests, in a way that may make the actions of a group look more important than the size of the group might support.

Francis Bacon is credited with the expression 'Knowledge is Power' and the growth in use of Web technologies has brought this to the fore with the creation of new classes of information rich and information poor depending on their access to this technology. Access to huge amount of information is now possible, much than anyone could have dreamed of a few years ago. With access to the Web many people now also have access to encyclopaedias, news, research results and much more. Those without access to the Web miss out.

Undesirable Impacts

There are also many undesirable impacts of the growth of the Web. Many people would see the increase in Internet gambling as one of these and most would see as a bad thing the Web's role in making much easier the dissemination of pornography. Theft and fraud resulting from improper Internet financial transactions are a major problem, and identity theft has been made easier by the increase in general use of the Web.

Another problem in general use of the Web is the increased possibility of misinterpretation of information. Medical Web sites abound on the Internet with almost every major disease being represented by at least a support group portal. These Web sites offer everything from emotional support and possible treatment advice to contacts within the medical community (Davey & Tatnall, 2007), but this is not always seen as useful and Theodosiou and Green (2003) identify five important problems with patients using medical portals to satisfy their needs:

- Potentially dangerous drugs and other substances may be bought by individuals for themselves or their children.
- Individuals can spend a lot of money on products or diagnostic procedures that have no scientific backing or benefit.
- The information may be more negative than the reality of the situation.
- Individuals may abandon treatment programmes of proven efficacy to pursue less-mainstream approaches.

Information that had previously been sold by its originators soon began to be dispersed without charge via the Web and the owners of the software, music and videos being freely distributed invented the term Internet piracy. Yar (2005) indicates the size of the possible loss to copyright owners in terms of billions of dollars. Plagiarism is another evil that has been made much easier and more common by use of the

Web. When it is so easy for a student to copy a passage from a journal article found in a digital library and paste it into their essay it is little wonder that plagiarism has become a major problem in universities and other educational institutions.

Finally, there is the question of whether spending large periods of time in front of a computer should be seen as a non-human activity that discourages normal social interaction. While the use of social networking sites offers many possibilities for social interaction of a sort, unless this is followed up by face-to-face contact of the normal type there could be something unhealthy about this use of the Web.

THE SOCIO-TECHNICAL NATURE OF WEB SYSTEMS

Sometimes those of us who design information systems of any type get lost in the elegance and intrinsic worth of our design efforts and forget that for an information system to have any value, it must first be adopted and put into use. There is an old Anglo Saxon saying that 'You can lead a horse to water, but you can't make it drink', and something similar can be said about the adoption and use of information systems. Because of their human and non-human aspects, Web-based information systems should be considered as socio-technical entities, and it is most important that both of these aspects are kept in mind throughout the process of systems design and implementation. This is especially true of Web-based information systems.

Building any type of information system is a difficult task, partly due to the problems in ascertaining the requirements of the intended users, but also because of the complexity of the large number of human-machine interactions (Tatnall & Davey, 2005) that such systems typically involve. This complexity is reflected in the difficulty of building these systems to operate free from error and to perform as intended (Tatnall, 2009a), and building Web based systems is no less complex.

The Information Systems (IS) discipline is primarily concerned with the ways that people build and use computer-based systems to produce useful information. It thus always has to deal with issues involving both people and machines, and with the multitude of human and non-human entities that comprise an information system (Tatnall, 2003). The discipline of Information Systems is thus neither merely a technical discipline nor a social one, but one that is truly socio-technical. Systems developers and researchers all face the problem of how to handle the many complexities due to interconnected combinations of people along with computers, peripherals, procedures, operating systems, programming languages, software, broadband connections, switching technology, data, databases and many other inanimate objects. They need to consider how all these elements relate to humans and human organisations, and how humans relate to them (Longenecker, Feinstein, Couger, Davis & Gorgone, 1994; Tatnall, 2009a).

ADOPTION OF WEB TECHNOLOGY: INNOVATION

Just because a new Web technology exists it cannot automatically be assumed that organisations or individuals will want to adopt or to use it. A Web technology will only be adopted if potential users make a decision to do so and the adoption of a technological innovation, such as a Web technology, occurs for a variety of different reasons. Thus the first step in researching the use of a new Web technology by an organisation (or an individual) is to investigate why it was adopted, and so consider the Web technology as a technological innovation. This can be done by examining the adoption of the new Web technology through the lens of innovation theory.

It is important at this stage to distinguish between invention and innovation. Invention refers to the construction of new artefacts or the discovery of new ideas, while innovation involves making use of these artefacts or ideas in commercial or organisational practice (Maguire, Kazlauskas & Weir, 1994). Invention does not necessarily invoke innovation and it does not follow that invention is necessary and sufficient for innovation to occur (Tatnall, 2009b). Clearly the Web technology can be seen as an invention, but the point here is that it will not be used unless it is adopted, and that means looking at it also as a technological innovation. Of course, the application of innovation theory to the adoption of a new Web technology assumes that the potential adopter has some choice in deciding whether or not to make the adoption. In the case of an organisation or individual considering the adoption and use of a Web technology, however, it is difficult to see any reason why they would not have a large measure of choice in this adoption decision. This makes the application of adoption theory quite appropriate when considering the use of Web technology.

A number of approaches exist to modelling how technological innovation takes place, including the Theory of Reasoned Action (Ajzen & Fishbein, 1980), the Theory of Planned Behavior (Ajzen, 1991), the Technology Acceptance Model (Davis, 1986), Diffusion of Innovations (Rogers, 1995; Rogers, 2003) and Innovation Translation (Callon, 1986b; Latour, 1996). In particular, both the Diffusion of Innovations and the Technology Acceptance Model (TAM) are very well known and widely used approaches to theorising technological innovation.

RESEARCHING WEB TECHNOLOGY AND ITS APPLICATIONS

This multi-volume book series contains a large number of research articles on a variety of topics relating to Web technologies, and in this section I will look at how research in Web technology can be categorised (Tatnall, 2009b) and how it is being undertaken. While some of articles on Web technology are fairly descriptive, perhaps examining a commercial product in some depth, others involve detailed analytical research. It appears to me that most of the articles outlining research on Web technology could be classified into three main areas which could, of course, be further subdivided:

- **Web technology:** research on the technical design of computer hardware and software for Web applications.
- **Web implementation:** research issues involved in the implementation of Web sites in specific organisations.
- **Web applications:** research into the many uses of the Web, and the social and political consequences of this use.

Web technology itself has been researched by many scholars, especially those involved in the design and implementation of Web-based systems. This is a wide field and offers many possibilities for fairly technical research articles dealing with various aspects of these systems. Commercial Web technology products and vendors each have an important role to play and their investigation and evaluation provides a profitable avenue of research. Quality issues and standards as well as measurement of effectiveness could also be validly considered. Another important consideration here is whether certain implementation factors are more likely to lead to successful *adoption* of Web technology than others. I will take this point further shortly.

As well as the technology itself, there needs to be some consideration of issues involved in the organisation implementing its Web-based application. Whether the Web technology is based on a commercial product (such as Front Page) or if it is programmed from scratch (in a programming language such as Java or Visual Basic) is itself an important implementation issue. Articles dealing with implementation issues often involve a case study of the organisation implementing the Web-based application and discuss the goals and intentions of using this: how the technology was chosen, the implementation approach used and the problems incurred during the implementation. Not all implementation issues, of course, are technical ones as implementations all involve people.

No doubt the largest area, however, consists of articles that investigate Web-based systems uses and applications, of which there are many. A major user of Web technology around the world is governments and the public sector and most government departments now have their own Web sites. Social and community-based Web sites are also common. At the personal level, research is often conducted into topics including Weblogs, widgets and MP3 players. Medical, health and bio-informatics Web sites form another significant group of applications as do the business and industrial sectors. Much research investigates organisational and management issues regarding Web technology use, human resources and enterprise information portals, Web sites for small to medium enterprises and more specific topics including shopping, the automotive industry and wine industry Web sites. The economics of setting up and using these Web sites is also discussed, as are issues of strategic planning, user acceptance, security and the law. Applications in areas such as business intelligence, artificial intelligence, semantic portals, intelligent agents and mobile technology could also be included here.

The approach used in Web-based systems research depends, of course, on the type of research being undertaken and will most likely differ depending on whether this is related to Web technology itself, implementation or applications. Should a purely descriptive article be considered as research? It is not the place of this chapter to discuss this, but the sort of article I will discuss involves at least some degree of analysis. Articles considering specific Web technology products and vendors will most likely look at some form of analysis of the benefits of this feature or that design approach and may be quite general in nature. As mentioned previously though, apart from articles discussing Web technology itself, most research will in some way relate to the use of this Web technology by some particular organisation and so will probably involve some form of case study.

One research approach that attempts to take proper account of both the social and technical aspects of Web-based information systems is that of Soft Systems Methodology (SSM), a variant on action research developed by Peter Checkland and his colleagues (Checkland & Scholes, 1991) from Lancaster University. SSM attempts to give due recognition to both the human and technological aspects of a system and acknowledges both human and non-human aspects of these systems, which it considers as entirely separate types of entity.

Another research approach is actor-network theory (ANT) that also considers both social and technical aspects of an information system. ANT, however, considers that it is not possible to distinguish between these two types of entities and so attempts to consider and handle both human and non-human entities in the same way.

Actor-network theory proposes that everything we do involves hybrid entities (Latour 1993) containing both human and non-human elements. ANT was developed to analyse situations where separation of these elements is difficult (Callon 1999; Tatnall and Gilding 1999). One could question, for instance, which parts of an information system are just inanimate objects and which are the result of human interactions. If we consider a Web portal, for example, it is difficult to differentiate its technical aspects from the influence exerted by the socio-cultural background of the development team (Cusumano & Selby, 1997; Sahay, 1997). What seems, on the surface, to be social is partly technical, and what may appear to be only technical is partly social.

ANT handles this difficulty in determining the social-technical divide by refusing to allow the possibility that either purely technical or purely social relations can exist. It instead offers the notion of heterogeneity to describe situations such as the development of a website or other web-based system that involves computer hardware, computer software, communications hardware, communications software, telecommunications authorities, cables, portal hosting sites, Internet service providers, data, human analysts, human programmers, human users and so on. The use of heterogeneous entities that can be partly social and partly technical (Bijker, Hughes & Pinch, 1987) then avoids questions of: 'is it social?' or 'is it technical?' as missing the point, which should be: "is this association stronger or weaker than that one?" (Latour, 1988, p. 27). ANT considers both social and technical determinism to be flawed and proposes instead a socio-technical account (Latour, 1986; Law & Callon, 1988) in which neither social nor technical positions are privileged.

"ANT was developed to analyse situations in which it is difficult to separate humans and non-humans, and in which the actors have variable forms and competencies." (Callon, 1999, p:183)

In order to address the need to treat both human and non-human actors fairly and in the same way, ANT was designed upon three fundamental principles (Callon, 1986b):

- **Agnosticism:** analytical impartiality is demanded towards all the actors involved in the project under consideration, regardless of whether they are human or non-human.
- **Generalised symmetry:** explains the conflicting viewpoints of different actors in the same terms by use of an abstract and neutral vocabulary that works the same way for human and non-human actors. Neither the social nor the technical elements in these 'heterogeneous networks' (Law, 1987) are then given any special explanatory status.
- **Free association:** requires the abandonment of all distinctions between the technological and the social (Callon, 1986b; Singleton & Michael, 1993).

Actor-network theory thus attempts impartiality towards all actors in consideration, whether human or non-human, and makes no distinction in approach between them. Callon (1986b, p. 200) puts it this way: "The rule which we must respect is not to change registers when we move from the technical to the social aspects of the problem studied." Callon (1987) further proposes that entities gain strength by gathering a 'mass of silent others' into a network to give them greater strength and credibility. This network then becomes durable partly due to the durability of the bonds that hold it together.

In ANT, an actor is any human or non-human entity that is able to make its presence *individually felt* by the other actors (Law, 1987). It is made up *only* of its interactions with these other actors (de Vries, 1995), and Law (1992) notes that an actor thus consists of an association of heterogeneous elements constituting a network. It should not be considered as a 'point object' but rather as an association of heterogeneous elements themselves constituting a network. Each actor can thus also be considered to constitute a simplified network (Law, 1992). In this sense, an actor can in many ways also be considered as a black-box, and when the lid of the box is opened it will be seen to constitute a whole network of other, perhaps complex, associations (Callon, 1986a). In many cases, details of what constitutes an actor – details of its network – are a complication we can avoid having to deal with all the time. We can usually consider this entity just as an actor, but when doing this it must be remembered that behind each actor there hide other actors that it has, more of less effectively, drawn together (Callon, 1987).

After development, as technological innovations are often not adopted in their entirety or in exactly the form that their proponents suggested, ANT makes use of a theory of Innovation Translation (Latour, 1986; Law & Callon, 1988; Latour, 1996) which suggests that before adoption, an innovation is first 'translated into a form which is more appropriate for use by the potential adopter. Callon et al. (1983) propose that translation involves all the strategies through which an actor identifies other actors and reorganises them in its own way.

The main advice on method suggested by the proponents of actor-network theory is to "follow the actors" (Latour, 1996) and let them set the framework and limits of the study themselves. The process followed in an ANT analysis is thus far from linear, and not just simply a matter of collecting data and then analysing it.

I will argue that two things make actor-network theory very appropriate for researching and discussing web-based systems and for framing all but the type of research that only describes the technology itself. First is its treatment of both human and non-human actors and of the interactions between them. There is no doubt, in my view, that both human and non-human actors including the hardware, software and communications technologies themselves, jointly effect the outcome of implementing any web-based system. To consider one and ignore the other would give a very one sided view of what is going on. For example, suppose that a particular technology is too expensive or too difficult to use. There is then little likelihood of it being adopted and used properly. On the other hand suppose that this technology does not fit with the way that a particular organisation does its business. This is not necessarily due to the nature of the technology alone any more than it is due alone to the human factors of the way that business is done. It is more likely due to the interaction of each of these factors. In another situation, what one person or organisation considers the perfect technological solution will not necessarily suit another. The reasons it might not suit them could be complex and have more to do with cultural or social issues than technical issues. A consideration of both human and technological aspects of these systems then is essential.

Another advantage of using ANT is in its treatment of the adoption of innovations. Most other approaches to innovation adoption consider only whether some technology is adopted or not and do not give put much store on partial adoption. When investigating this area it is often the case that one person or company will adopt certain features of Web technology and not others that it could have been expected to adopt. There will then often be another similar company that adopts other different features of the same technology. One must then question the reason for this, and actor-network theory's approach using Innovation Translation has the ability, I would suggest, to explain this well. ANT suggests that before adoption, a technological innovation must first be 'translated' into a form appropriate for adoption by the organisation concerned. It so offers the means for explaining partial adoptions. In my own research I have found actor-network theory a very useful research methodology in working with socio-technical systems involving Web technology (Tatnall, 2009c).

ISSUES AND TRENDS: CONCLUSION

It is always dangerous to attempt any predictions into the future of technology. Thomas Watson (Senior), then President of IBM, is alleged to have stated in 1943: "I think there is a world market for maybe five computers". Whether Watson made this statement of not, and there is little evidence that he actually did, it highlights the danger in making future predictions related to information and communications technologies. I am, however, I think fairly safe in saying that for the future of the Web three things seem fairly clear.

1. The Web will continue to grow in size as move users (particularly in developing countries) move to adopt web technologies and more information is put onto the Internet. No doubt this will put considerable pressure on the infrastructure, but with new technologies this problem should be overcome.
2. The speed of access to the Web will continue to increase as governments around the world put in new higher speed broadband infrastructure. This will mean that websites can contain even more multi-media material.
3. More people will access the Web with the aid of mobile computing technologies, meaning that the Web will become more accessible everywhere.

Questions that do not have a clear answer are whether access to most websites will continue to be free of cost and whether governments around the world will keep their hands off the Web or want to exercise a greater degree of control as some governments now do. Given that the Web has been around for only about 15 years it will be interesting to see what a book about the Web in 15 years time will say – perhaps there will then be no need for such a book as the Web will have become so well know that there will be no need for one.

REFERENCES

Abdelhakim, M. N. A., & Shirmohammadi, S. (2007). *A Web-Based Group Decision Support System for the Selection and Evaluation of Educational Multimedia*. International Multimedia Conference Augsburg, Germany ACM.

Ajzen, I. (1991). The Theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179-211.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. London, Prentice-Hall, Englewood Cliffs.

Al-Hajri, S. (2005). Internet technology adoption in the banking industry. *Business*. PhD Thesis, Victoria University, Melbourne. .

Barras, C. (2009). Internet wonders: How big is the Net? *New Scientist, 202*, 30.

Bijker, W. E., Hughes, T. P., & Pinch, T. J. (Eds.) (1987). *The social construction of technological systems: new directions in the sociology and history of technology*. Cambridge, MA: MIT Press.

Callon, M. (1986a). The sociology of an actor-network: The case of the electric vehicle. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology* (pp. 19-34).

Callon, M. (1986b). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. In J. Law (Ed.), *Power, action & belief. A new sociology of knowledge?* (pp. 196-229). London: Routledge & Kegan Paul.

Callon, M. (1987). Society in the making: The study of technology as a tool for sociological analysis. In W. E. Bijker, T. P Hughes, & T. P Pinch (Eds.), *The social construction of technological systems* (pp. 85-103). Cambridge, MA: The MIT Press..

Callon, M. (1999). Actor-network theory - The market test. In J. Law & J. Hassard (Eds.), *Actor network theory and after* (pp. 181-195). Oxford: Blackwell Publishers..

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information, 22*(2), 191-235.

Checkland, P., & Scholes, J. (1991). *Soft systems methodology in action*. Chichester: Wiley.

Cusumano, M. A., & Selby, R. W. (1997). How Microsoft builds software. *Communications of the ACM, 40*(6), 53-61.

Daniel, E. (1999). Provision of electronic banking in the UK and the Republic of Ireland. *International Journal of Bank Marketing, 17*(2), 72-82.

Davey, B. and Tatnall, A. (2007). The paradox of social portals. In A. Tatnall (Ed.), *Encyclopedia of Portal Technology and Applications* (pp. 689-693). Hershey, PA, Information Science Reference.

Davis, F. D. (1986). *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. Unpublished doctoral thesis, MIT, Boston.

Davison, A., Burgess, S., & Tatnall, A. (2008). *Internet technologies and business*. Melbourne: Data Publishing.

de Vries, G. (1995). Should we send Collins and Latour to Dayton, Ohio? *EASST Review, 14*(4).

George, C. (2005). The internet's political impact and the penetration/participation paradox in Malaysia and Singapore. *Media, Culture & Society, 27*(6), 903-920.

Goh, D. H.-L., Theng, Y.-L., Lim, E.-P., Zhang, J., Chang, C. H., & Chatterjea, K. (2007). Learning geography with the G-Portal Digital Library. In A. Tatnall (Ed.), *Encyclopaedia of Portal Technology and Applications* (pp. 547-553). Hershey, PA, Information Science Reference. Hoffman, D. L., & Novak, T. P. (1996). A new marketing paradigm for electronic commerce. Retrieved 2009, from http://sloanwordpress.ucr.edu/1996/10/17/working-paper-hoffman-and-novak-february-1996

Hutchins, B., & Lester, L. (2006). Environmental protest and tap-dancing with the media in the information age. *Media, Culture & Society, 28*(3), 433-451.

Karjaluoto, H., Mattila, M., & Pento, T. (2002a). Electronic banking in Finland: Consumer beliefs and reactions to a new delivery channel. *Journal of Financial Services Marketing, 6*(4), 346-361.

Karjaluoto, H., Mattila, M., & Pento, T. (2002b). Factors underlying attitude formation towards online banking in Finland. *International Journal of Bank Marketing, 20*(6), 261-272.

Latour, B. (1986). The Powers of association. In J. Law (Ed.), *Power, action and belief. A new sociology of knowledge? Sociological Review monograph 32* (pp. 264-280). London: Routledge & Kegan Paul.

Latour, B. (1988). The Prince for machines as well as for machinations. In B. Elliott (Ed.), *Technology and social process* (pp. 20-43). Edinburgh: Edinburgh University Press.

Latour, B. (1993). *We have never been modern*. Hemel Hempstead, Harvester Wheatsheaf.

Latour, B. (1996). *Aramis or the love of technology*. Cambridge, MA: Harvard University Press.

Law, J. (1987). Technology and heterogeneous engineering: The case of Portuguese expansion. In W. E. Bijker, T. P. Hughes, & T. J. Pinch (Eds.), *The social construction of technological systems: New directions in the sociology and history of technology* (pp. 111-134). Cambridge, MA: MIT Press.

Law, J. (1992). Notes on the theory of the actor-network: Ordering, strategy and heterogeneity. *Systems Practice*, *5*(4), 379-393.

Law, J., & Callon, M. (1988). Engineering and sociology in a military aircraft project: A network analysis of technological change. *Social Problems, 35*(3), 284-297.

Lawrence, E., Newton, S., Corbitt, B., Braithwaite, R., & Parker, C. (2002). *Technology of Internet business*. Sydney: John Wiley & Sons.

Longenecker, H. E. J., Feinstein, D. L., Couger, J. D., Davis, G. G., & Gorgone, J. T. (1994). Information Systems '95: A Summary of the Collaborative IS Curriculum Specification of the Joint DPMA, ACM, AIS Task Force. *Journal of Information Systems Education, Winter 1994-95*, 174-186.

Ma, J., Zhang, Q., & Fan, Z. (2009). Web-based group decision support for R&D project outcome assessment in government funding agencies. Retrieved 2009, from http://www.is.cityu.edu.hk/Research/WorkingPapers/.

Maguire, C., Kazlauskas, E. J., & Weir, A. D. (1994). *Information services for innovative organizations*. San Diego, CA: Academic Press.

McCarthy, C. (2007). Portal features of major digital libraries. In A. Tatnall (Ed.), *Encyclopaedia of Portal Technology and Applications* (pp. 724-736). Hershey, PA: Information Science Reference.

Network Working Group. (1974). Specification of Internet transmission control program. Retrieved 2009, from http://tools.ietf.org/html/rfc675

O'Hara, B., & Petrick, A. (2005). *802.11 Handbook: A Designers Companion*. Piscataway, NJ: IEEE Standards Association.

Qiang, X. (2005). China's first Internet-organized protests. *The Wall Street Journal*. New York.

Rogers, E. M. (1995). *Diffusion of innovations* (4th ed.). New York: The Free Press.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: The Free Press.

Sahay, S. (1997). Implementation of information technology: A space-time perspective. *Organization Studies*, *18*(2), 229-260.

Sale, T. (2006). Codes and ciphers in the Second World War. Retrieved July 2006, from http://www.codesandciphers.org.uk

Singleton, V., & Michael, M. (1993). Actor-networks and ambivalence: General practitioners in the UK cervical screening programme. *Social Studies of Science, 23*(2), 227-264.

Sklar, J. (2009). *Principles of Web design*. Boston, MA: Course Technology, Cengage Learning.

Stohl, C. and Ganesh, S. (2009). Understanding relationships between the Internet and protest strategies and frames in the antiglobalisation movement. Retrieved 2009, from http://www.allacademic.com/meta/p172140_index.html

Tatnall, A. (2003). Actor-network theory as a socio-technical approach to information systems research. In S. Clarke, E. Coakes, M. G. Hunter, & A. Wenn (Eds.), *Socio-technical and human cognition elements of information systems* (pp. 266-283). Hershey, PA: Information Science Publishing.

Tatnall, A. (2005). *Web portals: From the general to the specific*. 6th International Working for E-Business (We-B) Conference, Melbourne, Victoria University.

Tatnall, A., Ed. (2007a). *Encyclopedia of portal technology and applications*. Hershey, PA: Information Science Reference.

Tatnall, A. (2007b). The ubiquitous portal. In A. Tatnall (Ed.), *Encyclopaedia of portal technology and applications* (pp. 1040-1044). Hershey, PA: Information Science Reference.

Tatnall, A. (2009a). Actor-network theory applied to information systems research. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (2nd ed.) (pp. 31-35). Hershey, PA: IGI Global.

Tatnall, A. (2009b). Gateways to portals research. *International Journal of Web Portals, 1*(1), 1-15.

Tatnall, A. (2009c). Information systems, technology adoption and innovation translation. *International Journal of Actor-Network Theory and Technological Innovation, 1*(1), 59-74.

Tatnall, A. (2009d). Web portal research issues. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (2nd ed.) (pp. 3642-3647). Hershey, PA: IGI Global.

Tatnall, A. and Davey, B. (2005). A new spider on the Web: Modelling the adoption of Web-based Training. In P. Nicholson, J. B. Thompson, M. Ruohonen, & J. Multisilta, (Eds.), *E-training practices for professional organizations*. Assinippi Park, MA: Kluwer Academic Publishers. Tatnall, A., Davey, B., Burgess, S., Davison, A., & Wenn, A. (2002). *Management information systems - Concepts, issues, tools and applications*. Melbourne: Data Publishing.

Tatnall, A., & Gilding, A. (1999). *Actor-network theory and information systems research*. 10th Australasian Conference on Information Systems (ACIS), Wellington, Victoria University of Wellington.

Theodosiou, L., & Green, J. (2003). Emerging challenges in using health information from the Internet. *Advances in Psychiatric Treatment, 9*, 387-396.

Wave Report. (2001). IEEE 802.11 Standard Tutorial. Retrieved May 2004, from http://www.wave-report.com/tutorials/ieee80211.htm.

Woo-Young, C. (2005). Online civic participation, and political empowerment: Online media and public opinion formation in Korea. *Media, Culture & Society*, *27*(6), 925-935.

Yar, M. (2005). The global 'epidemic' of movie 'piracy': Crimewave or social construction? *Media, Culture & Society*, *27*(5), 677-696.

ENDNOTE

1 Conseil Eroeen pour la Recherche Nucleaire (CERN) – an international scientific organisation based in Geneva, Switzerland.

About the Editor

Arthur Tatnall (BSc, BEd, DipCompSc, MA, PhD, FACS) is an associate professor in the Graduate School of Business at Victoria University in Melbourne (Australia). He holds bachelor's degrees in science and education, a graduate diploma in computer science, and a research MA in which he explored the origins of business computing education in Australian universities. His PhD involved a study in curriculum innovation in which he investigated the manner in which Visual Basic entered the curriculum of an Australian university. He is a member of three IFIP working groups (WG3.4, WG3.7 and WK9.7) and is also a fellow of the Australian Computer Society. His research interests include technological innovation, information technology in educational management, information systems curriculum, project management, electronic commerce, and Web portals. He has written several books relating to information systems and has published numerous book chapters, journal articles and conference papers. He recently edited *Encyclopedia of Portal Technology and Applications* for IGI Global.

Section I
Fundamental Concepts and Theories

This section serves as the foundation for this exhaustive reference tool by addressing crucial theories essential to the understanding of Web technologies. Chapters found within these pages provide an excellent framework in which to position Web technologies within the field of information science and technology. Individual contributions provide overviews of the mobile Web, semantic Web, and Web 2.0, while also exploring critical stumbling blocks of this field. Within this introductory section, the reader can learn and choose from a compendium of expert research on the elemental theories underscoring the research and application of Web technologies.

Chapter 1.1

ICT and Interculture Opportunities Offered by the Web

Laura Corazza
Università di Bologna, Italy

BACKGROUND

In Italy, as in other European countries such as Germany and France, the words "multicultural" and "intercultural" have distinct meanings[1]. In some ways the passage the one from the other indicates the evolution of a concept that, when examined in detail, forces us to re-examine educational political, and cultural choices in a society.

"In Italy, especially in the field of pedagogy, the term multicultural indicates a situation where cultures co-exist side by side but as yet, have not found a means of comparison and relating to each other . <…>. The term *intercultural*, which originated in France and then spread to other European countries, on another hand, describes a situation of interaction between different cultures, often describing a prospect, a possible point of arrival or an objective to strive for; a situation in which the people of different cultures begin to open up to reciprocal relationships, thereby bringing about the possibility of integration, characterised by a political and cultural pluralism which aims at reciprocal respect and appreciation between the respective cultures"[2]. With Multiculturalism, people accept to live with those from other cultures with tolerance but without promoting forms of exchange and democratic co-existence. The form of social organization that derives from this is that of the "melting pot" which encourages the development of ghettos or "Little Italy" and "China towns". However the Intercultural approach is reciprocal. It is for those who accept and listen, those who are tolerant, those who are not afraid of "contamination" but constantly seek to mediate between different points of view and backgrounds.

Among various other factors, it is globalization, encouraged by new communication and information technologies, that has contributed to the transformation of contemporary world-wide society into a multicultural society. These technologies have made it possible to communicate easily and at low cost with every corner of the planet. We can observe events which are taking place all over the world and take part in collective cultural processes. The Internet is made up of interconnected nodes without a centre. It encourages the growth of new trans-national communities on various levels, ranging from the political and professional to the scientific. It produced some elements of standardization, such as the use of a single language, English and has lead to uncommon social contacts on a worldwide level. At the same time however, these new communication technologies can also be viewed as a further cause of the divide between those who can access information and participate in the creation of knowledge and those who cannot.

FOCUS

The concept of freedom of access to information is an integral part of the philosophy and history of the Internet. It is also guaranteed by the characteristics of computer mediated communication and tools such as e-mail, forums, mailing lists, blogs and portals. CMC has changed the concept of communication itself. Leaving behind the one-way logic of communicator- receiver (typical of traditional mass media) the situation has become one where individuals are nodes in a network, part of an interconnected environment. The individual has the active role of social participant in communication rather than being a passive consumer (as in the case of television).

In addition to this, there are a number of new opportunities available for the user. It is now possible to influence the information circuit significantly. People can provide information as well as obtain it, they can teach as well as learn and participate not only as an individual, but as part of a group. From the moment a person becomes aware of this new communicative reality and has the possibility of taking part in it, he or she possesses new rights and also responsibilities. The rights include the opportunity to access the Web and its contents independently of limits (economic, time, and movement) and diversity (cultural, social, cognitive, or physical). To this we can add the right to learning. This must take into consideration, not only the acquisition of technical and procedural knowledge but also cognitive, emotive and social competences which are needed in order to participate with full awareness in a form of communication which is also building knowledge. There also the ever present problem of standardized positions, so the individual must be aware of personal expectations and needs, and adopt a critical, constructive attitude and be able to question, make hypotheses, choices and checks. Moreover, it is necessary to do this in a social setting which requires emotional control and an ability to collaborate, mediate and negotiate.

The communicative environment is that of the Web where the passage between the exchange of information and the building of knowledge is potentially fast, thanks to the network that allows for an exchange rather than a one-way flow of information. This environment is one of democratic co-existence based on mutual respect. It appreciates individuality and cultural and social diversity. At the same time the right to access is also accompanied by certain responsibilities towards others. Anyone who takes part in this communicative process becomes an adherent of the philosophy of Internet. As a direct consequence of its genesis and evolution, it appears to be more closely associated with the technology of freedom rather than that of control, and with democratic relationships rather than ideological imperialism.

There are various types of free information on the web today. They can be placed into two categories: new forms of news agencies and new forms of publishing. Both fit into to the intercultural and democratic view which requires free information, originating “from below” that is, from all the links in the network, combined with fast transmission speeds and a high level of diffusion. Examples of the first category are blogs and more organised agencies such as Misna, an international press agency run by congregational missionaries in Africa, Asia, Latin America and Oceania. Though possessing few resources it can boast many successes and has earned “recognition after only a brief period of activity including the award of the ‘St Vincent 2002’ journalism prize”[3].

The archives of scientific literature (papers, journal articles, addresses made at conferences) are another important example of freely available information on the Net. Offered in an electronic format at no cost, by the very universities that finance the research and believe in the principle of free access to information, they are potentially useful to everyone from college students in America to the doctor in a hospital in Tanzania

Thus we can state the Internet’s contribution to interculturalism has manifested itself in three ways: the creation of a cyber-culture; in a new concept of knowledge; and in the participation “from below” in the building of learning:

Internet as a place of communicative interaction and universal culture.

For Lévy, what is interesting about cyberculture is the coming together of all differences and heterogeneity. He believes in the ability of cyberspace to bring out the best of human intelligence in every person through communicative relationships which create a collective intelligence and a cyber-culture. In this sense we can assert that cyber-culture is both a transculture and an inteculture, since it is universal without being oppressively uniform.

If we take cyberspace to be the *location*, then communication is the *means* that permits the creation of this collective intelligence and cyber-culture. Computer users can communicate with each other in a novel way on the Internet since they can be more than mere passive users and isolated consumers like today’s television viewers. Communication is not limited to “one to one” exchange. In cyber culture it is reciprocal, interactive and communitarian. Moreover it is universal as anyone can be an active communicator. The expansion of interconnections displays the fact that there is one general humanity. This is what Levy means by “universal”: the idea of universality in the Enlightenment sense of the word, where our goal is the unity of human kind and the affirmation of universal principles such as the rights of Man. All human beings can come into reciprocal contact virtually and become conscious collectively of their existence. The more we add links, the more diverse and heterogenic material is circulated on the net. Humanity is becoming aware of itself (universality) but this awareness has no one meaning. It does not pass through one central point and is not bound by any set of laws unlike science where, for example, universal gravitation is the same everywhere. What is interesting in cyber-culture is the bringing together of all differences and heterogeneity

Internet as a location for new knowledge which is open, complex, multidisciplined, individual, and collective at the same time.

Morin (2000) has highlighted the problem of the inadequacy of knowledge (which is *divided* by the boundaries between disciplines while the reality of the world is increasingly global and interconnected) and the challenge of complexity. Complexity is a method and a form of knowledge which requires a dialogical approach. Gregory Bateson also spent his life demonstrating the interdependence between elements and the interconnections between different worlds and disciplines. He theorises about the ecology of ideas, an ecosystem in which there is a plurality of levels of application of ideas, using a systematic approach which, from the point of view of learning means giving importance to contexts, relationships and functions.

So the new idea of knowledge, does not only refer to concepts that are to be *transmitted* but, above all, to the itineraries and the network of individual and collective experiences which are in a state of perpetual change.

This is the challenge that the individual can take up, not withstanding the risks of cognitive standardisation on one hand, and fragmentation or individualism on the other. Morin's "tête bien faite" requires the full use of intelligence in the organisation of ideas, in the search for links, routes and new itineraries, knowing full well that it is not possible to attain totality but that we must put into *practice* collective and individual cognitive *practices* at the same time because they exploit the power of connectivity. "If a document is put on the World Wide Web you are doing two things at the same time: firstly, you are increasing the amount of information that is available, but secondly, you are doing another thing with the links between your document and the others: you offer the surfer who finds your document your point of view. So you are not merely offering information but a point of view in a collection of information. The World Wide Web is not only an enormous mass of information, it is the expression of thousands of different points of view. It should be viewed from this aspect" (Levy, 1995). There is space for every diversity and every point of view on the Web which doesn't have to become individualism, if you allow yourself to become enchanted by hypertextual logic, by the links and maps and if you allow yourself to feel part of a whole, a collective intelligence and a universal culture.

Internet as a place for forming an opinion, expressing a critical judgement, and participating in the building of knowledge.

A surprising example of the impact of the Internet on the power of the individual and thus of the group, is the increase in public access to health information. This is not just another step in the spread of knowledge, but a political change which introduces a new balance of power. More and more patients arrive at their doctor's with information they have found on the Web. It may be incomplete or taken from dubious sites, but they begin their treatment in a more knowledgeable and participative way. Even though there are limits and the information is certainly not enough to provide specialist knowledge, it is enough to modify the power/knowledge rapport in the doctor/patient relationship.

The potential for forming opinions, and thereby expressing judgements and making demands aided by the Web, is becoming more apparent in the political development of citizens. People can participate in public debates as they once did in the classical agora. They can inform themselves and take up positions through blogs, messaging systems and online journalism.

If we say that technology is not neutral and the choice of technology produces profound effects, the most evident effect of the Internet is the transition from vertical to horizontal communication. The first has an undeniably authoritarian nature, even if the intentions of those using it are democratic, since it requires a silent, passive viewing. The second however, permits active participation, and while this alone is not enough to confer unquestionably democratic credentials, it certainly transforms the quality of communication. In general, the passage from one-way communication to interactivity does not automatically guarantee a growth in democracy. On the contrary it can increase the use of extorted consensus to legitimise solutions taken without the effective participation of citizens.

The key question is thus the following: Can the Internet be used to manipulate consensus? Is there a social use of new technologies? Up to now citizens have been subjected exclusively to vertical information offered by television, which creates passivity in viewer. The condition of citizens using the web, benefiting from horizontal communication which blurs the distinction between producers and consumers and the supremacy of the former over the later, is totally different. Everyone becomes a supplier of information, and as such, an active player in communication. The most marginalized of Internet users can have word power equal to that of a government, especially when the news they wish to communicate is information that others want to suppress at any cost. Censorship becomes much more difficult, but with one condition: that there is a real mass internet information literacy program and a true universal service. Without this there is the risk that an information apartheid emerges. Moreover, there is also the danger of *hyper-information,* where citizens are unable to pick out the important information and check its validity due the potentially unlimited quantity of information available. It is for this reason too that the role of education in permitting a social use of new technologies is increasing. It has an important part to play in training active, knowledgeable citizens, to guarentee an adequate command of the tools necessary for developing the competences required by society.

According to Rifkin, cultural production is the first level where the economic life of a country is played out. In consequence of this, there is a strong move to control such production in the direction of a standardised predetermined models (Rifkin, 2000).

But if it is true that the Internet is disliked by governments because it cannot be censored or controlled easily, it can also become a tool in the democratic education of citizens, if the move towards liberty prevails over that towards standardisation. A conscious citizen is a person who participates in the building of a free, universal society which is constantly developing and changing. The process of building knowledge that can come about through the use of the network of ideas and information available on the Internet produces social progress. This is what Dewey calls "social efficiency": everything that makes an experience valid for the community, as well as oneself; everything that makes it communicable and useful in the demolition of social barriers. The same efficiency that he considers to be the final aim of education in a democratic society.

CURRENT LIMITS AND FUTURE DEVELOPMENTS

Technology of Freedom or Technology of Control?

Whether we consider technology to be neutral, or something that can structure the user's activities, we cannot avoid reflecting on the responsibilities of those who use it and the need for user-education. Whilst there is a great potential for helping humanity, we face a number of new questions. A critical analysis can reveal both positive and negative aspects, opportunities and risks, and a possible direction to take in order to achieve a democratic use for the civic education of the citizen. The key question is that posed by Stefano Rodotà: Are we using the technology of freedom or the technology of control? Can the use of technology be free from rules? Is it necessary to reflect on which direction we should be striving for?

We need to free ourselves from some vices of form and clichés that linger, such as technological arrogance (and the idea that technology does not need rules), the optimism of the market (which in reality favours economic interests over social equality) and the political-ideological simplification which sees technology as a cure for all evils (in reality technology is not neutral but requires careful use)

The Digital Divide

Regarding access to the new communication technologies and the internet in particular, there are no equal opportunities between the technologically advanced rich countries and countries which are technologically behind. In a message of 2002, the secretary general of the United Nations Kofi Annan underlined the importance of communication technologies and exhorted poor countries on the African continent to unite to create a digital revolution which had become indispensable. In collaboration with the ITU International telecommunications union the UN organised the first World Summit on the Information Society, WSIS with the aim of building a common vision of the information society and adopt a plan of action to bring it about[4].

The first principle of the Geneva declaration states:

We, the representatives of the peoples of the world, assembled in Geneva from 10-12 December 2003 for the first phase of the World Summit on the Information Society, declare our common desire and commitment to build a people-centred, inclusive and development-oriented Information Society, where everyone can create, access, utilize and share information and knowledge, enabling individuals, communities and peoples to achieve their full potential in promoting their sustainable development and improving their quality of life, premised on the purposes and principles of the Charter of the United Nations and respecting fully and upholding the Universal Declaration of Human Rights.

It is not enough to be connected to resolve the fundamental problems of underdevelopment and to ensure that the Information Society becomes a vehicle for democracy, justice, equality and respect for individuals and their personal and social development. Beyond the mere physical availability of a computer, other factors such as economic/social resources and levels of literacy, influence the successful use technology. It is the development of the necessary competences and a fruitful, informed use of the internet that is the challenge for the current Information Society and lifelong learning in particular.

CONCLUSION

In conclusion Internet is a wonderful opportunity (comparable with the impact on culture of the invention of the printing press) not only when we consider the technical changes in the exchange of information and the spread of knowledge, but also for the political, social and educational implications. In substance we have to start to conceive and use new democratic environments for the production of culture, new forms of diffusion which offer efficient tools for communication between individuals, disciplines and political and social points of view. To use a more traditional pedagogical language, what emerges from the analysis of the spontaneous (and in some cases experimental) use of the internet is the necessity to use the great potential for communicative and didactic renewal which is beginning to manifest itself, in order to produce a culture that overcomes individualism and is oriented towards forms of collaboration which widen the opportunities for democratic participation in the information/knowledge process. By moving in this direction perhaps we can cement the relationships between individuals, groups, states, and political organisations providing a glimpse of the added value of cooperation, reciprocity, peace, and thus interculture.

REFERENCES

Castells, M. (2002a). *La nascita della società in rete.* Milano: Università Bocconi.

Castells, M. (2002b). *Galassia Internet.* Milano: Feltrinelli.

Castells, M. (2002c). *Rivoluzione tecnologica e identità,* intervista a “Mediamente”, Rai Educational, URL: http://www.mediamente.rai.it/articoli/20020530b.asp.

Corte, M. (2006). *Comunicazione e giornalismo interculturale. Pedagogia e ruolo dei mass media in una società pluralistica.* Padova, Cedam.

Dewey, J. (1954). *Il mio credo pedagogico. Antologia di scritti sull’educazione,* trad. it. a cura di Lamberto Borghi, Firenze, La Nuova Italia.

Dewey, J. (1967). *Scuola e società*, Firenze, La Nuova Italia. *The school and the society*. Chicago: The University of Chicago Press.

Dewey, J. (1992). *Democrazia e educazione*, Firenze, La Nuova Italia. *Democracy and education*: *An introduction to the philosophy of education*. New York: Macmillan.

Genovese, A. (2003). *Per una pedagogia interculturale. Dalla stereotipia dei pregiudizi all'impegno dell'incontro*, Bologna, BUP.

Harnad, S. (2003). For whom the gate tolls? In D. Law & J. Andrews (Eds.), *Digital libraries: Policy planning and practice*. Ashgate.

Lévy, P. (1995a). *L'intelligenza collettiva*, intervista a "Mediamente", Parigi - European IT Forum, 4 settembre, URL: http://www.mediamente.rai.it/HOME/bibliote/intervis/l/levy.htm

Lévy, P. (1996). *L'intelligenza collettiva. Per un'antropologia del cyberspazio*. Milano, Feltrinelli.

Lévy, P. (1999). *Cybercultura. Gli usi sociali delle nuove tecnologie*. Milano, Feltrinelli.

Misna, (Missionary International Service News Agency) provides daily news from for and about the South of the World.

Morin, E. (1985). *Le vie della complessità*, in G. Bocchi e M. Ceruti (a cura di), *La sfida della complessità*, Milano, Feltrinelli.

Morin, E. (2000). *La testa ben fatta. Riforma dell'insegnamento e riforma del pensiero*, Milano, Cortina.

Open Archives Iniziative. http://www.openarchives.org

Pinto Minerva, F. (2002). *L'intercultura*, Roma-Bari, Laterza.

Rodotà, S. (2001). *Tecnologia e democrazia*, intervista a: "Enciclopedia multimediale delle scienze filosofiche. Il Grillo", prodotto da Rai Educational, puntata del 08/01/2001, URL: http://www.emsf.rai.it/grillo/trasmissioni.asp?d=742

Rodotà, S. (2004). *Tecnopolitica* (2d ed.), Roma, Laterza.

Rodotà, S. (2006). *Sviluppo telematico e democrazia*, intervista a:"MediaMente", Roma, 13 gennaio, URL: http://www.mediamente.rai.it/home/bibliote/intervis/r/rodota02.htm#link001.

Web sites **Misna,**www.misna.org

Willinsky, J. (2003). Democracy and education: The missing link may be ours. *Harvard Educational Review*, *72*(3), 40.

Willinsky, J. (2003). The nine flavours of open access scholarly publishing. *Journal of Postgraduate Medicine*, *49*(3), 263–267.

Willinsky, J. (2005). Scientific research in a democratic culture: Or what's a social science for? *Teachers College Record*, *107*(1), 38–51. doi:10.1111/j.1467-9620.2005.00455.x

KEY TERMS

Cyberculture: For Lévy, what is interesting about cyberculture is the coming together all differences and heterogeneity. He believes in the ability of cyberspace to bring out the best of human intelligence in every person through communicative relationships which create a collective intelligence and a cyberculture. In this sense we can assert that cyberculture is both a trans-culture and an interculture, since it is universal without being oppressively uniform.

Democratic Education: Morin has highlighted the problem of the inadequacy of knowledge (which is *divided* by the boundaries between disciplines while the reality of the world is increasingly global and interconnected) and the challenge of complexity. Complexity is a method and a form of knowledge which requires a dialogical approach. Gregory Bateson also spent his life demonstrating the interdependence between elements and the interconnections between different worlds and disciplines. So the new idea of knowledge, does not only refer to concepts that are to be *transmitted* but, above all, to the itineraries and the network of individual and collective experiences which are in a state of perpetual change. The process of building knowledge that can come about through the use of the network of ideas and information available on the internet produces social progress. This is what Dewey calls "social efficiency": everything that makes an experience valid for the community, as well as oneself; everything that makes it communicable and useful in the demolition of social barriers. The same efficiency that he considers to be the final aim of education in a democratic society.

Democratic Information: The concept of freedom of access to information is an integral part of the philosophy and history of the internet. It is also guaranteed by the characteristics of computer mediated communication and tools such as e-mail, forums, mailing lists, blogs and portals. CMC has changed the concept of communication itself. Leaving behind the one-way logic of communicator-receiver (typical of traditional mass media) the situation has become one where individuals are nodes in a network, part of an interconnected environment. The individual has the active role of social participant in communication rather than being a passive consumer (as in the case of television).

Information Society: It is now possible to influence the information circuit significantly. People can provide information as well as obtain it, they can teach as well as learn and participate not only as an individual but as part of a group. From the moment a person becomes aware of this new communicative reality and has the possibility of taking part in it, he or she possesses new rights and also responsibilities. The rights include the opportunity to access the web and its contents independently of limits (economic, time, and movement) and diversity (cultural, social, cognitive or physical). To this we can add the right to learning. This must take into consideration, not only the acquisition of technical and procedural knowledge but also cognitive, emotive and social competences which are needed in order to participate with full awareness in a form of communication which is also building knowledge.

Intercultural Education: In Italy, as in other European countries such as Germany and France, the words multicultural and intercultural have distinct meanings. In Italy, especially in the field of pedagogy, the term multicultural indicates a situation where cultures co-exist side by side but as yet, have not found a means of comparison and relating to each other. The term *intercultural*, which originated in France and then spread to other European countries, on the other hand, describes a situation of interaction between different cultures, often describing a prospect, a possible point of arrival or an objective to strive for; a situation in which the people of different cultures begin to open up to reciprocal relationships, thereby bringing about the possibility of integration, characterised by a political and cultural pluralism which aims at reciprocal respect and appreciation between the respective cultures.

ENDNOTES

[1] In North American and the majority of English speaking culture however the term *multicultural* is most widely used (especially in education). This can be observed when consulting ERIC, the bibliography data base of the U.S. Department of Education which specializes in topics relating to pedagogy and didactics and representative of North American scientific literature. In its thesaurus *intercultural education* is only considered as a synonym for *multicultural education*. The only exception is *intercultural communication*.

[2] Genovese A. (2003). *Per una pedagogia interculturale. Dalla stereotipia dei pregiudizi all'impegno dell'incontro,* Bologna, BUP, p. 181.

[3] http://www.emi.it. Description of the agency MISNA, under the heading of "cards, profiles," where they are actually exposed the various associations that work with MISNA.

[4] http://www.itu.int/wsis/

WEB SITES

The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The Open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication. Continued support of this work remains a cornerstone of the Open Archives program. The fundamental technological framework and standards that are developing to support this work are, however, independent of the both the type of content offered and the economic mechanisms surrounding that content, and promise to have much broader relevance in opening up access to a range of digital materials. As a result, the Open Archives Initiative is currently an organization and an effort explicitly in transition, and is committed to exploring and enabling this new and broader range of applications. As we gain greater knowledge of the scope of applicability of the underlying technology and standards being developed, and begin to understand the structure and culture of the various adopter communities, we expect that we will have to make continued evolutionary changes to both the mission and organization of the Open Archives Initiative.

Public Library of Science,http://www.plos.org/

PLoS, a nonprofit organization of scientists and physicians committed to making the world's scientific and medical literature a freely available public resource. As of 2006 it publishes *PLoS Biology*, *PLoS Medicine*, *PLoS Computational Biology*, *PLoS Genetics* and *PLoS Pathogens*. PLoS ONE is a new journal to be launched soon.

Our goals are (1) to open the doors to the world's library of scientific knowledge by giving any scientist, physician, patient, or student—anywhere in the world—unlimited access to the latest scientific research. (2) Facilitate research, informed medical practice, and education by making it possible to freely search the full text of every published article to locate specific ideas, methods experimental results, and observations. (3) Enable scientists, librarians, publishers, and entrepreneurs to develop innovative ways to explore and use the world's treasury of scientific ideas and discoveries.

PubMed Central (PMC), http://www.pubmedcentral.nih.gov/

PubMed is a free search engine offering access to the MEDLINE database of citations and abstracts of biomedical research articles. It is offered by the United States National Library of Medicine as part of the Entrez information retrieval system. MEDLINE covers over 4,800 journals published in the United States and more than 70 other countries primarily from 1966 to the present.

Ricerche di Pedagogia e Didattica,http://rpd.cib.unibo.it

Freely accessible journal of the university of Bologna on pedagogical and didactic topics

Unione Europea, http://europa.eu/index_it.htm

Portal: "Recommendations of the European Parliament and Council regarding key competences for lifelong learning", 11/20/2005.

This work was previously published in Encyclopedia of Information Communication Technology, edited by A. Cartelli; M. Palma, pp. 357-364, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 1.2
Mobile Social Web:
Opportunities and Drawbacks

Thorsten Caus
Georg August University of Göttingen, Germany

Stefan Christmann
Georg August University of Göttingen, Germany

Svenja Hagenhoff
Georg August University of Göttingen, Germany

ABSTRACT

As mobile Internet usage continues to grow, the phenomenon of accessing online communities through mobile devices draws researchers' attention. Statistics show that close to 60 percent of all mobile Internet traffic worldwide is related to the use of mobile social networks. In this chapter, the mobile social Web is defined, categories of mobile communities explained, and success factors and drawbacks discussed from the technical, social, and economic perspectives. Challenges, including low transmission rates, changes in usage patterns, search for new revenue sources, as well as the need for development of original mobile Web content and applications are addressed. The technical requirements for the mobile use of online communities are identified. The chapter closes with a summary of potential economic and social prospects of the emerging mobile social Web.

DOI: 10.4018/978-1-60566-368-5.ch002

INTRODUCTION

Until recently, the Internet was a domain restricted to stationary computers, but nowadays it can also be accessed through mobile devices equipped with web browsing capabilities. Now it is not only possible to surf the web using wireless access and mobile devices, but there is also a growing number of mobile Internet applications and services. Increasingly, mobile social networking applications have been made available to a large number of mobile phone users.

Internet users have accepted online communities and internalized the concept of the Social Web also referred to as Web 2.0 (Koesch, Magdanz, & Stadler, 2007). Private as well as business users have become familiar with various online communities (Patrzek, 2007; von Tetzchner, 2008). On the one hand, mobile social networks are becoming more widespread because of the increasing dissemination of new wireless communication technologies (Heng, 2006, p. 5). On the other hand, a large number of devices are designed to implement new communications technologies, for example, the Universal Mobile Telecommunications System (UMTS) in Europe (Heng, 2006, p. 1).

Studies conducted by Opera Software, an Internet software and services company, demonstrate that 40 percent of all mobile Internet traffic worldwide is related to the use of online communities. In some countries the share is as high as 60 percent: for example, in the United States, South Africa, and Indonesia (von Tetzchner, 2008). Research into the various ways of using the Social Web in a mobile context is now of paramount importance. In this chapter, *mobile social web* is defined, categories of mobile online communities and their success factors explained, and selected opportunities and drawbacks of the mobile online communities discussed from a technical, social, and economic perspectives.

BACKGROUND

The Social Web can be viewed as a concept and a platform that utilizes social software (e.g., forums, wikis, blogs, etc.) to fulfill or support some of the important human needs, such as: self-realization, acceptance, social connectedness, and safety (Maslow, 1943, p. 372-383). The purpose of the Social Web is to support human communication and facilitate social contact. The Social Web encompasses numerous Internet applications, such as social networking sites, massively multiplayer online role-playing games, photo and video sharing, online stores and auction houses, virtual worlds, and wiki collaborations. The most popular and widespread actualizations are online communities (e.g., MySpace, Facebook, StudiVZ or XING). The term "Social Web" is often used in everyday language as well as in scholarly literature as a synonym for "virtual" and "online communities" (Hummel, 2005, p. 5), although these terms do not differ greatly (Fremuth & Tasch, 2002, pp. 5-6).

In the past years many academic disciplines have dealt with the Social Web. Various attempts to provide a definition have resulted in three different approaches: technical, social, and economic. The technical approach focuses on the Internet as a medium or platform for a community. The sociological point of view stresses the forming and functioning of communities, whereas the economic perspective examines potential gains and intended profits (Hummel, 2005, p. 8-11).

These three perspectives have led to a variety of definitions of online communities with differing points of emphasis. A detailed overview of common definitions is given by Fremuth and Tasch (2002), Hummel (2005) and Markus (2002). In identifying an *online community* one perspective emphasizes that it is formed by a group of people, while another stresses its web platform. The definition used in this chapter combines both approaches, for an *online community* is seen as a social group that interacts through a web platform over an extended period of time.

An online community can be characterized by four elements (Gebert & von Rosenstiel, 1992, p. 122-123; Hamman, 2000, p. 225):

- group of people with shared objectives (e.g., interests, goals)
- interaction over an extended period of time
- closeness due to bonds and relationships
- shared space for interactions governed by certain rules (for example, role definitions).

Without shared objectives there would be no interaction and relationship and, subsequently, no community at all (Markus, 2002, p. 36). Interactions within the community are seen as topic-oriented communication as well as the execution of actions (Kim, 2000, p. 5). Both can take place independently of time and location (Winkler & Mandl, 2004, p. 14). The process of founding and maintaining such online communities usually takes place on the Internet (Eigner & Nausner, 2003, p. 58). The second defining characteristic is the web platform, which can be seen as an Internet communication system which acts as an intermediary. It enables and facilitates meetings, the maintenance of the community, and its interaction with other people (Reichwald, Fremuth, & Ney, 2002, p. 8).

There are different ways of categorizing online communities (Brunold, Merz, & Wagner, 2000, p. 30-37; Fremuth & Tasch, 2002, p. 21; Hummel, 2005, p. 46). A reasonable approach is to categorize them according to similarities, for they play a major role in online communities. Therefore, an online community can be geographic (bound to an area), demographic (classification according to nationality, age, gender), or based on shared interests or activities (Kim, 2000, p. 5).

Online communities can be viewed as social systems. Relationships and interactions can only develop once a web platform has been established, which makes it difficult to start a community (Leitner, 2003, p. 36). The network-effect character of online communities shows this very clearly. There will only be accelerated growth once a critical amount of relationships and interactions between users has been achieved. This is due to the fact that users do not benefit before this point is reached (Reichwald, Fremuth, & Ney, 2002, p. 8). Even though this development is difficult to predict, an operator is able to influence the development of an online community by making it more attractive (Reichwald, Fremuth, & Ney, 2002, p. 9-10). The success factors that have been identified are listed in Table 1.

THE MOBILE SOCIAL WEB

The mobile use of online communities can be referred to as the *mobile social web*. In this context, *mobility* is understood as the unrestricted transfer of text, voice or data independent of user's physical location (Kurose & Ross, 2005, pp. 536-538). Therefore, mobile online communities are free of interruption caused by the movement of the user from one location to another. The devices employed can be either wireless or wired (Kurose & Ross, 2005, p. 504). Suitable types of devices are mobile phones, smart phones (Michelsen & Schaale, 2002, p. 51) and personal digital assistants, as these can always stay switched on and do not need to be booted. Although mobility does not necessarily require wireless connections to these devices, wireless mobile networks are used most commonly (Gerum, Sjurts, & Stieglitz, 2003, p. 145). European standards, for instance, include the Global System for Mobile Communication (GSM) extensions, Enhanced Data Rates for GSM Evolution (EDGE) standard (Herzig, 2001, p. 399), General Radio Packet Service (GPRS) protocol (Stader, 2001, p. 37), and the 3rd generation Universal Mobile Telecommunications System (UMTS) technology (Kurose & Ross, 2005, pp. 534-535). The mobile social web involves opportunities as well as drawbacks, as will be discussed below.

Opportunities of the Mobile Social Web

The features of mobile communities generally correspond to those of traditional online communities. They are enhanced by a new way of accessing the community-web-platform through mobile, wireless devices. On closer examination,

Table 1. Success factors on online communities (adopted from Koch, Groh, & Hillebrand, 2002; Leitner, 2003)

Users' point of view	Corporations' point of view
• Advantages of usage, for example in the form of problem solving or entertainment • Simple compilation of contributions • Easy technical access, usage and adequate stability • Equality, credibility and trust • Non-commercial orientation	• Personal network and personal characteristics of entrepreneurial team • Product or service idea in business model • Available resources and capabilities • Marketing strategy with viral emphasis • Potential for speedy marketing

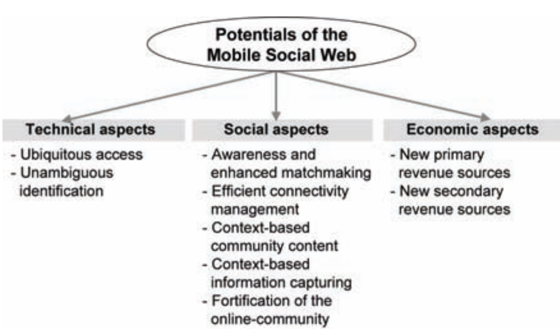
these communities do not only seem to benefit from mobile access, but also from additional potentials resulting from mobility and localization possibilities. Furthermore the question arises, whether this newly opened potential encompasses not only technical and social aspects but economic aspects as well (see Figure 1).

Technical Aspects

Mobile wireless devices facilitate ubiquitous access to online communities (Koch, Groh, & Hillebrand, 2002, p. 2). The user is able to gain access anywhere and any time, while being liberated from the world of stationary Internet and permanently installed devices (Reichwald, Fremuth & Ney, 2002, p. 6). Furthermore, users benefit from being reachable at all times by being armed with mobile wireless devices (Wiedmann, Buckler, & Buxel, 2000, p. 86). Therefore, the users can stay online and be accessible continuously and without interruption. No time is lost: for example, by turning the devices on or logging in. Thus, interaction becomes more spontaneous and expressive.

Mobile wireless devices make it possible to identify the user via his PIN and SIM card, both of which are requirements of mobile communications (Tasch & Brakel, 2004, p. 4). This explicit possibility of identification can be used to identify the user in an online community. The user's mobile device then serves as a membership card or an individual entrance ticket (Reichwald, Fremuth, & Ney, 2002, p. 7). Identifying the user can be automated, making it more reliable. It is also more authentic and results in more confidence within the community (Hummel, 2005, p. 72).

Figure 1. Overview of the mobile social web potentials



Social Aspects

A local context can be defined by identifying a user's current whereabouts (Koch, Groh, & Hillebrand, 2002, p. 3). A user's whereabouts could be presented to other users through geomapping or textually. Furthermore, an awareness service could be employed in addition to existing information services, such as "buddy lists" (Tasch & Brakel, 2004, p. 7). Contacts and friends in the vicinity of the user can be shown on the mobile device. Mobile online communities could thus improve interactions between community members and extend social ties. For example, people with similar interests can get together spontaneously. Contexts can be used to find out about the accessibility of mobile community users (Groh, 2003, p. 9). Depending on the user's current whereabouts a service can provide information about if and how a person may be contacted and the user can decide how to do this. The contact information can simply rely on such area aspects as availability of UMTS. Personal preferences can also be used for this purpose.

Another potential feature is to filter content according to current contexts (Groh, 2003, p. 9). A personal information service can select information about places of interest in the vicinity of the user's current location and report it to the user (context specific presentation of information). Moreover, context specific capturing of content is conceivable (Groh, 2003, p. 8). Metadata (e.g., location data or location names) and information and news services form the basis for this idea. For example, an entry reviewing the quality of food and drinks in a local café could be generated automatically by entering the name of the café or its location.

Personal meetings can be organized more easily as the awareness of a person's location increases, matching the users' interests is extended, and the management of accessibility becomes more efficient. Stronger relationships between persons will typically result when they are geographically close and have personal encounters (Larsen, Urry, & Axhausen, 2006, pp. 12-13). These aspects are not available in stationary online communities because they depend on time and location. Face-to-face communication can lead to an improved quality of relationships because interactions become closer and more intense. All of the above will result in the strengthening of a community (Schneider, 2003, p. 99).

Economic Aspects

Mobile communities allow an opportunity for spontaneous, affective, and meaningful community activities, which may result in closer relationships between community members than in the case of traditional online communities. Data input and output can be accomplished more easily through context specific services (Diekmann et al., 2006). Therefore, community mobile operators assume that users will be more willing to pay for participation in mobile communities than in stationary online communities (Reichwald, Fremuth, & Ney, 2002, p. 12). The following list provides details of primary and secondary revenue sources for mobile operators (Reichwald, Fremuth, & Ney, 2002).

Primary revenue sources arise from operating a mobile online community (Reichwald, Fremuth, & Ney, 2002, p. 11). Three possibilities can be identified:

- *To levy usage fees:* Users pay for the usage of mobile information and interactivity services and for community content. The fees depend on the usage of the services or they are usage-independent (Reichwald et al., 2002, pp. 24-25). Users pay either for the usage of each single service or for using services during a certain period of time. Both approaches have proved unsuccessful in the case of non-mobile online communities (Reichwald et al., 2002, pp. 23-24). It may be expected that the advantages of mobile communities and their improved attractiveness are great enough to generate an increased willingness to pay (Reichwald, Fremuth, & Ney, 2002, p. 12). Moreover, users are generally accustomed to paying for mobile services.
- *Advertising and sponsoring:* Even now, non-mobile online communities allow the generation of advertisements tailored for special target groups by analyzing user profiles. In comparison to mass advertisements, this approach involves less wastefulness (Schubert & Ginsburg, 2000, pp. 51-52). For example, depending on the user's current location, advertisements can be provided. Companies willing to sponsor certain activities are another potential source of revenue (Reichwald et al., 2002, p. 20).
- *Data transmission fees:* Fees for using telecommunication (TC) services have to be paid to the TC companies. The operator of a mobile online community could negotiate a share of the revenues. For this revenue sharing purpose, cooperation between the TC company and the community operator is needed, and accounting models have to be developed.

The users themselves can become the potential generators of possibilities to earn money for the community. Their content contributions can be used for market research as well as ideas about new services and products. Documented behavior, usage patterns, and preferences can be used to filter properties of target groups. The main potential for community operators has to be made up by selling the gathered knowledge as consulting services or by using it for themselves (Reichwald, Fremuth, & Ney 2002, p. 12). Mobile communities can facilitate activities at a higher level of utility and quality, as data are potentially more differentiated and reliable compared to stationary activities.

Drawbacks of the Mobile Social Web

Mobile online communities have drawbacks, as summarized in Figure 2.

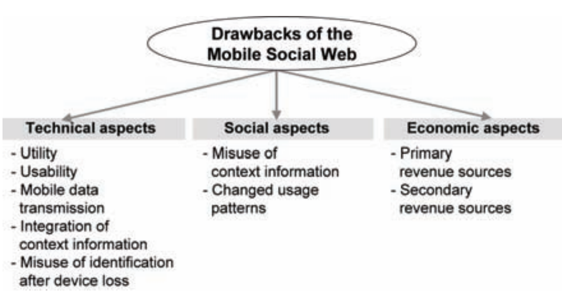
Technical Aspects

The utility of a service is a success factor as well as a challenge. It has to be decided which services should be accessible in a mobile community to make it attractive (Yom, 2002, p. 177). Not only information and interaction services have to be selected, content needs to be chosen, too. Restricted possibilities of data presentation and device handling as well as limited capacities of mobile devices affect the utility.

Other factors influencing usability are data transmission costs, low data transmission rates, device handling problems, and inconvenient data input and output possibilities (Bliemel & Fassott, 2002, p. 14). The fees of mobile network operators are mostly based on data volume. Therefore, a community user will have to pay fees depending on the intensity of his or her usage.

Low data transmission rates are often a problem for mobile device users. Established GSM services, such as Short Message Service (SMS), provide an adequate transmission rate; however, there are only limited possibilities for data presentation. Consequently, it makes more sense to opt for 3rd generation technologies like GPRS and UMTS, especially as the increasing dissemination of these technologies renders this alternative increasingly feasible.

Figure 2. Overview of the drawbacks of the mobile social web



An appropriate localization technology must be integrated in order to realize enhanced mobile services and provide new functionalities of mobile online communities, such as contextual community content, context-oriented information collection, or efficient connectivity management. The localization of end devices connected to cellular networks via their radio cell is currently still expensive. However, free and precise localization using Global Positioning System (GPS) is not yet feasible, as the vast majority of users do not have GPS-compatible end devices.

The possibility of unambiguous identification poses another problem. In case a mobile device gets lost, unauthorized persons may pretend to be the original user, as mobile devices usually have only a very low degree of security (Reisinger, 2007). This is especially a problem when users believe in the trustworthiness of the identification using mobile end devices.

Social Aspects

The electronic capturing and processing of the user context as well as the opening of communities for mobile usage can lead to negative consequences. Publishing the actual location of a user means an intrusion into his or her privacy and a limitation to intimacy. Being spied upon undiscovered could be the result of using contextual services. Parents could use these services to locate their children. Partners in a relationship could use it to track each other. The risks involved may lead to a fear of misuse or limited intimacy, resulting in the merely conditional use of contextual services. In stationary online communities, we observe that little emphasis is put on intimacy - users publish a multitude of personal data (Gross & Acquisti, 2005, p. 4-8). Hence, it remains open whether the added value of contextual services or the fear of misuse and limited intimacy will prevail.

At present mobile radio services, including SMS or telephony, are preferably used for contacting persons one already knows (Tasch & Brakel, 2004, p. 4). Stationary online communities are commonly used to create new relationships: i.e. contacting persons hitherto unknown to the user (Fremuth & Tasch, 2002, p. 24). The projection of mobile radio usage patterns onto stationary online communities could lead to the change from a preferably theme-oriented usage towards a person-oriented or communication-oriented usage of mobile online communities (Reichwald et al., 2002, p. 13). Mobile community activities would then be realized preferentially with already known persons. Such trends could jeopardize the establishment of theme-orientated mobile online communities.

Economic Aspects

A commercial design of mobile online communities has to identify primary and secondary revenue sources for their operators. Commercial intentions should be declared and openly communicated to the community (Leitner, 2003, pp. 43-44). This is the only way to grant trust and authenticity from the very beginning. Using primary revenue sources involves the following challenges:

- **Collection of usage fees:** So far, the stationary Internet largely provides free content and services (Reichwald, Fremuth, & Ney, 2002, p. 11). Internet users are accustomed to free services. There is a possibility that this factor will have a negative impact on the willingness to pay for mobile services. Users already pay for mobile data services such as SMS or mobile Internet access (Reichwald et al., 2002, p. 27). Fees for the use of services or content would increase these costs.
- **Advertising and sponsoring:** Advertising in online communities was often frowned upon in the past (Leitner, 2003, pp. 41-42). Even now, although an increased number of advertising banners are placed, the acceptance of advertising does not seem to be self-evident. Moreover, advertising messages on mobile devices are still fairly uncommon. It is unclear whether a satisfactory advertising effect can be achieved by mobile advertising in online communities. It is assumed that the perception duration of advertisement tends to drop with mobile usage compared to the stationary Internet (Heinonen & Stransdvik, 2007, p. 610; Reichwald et al., 2002, p. 22). Even when advertisements are noticed, little space for advertising messages is available due to the small displays of mobile devices (Michelsen & Schaale, 2002, p. 20).
- **Mobile data transmission fees:** As yet volume-billing models for the use of the mobile Internet are widespread; this could be a restriction to the time-consuming use of a mobile online community (Reichwald et al., 2002, p. 28). This can lead to lower data volumes, resulting in fewer payments to mobile radio operators. As a consequence, these operators would be less willing to forward payments to a community operator. It remains to be seen how this revenue potential develops through billing models for mobile Internet usage and cooperation.

There are also new challenges for the secondary revenue sources: Using a mobile online community as an instrument for market research can be profitable for an operator, but it seems reasonable that the users need to know that this is drafted on the basis of content and the analysis of their usage habits (Leitner, 2003, p. 43-44). This could especially affect success factors of an online community like equality, credibility and trust. Sharing part of the revenues with users could be a solution.

FUTURE TRENDS

A growing number of Internet services make their applications available to mobile users. This can be attributed to the increasing proliferation of mobile broadband Internet access (especially UMTS in Europe). Three variants of mobile implementation are available. Users can browse profile pages and photo albums via WAP and add new images and texts by means of SMS and Multimedia Message Service (MMS). The ShoZu service,[1] in contrast, performs as an integration platform and allows mobile users to upload and download content at multiple online communities with a single message. One Connect™ provided by Yahoo is another illustration which integrates social communities (e.g., MySpace, Dopplr, Facebook or Last.fm), instant messaging services (e.g., Yahoo! Messenger, MSN Messenger), and the ordinary communication channels of mobile phones. Finally, as the example of the COSMOS project[2] shows, a comprehensive use of mobile technical and social opportunities appears to be no so distant future—the COSMOS project integrates contextual services into the existing mobile social web. Users are informed about the geographical distance between them and may send messages to contacts within their own vicinity. The above-mentioned trends are expected to advance in the future; yet, the possible dominance of one of them cannot be predicted.

CONCLUSION

The analysis of the potentials of mobile online communities leads to the conclusion that the social significance of the mobile Internet goes further than providing communities with an additional access channel. Besides mobile access and instant connectivity, unambiguous identification and contextual services can also be realized.

The mobile web platform gains efficiency and facilitates not only flexible, spontaneous, and emotional interactions, but also credible and intensive ones. Reckoning with these possibilities, a community mobile operator can take an economic perspective and identify revenue sources that make the commercialization of mobile communities feasible.

However, the specific technical and social characteristics of mobile communities can cast doubt upon their prospective potentials. Commercialization appears less promising as the willingness of users to pay fees is rather low. Mobile advertising poses additional problems despite its context-relatedness. Moreover, the realization of secondary revenue sources can lead to problems of trust.

The assumption that the social significance of the mobile Internet will go beyond providing an additional access channel to online communities has so far only partially been confirmed. Many challenges to mobile communities remain to be resolved in the future for the economic perspective to gain relevance and the commercial interpretation to become a success.

REFERENCES

Bliemel, F., & Fassott, G. (2002). Kundenfokus im mobile commerce: Anforderungen der kunden und anforderungen an die kunden. In G. Silberer, J. Wohlfahrt, & T. Wilhelm (Eds.), *Mobile commerce - grundlagen, geschäftsmodelle, erfolgsfaktoren* (pp. 3-23). Wiesbaden, Germany: Gabler.

Brunold, J., Merz, H., & Wagner, J. (2000). www.cybercommunities.de*: Virtual communities: Strategie, umsetzung, erfolgsfaktoren.* Landsberg, Germany: Verlag Moderne Industrie.

Diekmann, T., Kaspar, C., Seidenfaden, L., & Hagenhoff, S. (2006). *Kontextbewusste informationsdienste auf grundlage von information beacons*. Retrieved June 23, 2008, from http://www.gi-mms.de/mms2006/kurzbeitraege/diekmann.pdf

Eigner, C., & Nausner, P. (2003). Willkommen, 'social learning'! In C. Eigner, H. Leitner, P. Nausner, & U. Schneider (Eds.), *Online-communities, weblogs und die soziale rückeroberung des netzes* (pp. 52-94). Graz, Austria: Nausner & Nausner.

Fremuth, N., & Tasch, A. (2002). *Virtuelle und mobile communities - begriffsklärungen und implikationen für geschäftsmodelle* (Arbeitsberichte des Lehrstuhls für Allgemeine und Industrielle Betriebswirtschaftslehre, No. 35). München: Lehrstuhl für Allgemeine und Industrielle Betriebswirtschaftslehre.

Gebert, D., & Von Rosenstiel, L. (1992). *Organisationspsychologie: Person und organisation*. Köln, Germany: Kohlhammer.

Gerum, E., Sjurts, I., & Stieglitz, N. (2003). *Der mobilfunkmarkt im umbruch - eine innovationsökonomische und unternehmensstrategische analyse.* Wiesbaden, Germany: Deutscher Universitäts-Verlag.

Groh, G. (2003). *Ortsbezug in kontext-sensitiven diensten für mobile communities*. Retrieved June 20, 2008, from http://www11.informatik.tu-muenchen.de/lehrstuhl/personen/groh/pub/gis.pdf

Gross, R., & Acquisti, A. (2005). *Information revelation and privacy in online social networks (the Facebook case)*. Retrieved June 22, 2008, from http://www.heinz.cmu.edu/~acquisti/papers/privacy-facebook-gross-acquisti.pdf

Hamman, R. B. (2000). Computernetze als verbindendes element von gemeinschaftsnetzen. In U. Thiedeke (Ed.), *Virtuelle gruppen. Charakteristika und problemdimensionen* (pp. 221-243). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Heinonen, K., & Strandvik, T. (2007). Consumer responsiveness to mobile marketing. *International Journal of Mobile Communications*, *5*(6), 603–617. doi:10.1504/IJMC.2007.014177

Heng, S. (2006). *Entgegen vielen erwartungen! Breitbandige mobilfunktechnologie UMTS ist realität.* Retrieved June 5, 2008, from http://www.dbresearch.com/PROD/DBR_INTERNET_DEPROD/PROD0000000000198071.pdf

Herzig, M. (2001). Basistechnologien und standards des mobile business. *Wirtschaftsinformatik*, *43*(4), 397–404.

Hummel, J. (2005). *Online-gemeinschaften als geschäftsmodell - eine analyse aus sozio-ökonomischer perspektive.* Wiesbaden, Germany: Deutscher Universitätsverlag.

Kim, A. J. (2000). *Community building on the Web: Secret strategies for successful online communities.* Berkeley, CA: Peachpit Press.

Koch, M., Groh, G., & Hillebrand, C. (2002). *Mobile communities - extending online communities into the real world.* Retrieved May 27, 2008, from http://www11.informatik.tumuenchen.de/publications/pdf/Koch2002c.pdf

Koesch, S., Magdanz, F., & Stadler, R. (2007). *Soziale netzwerke - mobile kontaktbörsen.* Retrieved June 15, 2008, from http://www.spiegel.de/netzwelt/mobil/0,1518,470250,00.html

Kurose, J. F., & Ross, K. W. (2005). *Computer networking - a top-down approach featuring the Internet*. Boston: Pearson Education.

Larsen, J., Urry, J., & Axhausen, K. W. (2006). *Mobilities, networks, geographies*. Aldershot, UK: Ashgate.

Leitner, H. (2003). Online-community, 'hands on!'. In C. Eigner, H. Leitner, P. Nausner, & U. Schneider (Eds.), *Online-communities, weblogs und die soziale rückeroberung des netzes* (pp. 11-51). Graz, Austria: Nausner & Nausner.

Markus, U. (2002). Integration der virtuellen community in das CRM: Konzeption, rahmenmodell, realisierung. *Electronic Commerce*, *15*.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*, 370–396. doi:10.1037/h0054346

Michelsen, D., & Schaale, A. (2002). *Handy-business: M-commerce als massenmarkt*. München, Germany: Financial Times Prentice Hall.

Petrzek, D. (2007). Süchtig nach Myspace & Co. *Internet World Business*, *26*, 1.

Reichwald, R., Erben, R., Fremuth, N., & Tasch, A. (2002). Mobile communities: Phänomen und erlösungspotenziale. In R. Reichwald (Ed.), *Arbeitsberichte des Lehrstuhls für Allgemeine und Industrielle Betriebswirtschaftslehre* (no. 36). München, Germany: Lehrstuhl für Allgemeine und Industrielle Betriebswirtschaftslehre.

Reichwald, R., Fremuth, N., & Ney, M. (2002). Mobile communities - erweiterung von virtuellen communities mit mobilen diensten. In R. Reichwald (Ed.), *Mobile kommunikation* (pp. 521-537). Wiesbaden, Germany: Gabler.

Reisinger, D. (2007). *'Bluejacking,' 'bluesnarfing' and other mobile woes*. Retrieved June 18, 2008, from http://news.cnet.com/8301-10784_3-9764450-7.html

Schneider, U. (2003). Online-community - neues medium und/oder neue sozialform? In C. Eigner, H. Leitner, P. Nausner, & U. Schneider (Eds.), *Online-communities, weblogs und die soziale rückeroberung des netzes* (pp. 95-114). Graz, Austria: Nausner & Nausner.

Schubert, P., & Ginsburg, M. (2000). Virtual communities of transaction: The role of personalization in electronic commerce. *EM Electronic Markets*, *10*(1), 45–56. doi:10.1080/10196780050033971

Stader, R. (2001). Client- und endgerätetechnologien für mobile community-support-systeme. In U. Baumgarten, H. Krcmar, R. Reichwald, & J. Schlichter (Eds.), *Community online services and mobile solutions - projektstartbericht des verbundvorhabens COSMOS* (pp. 31-44). München, Germany: Institut für Informatik, Technische Universität München.

Tasch, A., & Brakel, O. (2004). *Location based community services – new services for a new type of Web communities*. Retrieved May 25, 2008, from http://www.cosmoscommunity.org/downloadFiles/Lisbon-format-final.pdf von Tetzchner, J. S. (2008). *Mobile browsing report, state of the mobile Web: First quarter, 2008*. Retrieved June 23, 2008, from http://www.opera.com/mobile_report/

Wiedmann, K.-P., Buckler, F., & Buxel, H. (2000). Chancenpotentiale und gestaltungsperspektiven des m-commerce. *Der Markt*, *39*(153), 84–96. doi:10.1007/BF03036349

Winkler, K., & Mandl, H. (2004). *Virtuelle communities - kennzeichen, gestaltungsprinzipien und wissensmanagement-prozesse*. Retrieved June 12, 2008, from http://epub.ub.unimuenchen.de/archive/00000323/01/ FB_166.pdf

Yom, M. (2002). Utility und usabilty im mobile commerce. In G. Silberer, J. Wohlfahrt, & T. Wilhelm (Eds), *Mobile commerce - grundlagen, geschäftsmodelle, erfolgsfaktoren* (pp. 173-184). Wiesbaden, Germany: Gabler.

KEY TERMS AND DEFINITIONS

Cosmos: The Community-Driven Systems Management in Open Source (COSMOS) project is a collaboration of software developers aimed to create standards-based tools for system management.

Mobile Internet: Use of TCP/IP based services and protocols with mobile devices via wireless communications technologies.

Mobile Social Web: Refers to mobile social networks and other Web 2.0-based applications in which people access and form online communities by using mobile devices.

Mobile Wireless Devices: Handheld electronic devices with wireless capability to connect to the Internet. Examples include mobile phones, smart phones, and personal digital assistants.

Mobility: Unrestricted transfer of text, voice or data independent of user's physical location.

Online Community: A social group that interacts through a web platform over an extended period of time.

Social Web: Refers to Web 2.0-based technologies and applications that are used to support communication and facilitate social contact, such as, social networking sites, massively multiplayer online role-playing games, photo and video sharing, online stores and auction houses, virtual worlds, and collaborative wikis.

ENDNOTES

1 http://www.shozu.com
2 See http://www.eclipse.org/cosmos

This work was previously published in Handbook of Research on Social Interaction Technologies and Collaboration Software; Concepts and Trends, edited by T. Dumova; R. Fiordo, pp. 11-21, copyright 2010 by Information Science Reference (an imprint of IGI Global).

Chapter 1.3
Social Semantic Web and Semantic Web Services

Stelios Sfakianakis
ICS-FORTH, Greece

ABSTRACT

In this chapter the authors aim to portray the social aspects of the World Wide Web and the current and emerging trends in "Social Web". The Social Web (or Web 2.0) is the term that is used frequently to characterize Web sites that feature user provided content as their primary data source and leverage the creation of online communities based on shared interests or other socially driven criteria. The need for adding more meaning and semantics to these social Web sites has been identified and to this end the Semantic Web initiative is described and its methodologies, standards, and architecture are examined in the context of the "Semantic Social Web". Finally the embellishment of Web Services with semantic annotations and semantic discovery functionality is described and the relevant technologies are explored

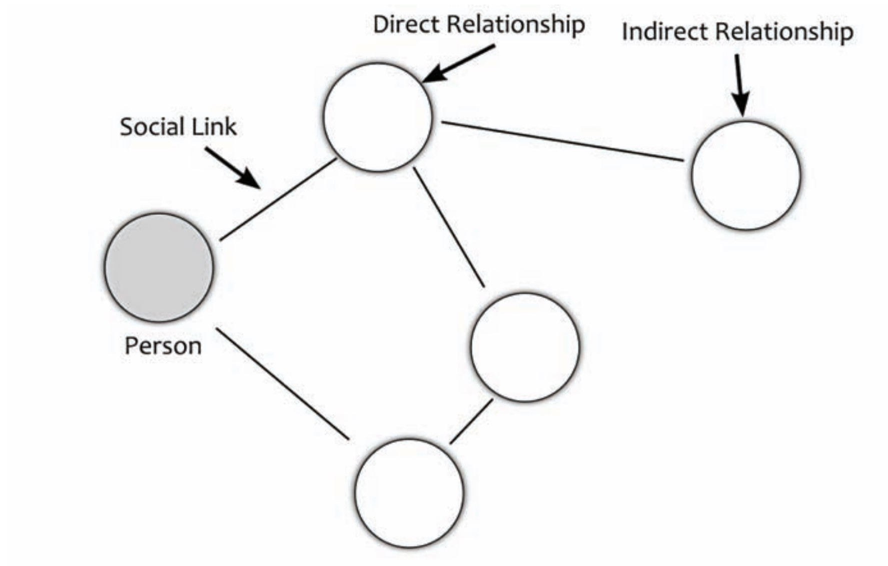
INTRODUCTION

The World Wide Web (WWW or, simply, the "Web") has been used extensively as a huge network of interconnected islands of data where documents are linked, searched for, and shared, forming a massive, albeit not always well organized, digital library. Sharing of digital content has always been the major requirement for the Web since its inception and will continue to be one of its most important features in the years to come. Nevertheless, what we experience nowadays is the endeavor for extending this sharing to cover also additional artifacts beyond plain documents, like data, information, and knowledge. The power of the hyperlinks, connecting different, possibly disparate entities, can also be exploited in order to connect information sources and people: not just "dumb" machine readable data but dynamic content like user profiles and ultimately people themselves for building virtual communities. The vision is that the current web of computers and documents will be broadened to the web of people. A "People Web" is the one where users are the nodes of the graph, the edges being their relationships and interactions in space and time, thus constructing new virtual societies (see Figure 1).

This new environment is leveraged by the introduction of an array of technologies collectively identified as Semantic Web (Berners-Lee, Hendler, & Lassila, 2001). The Semantic Web builds upon

DOI: 10.4018/978-1-60566-340-1.ch011

Figure 1. A social graph



the existing Web and provides the necessary substrate for giving "meaning" and "Semantics" to Web resources and Web interactions. The benefits will be many in a number of application domains and while the challenges, technological and other, are numerous, the momentum is strong and the Semantic Web slowly but steadily enters in a number of diverse domains like health and life sciences.

Furthermore the Semantic Web promises a great potential for supporting the construction and smooth operation of Web communities of people. In this chapter we study its fusion with social software and software for machine to machine communication over the Web for supporting this vision.

BACKGROUND

Since its launching in 1990, the Web has grown exponentially both in terms of size and in terms of use and utility to people and organizations. The inherent simplicity of hypertext and its feature limited, in comparison to previous hyper linking systems, one-way, inexpensive links (Universal Resource Identifiers – URIs) but also the employment of the Internet as its networking substrate led to its wide adoption and success.

In spite of its success and popularity the early version of the Web lacked in many respects, ranging from user accessibility and user interface design to the ability to repurpose and remix existing Web-based data in not pre-established ways. Although the hyper linking facility allowed the interconnection of different documents on the Web, the "traditional" Web suffers from fragmentation in the sense that the huge graph lacks any organization and discipline. This "anarchy" in the Web has also been its driving force for the success it has enjoyed so far but created the need for having special machinery, e.g. search engines like Google, to deal with the efficient indexing and discovery of the available information. Despite the fact that the search engine technology has made important steps in indexing and searching massive amounts of data on the Web, there's still the issue that keyword based searching is limited on its potential and usually finding "what the user wants" proves to be a tedious task. Another major limitation of this environment is that the people are not part of the equation. Users are expected to be the actors triggering the Web interactions but they are not allowed to participate and be involved enough in these interactions:

- Content delivered is not personalized. What the user gets back is usually not in-line with her interests or other preferences and there's no feedback link going from the user back to the system she interacts with so as to guide future interactions.
- Contextual information is not taken into consideration. The people as complex systems do not act in an easily predetermined way and the context of their actions is usually ignored or not taken advantage of. This context information ranges from the user's profile, which is also dynamic in nature, to the specific objective she/he is trying to achieve at a specific point in time.
- Content is passive and static, stored and maintained in back end databases, which the users do not have the ability to enrich or customize to their own needs
- Communication and collaboration of the users to build Web communities are not supported enough. Discussion forums were the sole way to build such communities but with no means to support intelligent integration of the different forums or to enhance the user collaboration experience.

These and other requirements are the ones that the Social Web tries to tackle. Social Web does not represent a shift or radical change in technology per se but rather a shift on the perception of the human – machine interaction by placing the users in the centre of the system and in control of these interactions. But from the other end of the spectrum there is also a clear need for making the Web itself more intelligent to support these machine facilitated social interactions. The Semantic Web could provide for such an enabling technology and recently the convergence of the Social and the Semantic Web and the experimentation of the two working in complementary ways have gained a lot of attention and research interest.

SOCIAL WEB OR WEB 2.0

The situation described in the previous section led to the emergence of a new breed of Web applications and sites that were collectively identified as "Web 2.0" by Tim O'Reilly (2005) and whose major design principle is to "*harness network effects to get better the more people use them"*. The value of "Web 2.0" sites and applications therefore comes to a large extent by the number of users participating and actively communicating and sharing through them so the term "*Social Web*" is actually a synonym. The social nature of this Web is evident when the collaboration of people and their active contribution is considered. The very essence of such sites is the building and maintenance of Web based *virtual communities* of people that produce and maintain *collective knowledge*. Examples of such community oriented and social networking sites include:

- Blogs, i.e. Web sites managed by individuals that provide news or opinions on certain subjects (typically personal online diaries), where typically other people are able to leave comments. In addition to comments, the hyperlinking facility of the Web has been extensively used to provide "trackbacks" (i.e. reverse hyperlinks that identify who is talking about me) and recommended blogs ("blogrolls"). Therefore blogging has been emerged as a method for anyone to publish content on the Web and building online communities of people that communicate, share, and integrate.
- "Social bookmarking" sites (e.g. http://del.icio.us/) where users can store and share their bookmarks with the additional

possibility to provide metadata through the means of tags, i.e. terms that denote concepts, meaning, intent, etc. These sites provide for user maintained and collaborative indexing of the Web content in a way that it may be more efficient to search there for something than in general purpose Web search engines.
- "Wikis" (e.g. http://en.Wikipedia.org), which are collaboratively built Web sites where the users, through custom made and user friendly interfaces, are able to create, share, enhance, and manage the content.
- Content sharing sites, e.g. YouTube (http://www.youtube.com/) for videos or Flickr (http://www.flickr.com/) for photographs, where the users upload their multimedia content and share it online with other users.
- Social networking sites, such as Facebook (http://www.facebook.com/) and MySpace, for online communities of people who share interests and activities or who are interested in exploring the interests and activities of others.
- Classified advertisement sites, e.g. Craigslist (http://www.craigslist.org), which offer advertisements for jobs, resumes, services, etc. grouped in categories.

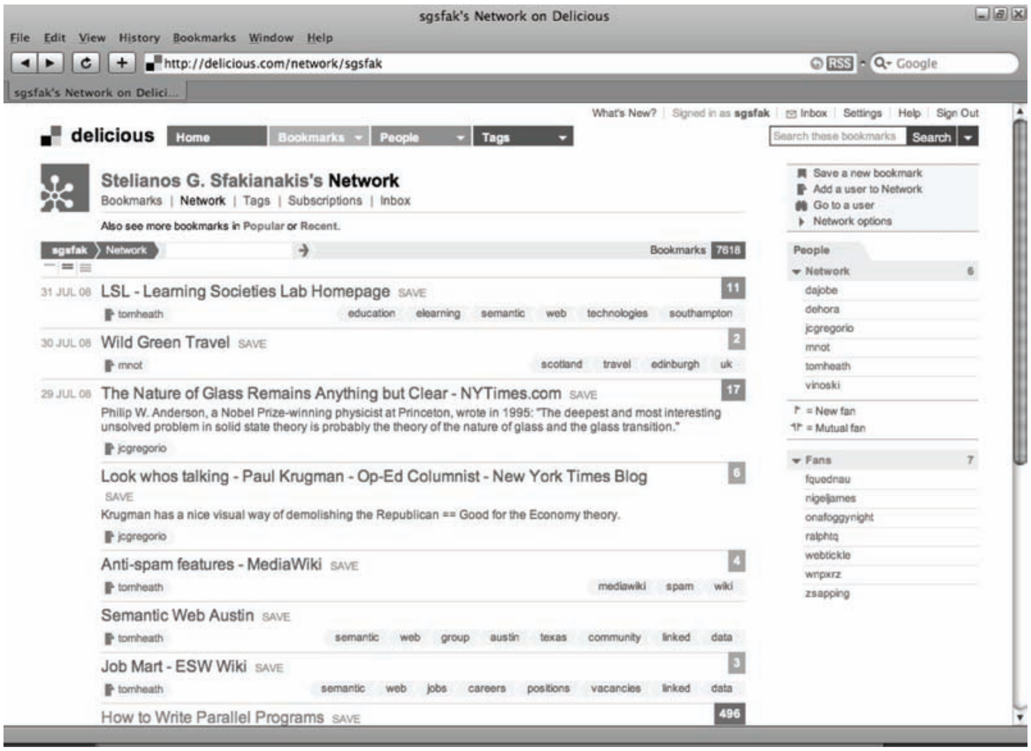
If we take only "Wikis" as an example we can see that these Web sites have been used in a multitude of ways:

- As online encyclopedias, e.g. Wikipedia
- As free dictionaries, e.g. Wiktionary (http://en.wiktionary.org)
- As free libraries of educational books, e.g. Wikibooks (http://en.Wikibooks.org)
- As software development repositories and issue tracking systems, e.g. Trac (http://trac.edgewall.org/)
- As open forums to promote research interests, like OpenNetWare (http://openwetware.org/) for biology and biological engineering
- As open educational centers to support learning activities, e.g. Wikiversity (http://en.Wikiversity.org/)
- As social event calendars, like Upcoming (http://upcoming.yahoo.com/)

The single distinctive feature of Wikis and a central trait of the social Web sites is the *user generated content* and its "open" editing: anyone can edit an existing Wiki article or create a new one for a particular topic if it doesn't exist already. The users therefore are responsible for adding content and maintaining the information that is available from these sites. Of course such an approach can raise a lot of concerns about the validity of the content, the lack of authority, etc.[1] and there have been cases in the past where such skepticism was proven true, such as the Seigenthaler incident[2]. Nevertheless this open model has worked quite well in practice and in general so that at the time of this writing Wikipedia is considered by many a serious competitor to the Encyclopedia Britannica. The reason for this can be explained as another instantiation of the "wisdom of crowds" phenomenon (Surowiecki, 2004): the participation of many people, possibly with different background, habits, way of thinking, and so on, in a decision making process usually yields better results than when the individual opinions are considered separately from one another.

The contribution of user content and the sharing of the uploaded information are the main forces for the formation of *online communities* of people. In Figure 2 an example of this community creation process is shown for the Del.icio.us online bookmarking site. Online bookmarking sites like this provide the means for storing and organizing bookmarks of Web sites on the Web instead of the users' desktop browsers. By storing their bookmarks in a central area the users are additionally enabled to create their online social networks by registering other users as members

Figure 2. Del.icio.us networks of users



of their network so that they can be notified about the bookmarking activity of these users. These networks therefore connect users with their friends, family, coworkers, or even totally strangers when they unexpectedly meet each other on the Internet and discover they have similar interests. Facilitated by these network links the users can subsequently observe each other's online behavior and even proactively send interesting Web sites addresses to their peers, easier and quicker than using email or instant messaging.

What the previous examples show is that in the Social Web users are in the limelight: they are the primary actors in the data sharing process through their contributions and online behavior. They are usually indulged by the low cost entry and participation in these Web sites, and, to a lesser extent, by the visual appeal the Web 2.0 sites offer to the viewer. The modern Web sites are actually Rich Internet Applications (RIA), where the majority of the business and control logic resides on the client (i.e. the Web browser), leveraged by technologies like AJaX[3] and Comet[4] which provide more responsive user interfaces.

The Social Web offers a meeting point for people to collaborate and share information in an open environment. The openness is a distinctive characteristic of Web 2.0 and it's supported by Open Data APIs like content syndication via RSS/Atom[5] and lightweight Web services interfaces like Open Search[6] . These technologies enable the view of Web sites as Web applications and their synthesis ("mashup") in more complex applications. An example of such combination of existing Web sites and their data to create new/aggregated content is Housing-Maps (http://www.housingmaps.com/) where houses to rent or buy are located through Craiglists and projected over geographic maps drawn from Google Maps (http://maps.google.com) so that a user can easily locate the information he wants in an interactive and visual way. A more general and reusable way to combine and "mix" content from different Web sites is offered by Yahoo! Pipes[7] which can be thought of a simple but effective way to build "workflows" and "dataflows" on the Web.

The above discussion shows that collaboration between people but also between Web sites/applications supports the notion of "collective intelligence" to the Social Web. An instance of this intelligence built collectively is the creation of "folksonomies" for categorization of resources. A quite popular way of classifying content in Web 2.0 Web sites is through "tagging". A tag is a keyword which acts like a subject or category. The user is allowed to attach whatever keywords she wants to identifiable content such as links in the case of social bookmarking, or videos and photographs in the case of digital content sharing. The important thing is that tags can be shared, used in searches, or recommended based on the choices of other users for the same content.

The new term "folksonomy", as a fusion of the words "folks" and "taxonomy", has been suggested to describe this method of classifying content through tags that are collaboratively generated and shared. Of course these "poor man's" classification schemes are informal in nature, could contain duplication in meaning, or be simply erroneous but again they are contributed by the users and the more people contributing the more robust and stable these "folksonomies" become. A self adapting and auto regulating method is usually followed through the use of tag clouds (Figure 3). In simple terms a tag cloud is a visual representation of a user's tags where each tag is weighted based on the user preferences and how many times he has used the tag. Through such an approach "good" tags are likely to prevail assuming that the user participation is high.

Collaboration, sharing, "mashing", annotating and "tagging" content are roughly the distinctive features of Web 2.0 and although in most of the cases the approach is not formal or the solutions are suboptimal the user participation and their socialization needs have driven the evolution of Web of documents to the Web of People (Ramakrishnan & Tomkins, 2007).

SEMANTIC WEB

To the other end of the spectrum, with roots in Artificial Intelligence research, the Semantic Web emanated as an extension to the current version of the Web that aims to enhance it by the promotion of higher level sharing and integration of data and information. According to Berners-Lee et al. (2001):

The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

The Semantic Web aims to support the representation and exchange of information in a meaningful way so as to make possible the automated processing of descriptions on the Web. The objective is to enrich the unstructured information in the current Web with machine processable descriptions of the Semantics in order to make its navigation and exploration by software agents as easy as it's for the human users today, or even easier. In this context Semantic Web promotes a shift from the current "syntactic" world to the future "Semantic" world of services, applications, and people and aims to make the machine to machine communication feasible so that not only data but also information and finally knowledge are shared.

The Semantic Web Technology Infrastructure

In technological terms the Semantic Web architecture consists of an array of technologies that can roughly be visualized in a layered design layout as depicted in Figure 4. The basic infrastructure in the bottom layers in this stack of technologies is

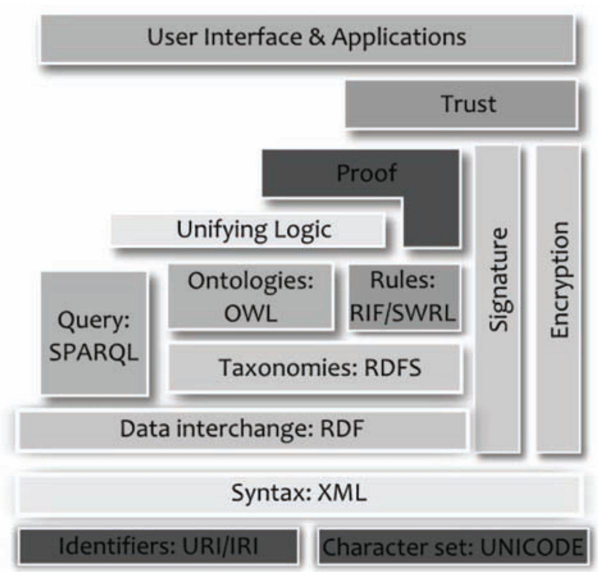
Figure 3. A tag "cloud"

the exactly the same to the syntactic Web: Uniform Resource Identifiers (URIs) used for identification of Web resources, universal encoding schemes for characters, i.e. Unicode, and XML and its related technologies (e.g. XML Namespaces) as a ubiquitous data serialization format. Some of the upper layers like Proof and Trust are missing or are work in progress. Here we will concentrate on the middle layers where the core infrastructure technologies of the Semantic Web reside: RDF, RDF Schema/OWL, and SPARQL.

The *Resource Description Framework* (RDF) is a syntax neutral data model that enables the description of Web resources in a simple way (Lassila, Swick, et al., 1999). At the core of RDF there is a model for representing and describing *resources* through named *properties* (also known as *predicates*) and their values. The resources can be anything that can be identified with a URI. Although in the initial specification of RDF resources were limited to Web documents and Web sites, it is possible and quite frequent in practice to describe, by the means of RDF and the various URI schemes, real world entities like people, or more abstract things like relationships and concepts. The use of URIs and especially the HTTP based ones for identifying persons or other physical entities may seem strange at first but this is in compliance with the architecture of the World Wide Web (Berners-Lee et al., n.d.) which strongly suggests the use of URIs for identifying anything that can be of importance irrespective of how abstract or tangible it may be.

The properties serve both to represent attributes of resources and to represent relationships between resources. They are also identified though URIs to make them unique. The combination of resources and the properties that connect them builds the simple RDF data model. In this data model the primary informational building block is the "triple" which denotes the subject – property - object expressions (Figure 5). The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a

Figure 4. The Semantic Web stack of technologies



relationship between the subject and the object. Since an object of a triple can be the subject of another one, a set of RDF triples forms a *directed graph* where the RDF resources, both subjects and objects, are the nodes of the graph and the predicates are the labeled arcs. As an example, in Figure 6 there's a simple RDF graph. The graph shown in the figure describes an entity identified through the URI "http://ssfak.org/stelios/", apparently denoting a person, which has a "name" property with the value "Stelios Sfakianakis", a property denoting the homepage of an organization a person works for relating it to the resource "http://www.ics.forth.gr/cmi-hta/", and a "maker" property that connects it (backwards, as an object) to the resource identified as "http://ssfak.org".

RDF as an abstract model is independent of any specific serialization syntax. The normative representation syntax for RDF graphs is XML but more lightweight formats, such as Turtle (Beckett & Berners-Lee, 2008), exist.

The simplicity and flexibility of RDF is evident but in certain cases its generality must be formally confined so that software entities are able to correctly exchange the encoded information. For example, stating that an animal is the creator of a Web page does not make sense in the real world but RDF does not forbid anyone for making such a claim. Ontologies (Uschold & Gruninger, 1996) provide such a tool to specify what can be expressed in the context of an application domain or in a real world scenario, what is the underlying meaning, and how the information presented can be further processed to generate more information.

Figure 5. RDF Data Model

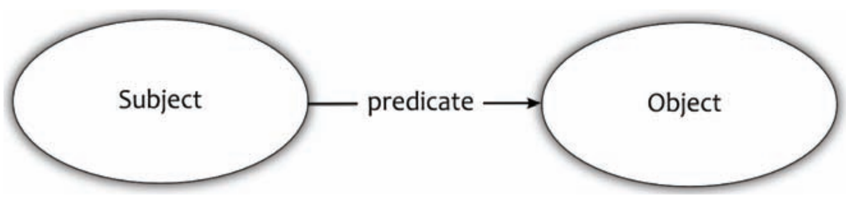
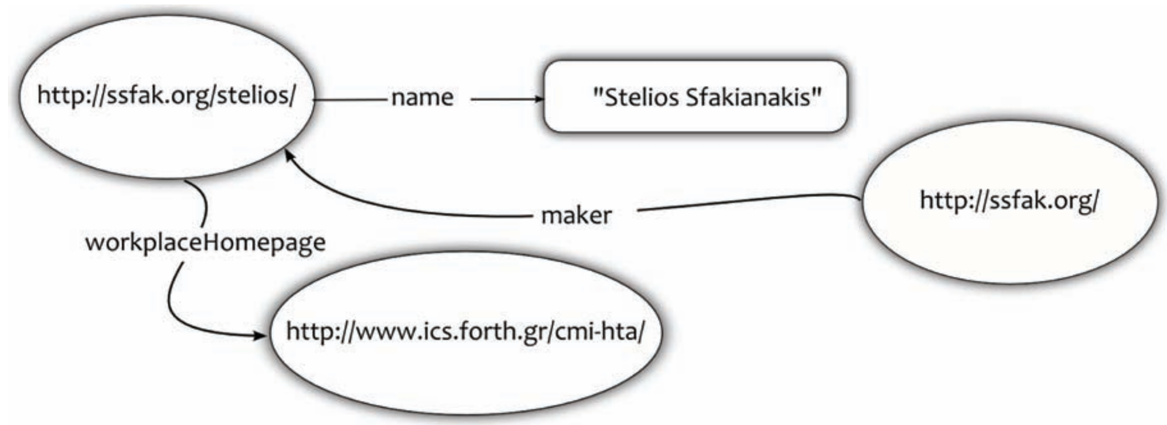


Figure 6. Abstract representation of RDF triples



Moreover ontologies and their less powerful relatives like taxonomies and thesaurus provide the means for achieving a common interpretation of a domain and a shared understanding of the concepts and relationships involved. In the Semantic Web there are two main technologies for providing such rigor: RDF Schema and OWL (Brickley & Guha, 2004; Dean, Schreiber, et al., 2004). RDF Schema provides the means for defining classes, class hierarchies, properties, property hierarchies, and property restrictions. Its expressive power is basically limited to the representation of concepts, their relations, and taxonomies of concepts. On the other hand the Web Ontology Language (OWL) was introduced to address the need for more expressiveness and extends the RDF Schema by providing three variants: OWL-Lite, OWL-DL, and OWL-Full. Without delving into details, the different species of OWL provide different degrees of expressiveness and are able to define existential restrictions, cardinality constraints in properties, property types like inverse, transitive, and symmetric, and a lot more. The added features of OWL allow the ontologies built in conformance to it to be formally treated and the data represented are amenable to "reasoning" and inference, i.e. they can be processed according to formal logic rules to deduce new information. All these happen on the basis of the Web infrastructure: RDF resources and their URI references are used, the open world assumption is followed, since partial information on the Web is a quite frequent phenomenon, and the ontologies themselves can be freely intermixed and meshed since hyperlinks are employed everywhere.

Since RDF is the common interchange and representation model of information, the Semantic Web transforms the hyperlinked syntactic World Wide Web to a huge database or a Global Giant Graph, as Tim Berners-Lee put it. The standard query language for this huge database is SPARQL (Prudhommeaux & Seaborne, 2008), which is similar to SQL. In addition to the query language the SPARQL standard defines an application protocol for the submission of queries to RDF sources and the retrieval of results. With the query language and the access protocol defined, the SPARQL specifies a Web friendly interface to RDF information, whether this is actually stored as RDF triples or not. It is therefore feasible to make SPARQL queries to relational or other databases through an appropriate wrapper or transformation process that translates, either online or in some preprocessing step, the internal data to an RDF compliant format. As a result these Semantic Web technologies enable the connection of data between different and heterogeneous data sources, effectively allowing data in one data source to be linked to data in another data source. (Bizer, Heath, Idehen, & Berners-Lee, 2008)

SOCIAL SEMANTIC WEB

In recent years the cross pollination of Semantic Web technologies and Social Networking has emerged as an interesting roadmap. The Semantic Web technology can significantly enrich and expedite the Social Web in order to establish the *Semantic Social Web* (*Greaves, 2007;*Gruber, 2007*).* In the Semantics-enabled social Web content can be easily connected, integrated, navigated, and queried so that the benefits of today's Social Web can be greatly enhanced and augmented beyond the limited user experience offered by social networking sites alone or the restricted keyword based search and matching.

What does the Semantic Web offer to the Social Web? First and foremost, the Semantic Web technologies can be used to provide rigor and structure to the content of the user contributions in a form that enables more powerful computation. Currently social Web applications are more focused on the distribution and the management of content and the social interactions around it rather than the provision of Semantically rich descriptions of the data. Although there are popular, "low end" technologies like "microformats" and tagging/"folksonomies" to cater for the annotation and the description of data, these seem to be ad hoc and unstructured efforts in comparison to the formal Web ontologies and metadata descriptions. On the other hand, as already described, the Semantic Web promotes the global distribution and integration of resources in a single, giant, interoperable graph. So, additionally, the standards and infrastructure of the Semantic Web can enable data sharing and computation *across* independent, heterogeneous social Web applications.

Furthermore, the Semantic Web can enhance the Social Web with additional intelligence as Jemima Kiss (2008) wrote:

If Web 2.0 could be summarized as interaction, Web 3.0 must be about recommendation and personalization.

An example of such added value is the case of Semantic Wikis (e.g. Schaert, 2006; Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., & Studer, R., 2006). The Semantic Wikis support the annotation with Semantics descriptions the links and the content they provide and take advantage of these annotations for providing more intelligent search and navigation. The annotation is usually done by some extended version of the Wiki editing syntax so that every link to another page or any important attribute of the current page is annotated with a property identifier. For example in a Semantic Wiki's page about the Europe the amount of its population, which is a number, can be wrapped with the appropriate metadata that denote that this number represents the population. Such metadata annotation makes structured search easy, e.g. for queries like what is the population of Europe, or which continents have population above a certain amount. Additionally it facilitates the users in providing more active content in the pages by incorporating "online queries" in the Wiki pages, in the sense that the page's content can be dynamically generated by the results of these queries on the metadata annotations. Although the details may vary from one implementation to another, there's usually an underlying model based on RDF and OWL to support these Wikis and the content can be exported in a Semantic Web compliant format. DBPedia is an interesting example of a truly Semantic Web Wiki which offers the content of Wikipedia in a machine-readable and searchable format (Auer et. al., 2007).

In another application area, Semantic Web technologies can facilitate the browsing experience of people and the searching capabilities of the Web search engines. Unlike traditional search engines, which "crawl" the Web gathering Web pages information, Semantic Web search engines index RDF data stored on the Web and provide an interface to search through the crawled data. Because of the inherent Semantics of RDF and the other Semantic Web technologies, the search and information retrieval capabilities of these search engines are potentially much more powerful than those of current search engines. Examples of such early Semantic Search Engines include the Semantic Web Search Engine (SWSE, http://www.swse.org/), Swoogle (http://swoogle.umbc.edu/), and Zitgist Search (http://www.zitgist.com/). These and other Semantic Web search engines explore and index the documents of the Semantic Web and its ontologies by the means of user friendly interfaces that hide the details and complexities of the technology.

Blogs, which are one of the most prominent examples of the Social Web, can also be enhanced with Semantics. Augmenting a blog with content and structural metadata is usually called Semantic Blogging (Cayzer, 2004; Bojars, Breslin, & Moller, 2006). Putting Semantics in a blog's contents means that the topic of the content is described in a machine processable way. On the other hand describing the structure of the blog Semantically entails the description of the entities that compose it: the posts, the comments, the users, etc. To this end there are a number of efforts to make the Semantic Web more social by building new ontologies to support people in their social interactions and provide Semantics to the Social Web. Two of such ontologies, SIOC and FOAF, are of particular importance in the context of Semantic Blogging and are described below.

SIOC

Existing online community sites usually provide rich information for specific interest groups but they are isolated from one another, which makes difficult the linking and merging of complementary information among different sites. The Semantically-Interlinked Online Communities (SIOC) project aims to link online community sites using Semantic Web technologies. It defines methods to describe the information that communities have about their structure and contents, and to find related information and new connections between content items and other community objects. SIOC again is based around the use of *machine-readable information* provided by these sites.

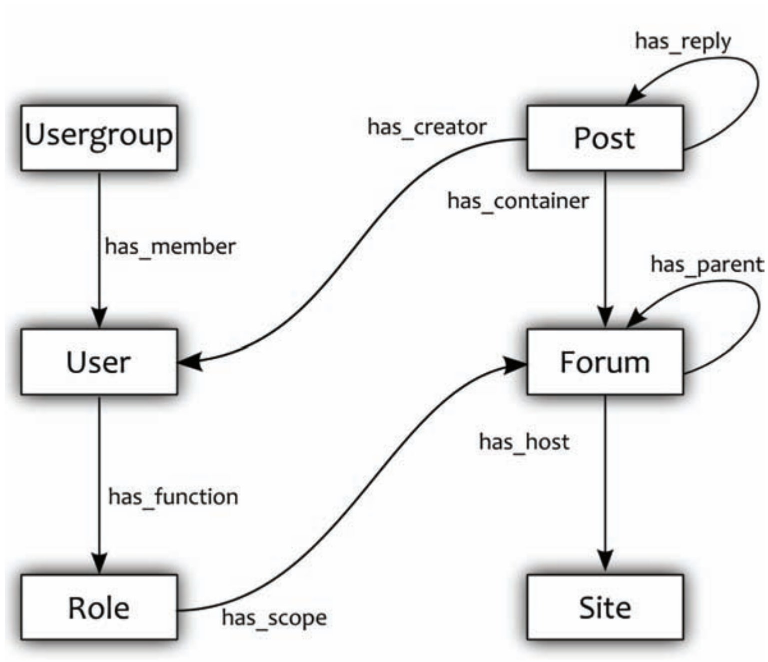
The main entities of SIOP are shown in Figure 7 and it's easy to see the role and the function of the main concepts. The entity Site refers to the location of an online community or set of communities, which hosts one or many blogs. A Forum can be thought of a discussion area on which posts are made. In a Forum a number of posts are contained where a Post represents an article or a message send by a user to the forum. Posts can be connected as people reply to previous posts and these connections can cross site boundaries since the identifiers of posts (as well as any Semantic Web resource) are universal and unique.

From the figure above it can be said that SIOP defines a common schema for the different blog sites and discussion forums. This of course needs not be their internal schema but a common, shared, and standard representation of their information model. Adopting SIOP therefore is a major step in achieving the integration of social content in Web 2.0.

FOAF

The Friend-Of-A-Friend (FOAF) project focuses on expressing mostly personal information and relationships in a machine-readable form. A central entity in the FOAF vocabulary and the one most frequently used is the Person (Figure 8). According to FOAF a Person may have names, e-mails, interests, publications, etc. It can also be connected to other resources like the Web site of the organization he/she works for (foaf:workplaceHomepage property), a personal blog site (foaf:weblog), the Website of his/her school (foaf:schoolHomepage), or to other people that he/she knows (foaf:knows).

Figure 7. The main classes and relationships of SIOC



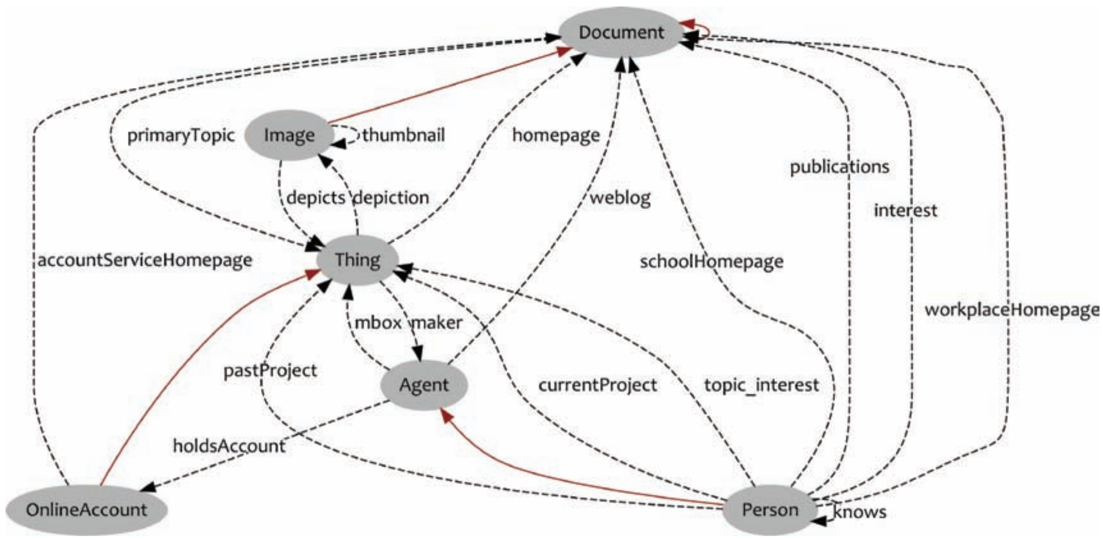
A lot of personal information can be therefore represented and parts of the real world's social graph can be inferred by following the foaf:knows relationship.

Of particular importance to the Social Web is the support the FOAF vocabulary offers to link the physical persons (foaf:Person) to the accounts they hold in a number of social Web sites (e.g. Flickr, Facebook) through the foaf:holdsAccount property. It is therefore possible through a single FOAF document that could be indexed in a Semantic Search engine to join all these different accounts and the information each of them exposes. Integration of different social content and behavior can be achieved and the resulting graph of information is searchable in an unambiguous and machine interpretable way.

Semantic Web Services

Application integration requires an agreed infrastructure to be in place for the exchange of information over the network. Over the last couple of decades there have been several attempts for defining such an infrastructure, such as Sun/RPC, CORBA, Microsoft's DCOM, Java RMI, and others. Currently Web Services are the favorite and most popular technology for building distributed systems over the Internet. As a middleware technology Web Services represent a new generation that tries to mitigate the problems of legacy integration technologies such as CORBA by adopting a more Web friendly substrate. Such a different approach seems to be needed in order to support business-to-business integration over the Internet where crossing organization borders has implications on the security, interoperability, scalability, maintenance, flexibility, and other aspects of application integration. In order to achieve these goals the Service Oriented Architecture (SOA) has been proposed. Informally speaking, in such architecture (Web) Services are network accessible entities that offer a number of functionalities to their callers. The SOA environ-

Figure 8. The main classes and relationships of FOAF



ment should be highly dynamic as suggested by a number of real world phenomena, like network instability, changing real world requirements and settings, etc. The need for "late binding" of services and clients is important and Figure 9, depicting the main entities of SOA and their interactions, shows that a middle service repository or registry is introduced. This repository stores "offers" of functionality as these are published by service providers, and subsequently performs matching with the corresponding "requests". After some matching has been performed the corresponding parties (services and their clients) are free to communicate and exchange data.

On the technology side Web Services put more emphasis on the following:

- Transport over widely accepted Web and Internet protocols like HTTP/HTTPS and SMTP
- XML message payloads to provide the extensibility, introspection, and interoperability required in building complex multi party systems
- Platform and programming language independence

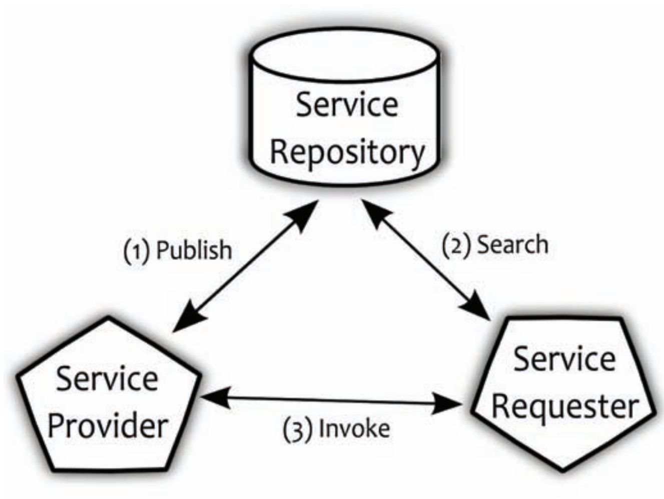
The Web itself is built around these very directions: open protocols, text based (markup, e.g. HTML) message and document content, and abstraction over implementation details. In essence the underlying infrastructure is roughly based on the following technologies:

- SOAP messaging format, which is based on XML, to provide a wrapper format and protocol for data interchange between Web services
- Web Service Description Language (WSDL) documents to describe the services' functionality and data exchange

On top of these a number of standard technologies have been specified for handling discovery (UDDI), security (WS-Security), trust (WS-Trust), composition (WSBPEL, WSCL), etc. Nevertheless for this discussion the WSDL standard is the most pertinent specification because it specifies in a machine readable format the structure of the XML messages exchanged.

Integration of computation and functionality is an additional field where Semantic Web shows a great potential of use because the Web

Figure 9. Web service architecture



Services, at their present incarnation, provide syntactic interoperability only. The WSDL service descriptions are restricted to the syntactic aspects of service interaction: how the service can be invoked, which operations may be called, what are the number and the type of the parameters each operation needs, etc. However, what the service does and in what order its operations have to be called in order to achieve certain functionalities is usually described only in natural language either in the comments of a WSDL description or in UDDI entries or other documentation. *Semantic Web Services* (McIlraith, Cao Son, & Zeng, 2001) is an "umbrella" definition to include the annotation of existing Web services with Semantics and their publication, discovery, and composition.

The vision again is to make feasible the machine to machine communication by providing machine interpretable descriptions of the services. Such descriptions will make possible the automatic discovery, composition, and invocation of services. Because of this a lot of efforts in metadata descriptions are centered on the Semantic Web and its technologies, namely RDF and OWL. In the area of the Semantic Web Services the following technologies and standards are relevant to the Semantic description of Web services:

- **UDDI (Universal Description, Discovery and Integration)** allows the discovery of potential business partners on the basis of the services they provide. Each business description in UDDI consists of a businessEntity element that describes a business by name, a key value, categorization, services offered (businessServices) and contact information for the business. Each businessService element contains descriptive information such as names and descriptions, and also classification information describing the purpose of the relevant Web service. Using UDDI, a Web service provider registers its advertisements along with keywords for categorization. A Web services user retrieves advertisements out of the registry based on keyword search. So far, the UDDI search mechanism relied on predefined categorization through keywords, but more recently specifications to use OWL in UDDI are emerging as a uniform way to express business taxonomies.
- **Semantic Annotations for WSDL and XML Schema** (SAWDL; Kopecký et. al., 2007) is a means to add Semantics inline to WSDL. It is actually a set of extensions to

WSDL 2.0 but can also be used for WSDL 1.1. With these extensions the service provider can attach references to Semantic concepts for the functionality of an operation or the type/meaning of a parameter and additional information for the transformation (mapping) of the XML data either to ("lift") or from ("lower") the corresponding Semantic terms. The Semantic domain model used is external to these annotations and could be expressed in OWL or other ontology language of choice.

- **OWL-S** (formerly DAML-S) builds on top of OWL and allows for the description of a Web service in terms of a Profile, which tells "what the service does/provides", a Process Model, which tells "how the service works", and a Grounding, which tells "how to access the service" (Martin et al., 2004). The service profile describes what is accomplished by the service, any limitations on service applicability and quality of service, and requirements that the service requester must satisfy in order to use the service successfully. The process model gives details about the Semantic content of requests, the conditions under which particular outcomes will occur, and, where necessary, the step by step processes leading to those outcomes. In the process model a service can be described as an atomic process that can be executed in a single step or a composite process that, similar to a workflow, can be decomposed in other processes based on control structures like 'if-then-else' and 'repeat-while'. Finally, Grounding descriptions supply information about the communication protocol and other transport information (such as port numbers) and the message formats and serialization methods used in contacting the service. The only currently specified grounding mechanism is based on WSDL 1.1 and will be extended to WSDL 2.0 as soon as it's finalized.
- **The Semantic Web Services Framework (SWSF)**, initiated by the Semantic Web Services Initiative (SWSI, 2004), includes the Semantic Web Services Language (SWSL) and the Semantic Web Services Ontology (SWSO). SWSL is a logic-based language for specifying formal characterizations of Web service concepts and descriptions of individual services. SWSO is an ontology of service concepts defined using SWSL and incorporates a formal characterization ("axiomatization") of these concepts in first-order logic.
- **WSMO (Web Services Modeling Ontology)** defines the modeling elements for describing several aspects of Semantic Web services (Feier et al., 2005). These elements are Ontologies, which provide the formal Semantics to the information used by all other elements, Goals which specify objectives that a client might have when consulting a Web service, Web services that represent the functional and behavioral aspects which must be Semantically described in order to allow semi-automated use, and Mediators that are used as connectors and they provide interoperability facilities among the other elements. It also defines the Web Service Modelling Language (WSML) which formalizes WSMO and aims to provide a rule-based language for the Semantic Web.
- **BioMOBY** (http://www.biomoby.org/) is a Web Service interoperability initiative in the field of bioinformatics aiming to facilitate the integration of Web-based bioinformatics resources. Currently there are two approaches to achieve such integration: The first approach, based on the Web Services paradigm, is referred to as "MOBY Services" (MOBY-S), while the second one is called "Semantic MOBY" (S-MOBY) and is based on concepts from

the Semantic Web. MOBY-S uses a set of simple, end-user-extensible ontologies as its framework to describe data Semantics, data structure, and classes of bioinformatics services. These ontologies are shared through a Web Service registry system, MOBY Central, which uses the ontologies to Semantically bind incoming service requests to service providers capable of executing them. S-MOBY on the other hand employs RDF and OWL and the document oriented infrastructure of the WWW (the GET/POST methods of HTTP) for publishing and retrieving information from its discovery servers.

As shown above this is an area of active research. So far SAWDL enjoys the approval of W3C being one of its recommendations but of course is lacking when compared with WSMO and OWL-S. Nevertheless SAWSDL can be combined with these most prominent technologies and it remains to be seen whether such approaches are adequate or something more powerful should be introduced.

CONCLUSION

The social aspects of the Web show an uprising evolution and all the indications imply that this trend will continue. The current Web 2.0 sites are quite successful in attracting people share their data and interests, and build online communities, but the next step will be to enrich them with more Semantics in the lines of the Semantic Web to provide a unifying platform for people and machines to use and collaborate. The need for Semantics (Breslin & Decker, 2007) is evident for enhancing the social networking sites with advanced filtering and recommendation services and also to provide data portability and integration between different sites and networks. This is an active area where the Semantic Web technologies can greatly help.

There have been a lot of discussions about what will be the "killer application" of the Semantic Web, which means some breakthrough in the domain that will show beyond any doubt the full potential of the Semantic Web. Nevertheless we think that Semantic Web technologies are used slowly and without much "noise" in a lot of different areas and as "extension to the existing Web" are not clearly visible but are certainly catching on. There is a common view nowadays that the Semantic Web will not supersede the Syntactic Web in any way but they will happily coexist in a symbiotic manner: the Web of documents will be enriched by the Web of data and information.

In terms of the core infrastructure what we see as emerging trend is the use of simple REST Web services (Fielding & Taylor, 2002) that present a small entry barrier and a transition from the SOAP and WSDL Web Services technologies backed by big commercial corporations like IBM and Microsoft to more flexible and agile architectures. These architectures are more bound to the existing Web and also are more Semantic Web friendly since they share common basic infrastructure and interaction protocols (e.g. Web protocols like HTTP used as application protocol and not for transport, full support for URI to access network resources and Semantic concepts, etc.). The whole history of the Web clearly shows that successful distributed systems of this scale are built on open access, open protocols, and open source methodologies combined with collaborative behavior by the people (developers, users) involved.

Research questions and issues for further investigation abound in this Semantic new world. First of all the issue of trust and security and how this is incorporated in the Semantic Web machinery should be tackled on. For example currently a user can claim anything in his FOAF document, or a malicious application can publish RDF information that contains false statements. The notion of identity and validation of the identity is important and there is ongoing work in this area, e.g. the incorporation of user certificates or Web based authentication mechanisms like OpenId[8]. Semantic Web has also increased demands for supporting indexing and reasoning over the managed content. The scalability concerns are real when we think about a Semantic Web search engine of the size of Google. Finally the adoption of these technologies by the users needs work to be done in the presentation layers as well. Easy to use, friendly, and functional user interfaces are necessary for making the transition to the Semantic Web more painless and transparent for the users.

REFERENCES

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a Web of open data. *The 6th International Semantic Web Conference (ISWC 2007)*.

Ayers, D. (2007). Evolving the link. *IEEE Internet Computing*, *11*(3), 94–96. doi:10.1109/MIC.2007.53

Beckett, D., & Berners-Lee, T. (2008). Turtle - Terse RDF triple language. *W3C Team Submission*.

Berners-Lee, T., Bray, T., Connolly, D., Cotton, P., Fielding, R., Jeckle, M., et al. (2004). *Architecture of the World Wide Web, Volume One.* W3C,

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 28–37.

Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. (2008). Linked data on the Web (ldow2008). In *WWW '08: Proceeding of the 17th International Conference on World Wide Web* (pp. 1265-1266). New York: ACM.

Bojars, U., Breslin, J., & Moller, K. (2006). Using Semantics to enhance the blogging experience. In *Proceedings of 3rd European Semantic Web Conference*,ESWC 2006, (pp. 679-696).

Breslin, J., & Decker, S. (2007). The future of social networks on the Internet: The need for Semantics. *IEEE Internet Computing*, 86–90. doi:10.1109/MIC.2007.138

Breslin, J., Harth, A., Bojars, U., & Decker, S. (2005). Towards Semantically-interlinked online communities. In *Proceedings of the 2nd European Semantic Web Conference* (ESWC05), Heraklion, Greece, LNCS, 3532, 500-514.

Brickley, D., & Guha, R. (2004). *RDF Vocabulary description language 1.0: RDF schema*. W3C Recommendation 10 February 2004. World Wide Web Consortium.

Cayzer, S. (2004). Semantic blogging and decentralized knowledge management. *Communications of the ACM*, *47*(12), 47–52. doi:10.1145/1035134.1035164

Dean, M., Schreiber, G., et al. (2004). *OWL Web ontology language reference.* W3C Recommendation, 10.

Decker, S. (2006). The social Semantic desktop: Next generation collaboration infrastructure. *Information Services & Use*, *26*(2), 139–144.

Feier, C., Roman, D., Polleres, A., Domingue, J., Stollberg, M., & Fensel, D. (2005). Towards intelligent Web services: The Web service modeling ontology (WSMO). *International Conference on Intelligent Computing (ICIC)*.

Fielding, R. T., & Taylor, R. N. (2002). Principled design of the modern Web architecture. *ACM Transactions on Internet Technology*, *2*(2), 115–150. doi:10.1145/514183.514185

Greaves, M. (2007). Semantic Web 2.0. *IEEE Intelligent Systems*, *22*(2), 94–96. doi:10.1109/MIS.2007.40

Gruber, T. (2007). *Collective knowledge systems: Where the social Web meets the Semantic Web*. To appear in Journal of Web Semantics, 2007

Kiss, J. (2008). Web 3.0 is all about rank and recommendation. *The Guardian*, February 4 2008,

Kopecký, J., Vitvar, T., Bournez, C., & Farrell, J. (2007). SAWSDL: Semantic annotations for WSDL and XML schema. *IEEE Internet Computing*, 60–67. doi:10.1109/MIC.2007.134

Lassila, O., Swick, R., et al. (1999). Resource description framework (RDF) model and syntax specification. *W3C Recommendation, 22*, 2004-03.

Martin, D., Paolucci, M., McIlraith, S., Burstein, M., McDermott, D., McGuinness, D., et al. (2004). Bringing Semantics to Web Services: The OWL-S Approach. In *Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004)*, (pp. 6-9).

McIlraith, Sheila A., Cao Son, Tran, & Zeng, Honglei. (2001). Semantic Web Services. *IEEE Intelligent Systems*, *16*(2), 46–53. doi:10.1109/5254.920599

Moller, K., Bojars, U., & Breslin, J. (2006). Using Semantics to enhance the blogging experience. In I*3rd European Semantic Web Conference (ESWC2006), LNCS, 4011*, 679-696.

O'Reilly, T. (2005). *What is Web 2.0: Design patterns and business models for the next generation of software.*

Prudhommeaux, E., & Seaborne, A. (2008). *SPARQL query language for RDF. W3C Recommendation 15 January 2008.*

Ramakrishnan, R., & Tomkins, A. (2007). Toward a peopleweb. *Computer*, *40*(8), 63–72. doi:10.1109/MC.2007.294

Schaert, S. (2006). IkeWiki: A Semantic Wiki for collaborative knowledge management. In *Proceedings of the 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*, (pp. 388-396).

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor. Paperback.

SWSI. (2004). *Semantic Web services initiative* (SWSI).

Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The Knowledge Engineering Review*, *11*(2).

Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., & Studer, R. (2006). Semantic Wikipedia. *Proceedings of the 15th international conference on World Wide Web*, (pp. 585-594).

KEY TERMS

Semantic Web (SW): The creator of the World Wide Web Tim Berners-Lee defines the SW as "a web of data that can be processed directly or indirectly by machines". A similar definition coming from the World Wide Web Consortium (W3C) describes the Semantic Web as "a Web that includes documents, or portions of documents, describing explicit relationships between things and containing Semantic information intended for automated processing by our machines."

Social Web/Web 2.0: The way people socialize and interact with each other through the World Wide Web. This term is also used to denote a large number of Web sites that are devoted to people and their social interactions through the creation of online communities of users that share digital content and information, discuss, or enable communication in any possible, Web-facilitated way.

Web Service (WS): A Web Service is defined by the World Wide Web Consortium (W3C) as "a software system designed to support interoperable machine-to-machine interaction over a network". Since this definition is quite general the term "Web Service" commonly refers to systems that communicate using XML messages that comply with the SOAP messaging format. In such systems, there is often machine-readable description of the operations offered by the service written in the Web Services Description Language (WSDL).

ENDNOTES

1. Despite the irony in itself as a fact, an extensive survey of this criticism can be found in the Wikipedia at http://en.WikipedSia.org/Wiki/Criticism_of_Wikipedia
2. http://en.Wikipedia.org/Wiki/Seigenthaler_incident
3. http://en.Wikipedia.org/Wiki/Ajax_(programming)
4. http://en.Wikipedia.org/Wiki/Comet_(programming)
5. http://en.Wikipedia.org/Wiki/Web_syndication
6. http://www.opensearch.org
7. http://pipes.yahoo.com
8. http://openid.net/

This work was previously published in Virtual Community Practices and Social Interactive Media: Technology Lifecycle and Workflow Analysis, edited by D. Akoumianakis, pp. 207-225, copyright 2009 by Information Science Reference (an imprint of IGI Global)

Chapter 1.4

Semantic Web in Ubiquitous Mobile Communications

Anna V. Zhdanova
The Telecommunications Research Center Vienna, Austria

Ning Li
University of Surrey, UK

Klaus Moessner
University of Surrey, UK

ABSTRACT

The world becomes ubiquitous, and mobile communication platforms become oriented towards integration with the web, getting benefits from the large amount of information available there, and creation of the new types of value-added services. Semantic and ontology technologies are seen as being able to advance the seamless integration of the mobile and the Web worlds. We provide background information on the Semantic Web field, discuss other research fields that bring semantics into play for reaching the ontology-enabled ubiquitous mobile communication vision, and exemplify the state of the art of ontology development and use in telecommunication projects.

DOI: 10.4018/978-1-60566-028-8.ch011

INTRODUCTION

Nowadays, mobile and Web environments converge in one shared communication sphere. Technologies stemming from Semantic Web and Mobile Communication fields get combined to achieve this convergence towards the vision of ontology-enabled ubiquitous mobile communication. Knowledge Management and Semantic technologies fields produce ways to describe, specify and manage information in a machine processable form, in particular, acquire, evolve, reuse, and combine knowledge (Fensel, 2001). Certain formats and protocols stemming from these fields are already being applied to telecommunications: vCard[1], CC/PP[2], UAProf[3]. However, these specifications are only applicable to a limited number of telecommunication scenarios, and management of information about resources in mobile environment could be substantially improved, e.g., by alignment of heterogeneous information sources in knowledge-based service enablers.

Ontologies and architecture knowledge layers play an ever-increasing role in service platforms and mobile communications. As integration of Telco, Internet and the Web takes place, in order to achieve interoperability, telecommunication systems and services tend to rely on knowledge represented with the use of shared schema, i.e., on ontologies similar to as envisioned on the Semantic Web (Tarkoma et al., 2007). However, specific ontology-based implementation solutions for mobile systems are rare, and best practices for such interoperability are not established. In this chapter, we address a problem of ontology-based interoperation in order to integrate independent components in a system providing value-added mobile services.

We present the overall state of the art ontology-related developments in mobile communication systems, namely, the work towards construction, sharing and maintenance of ontologies for mobile communications, reuse and application of ontologies and existing Semantic Web technologies in the prototypes. Social, collaborative and technical challenges experienced in the project showcase the need in alignment of ontology experts' work across the mobile communication projects to establish the best practices in the area and drive standardization efforts. We indicate certain milestones in integration of Semantic Web-based intelligence with Mobile Communications, such as performing ontology construction, matching, and evolution in mobile service systems and alignment with existing heterogeneous data models.

The chapter is structured as follows. In Section 2 we provide a motivation for discussing the convergence between the areas of Semantic Web and ubiquitous mobile communications. Section 3 gives an overview of the core ontology technologies involved, related and relevant research and development fields and challenges in the area. In Section 4, two illustrative case studies for the converged area are described. Section 5 concludes the chapter and Sections 6 indicates future research directions.

WHY SEMATICS IN UBIQUITOUS MOBILE COMMUNICATIONS?

In this section we motivate why combination of Semantic Web technology with ubiquitous mobile communications is beneficial. Semantic technologies in mobile communication have been somewhat considered to the less extent comparing to other fields, such as semantics in e-sciences, e-government, e-enterprise, e-communities, etc. However, as the mobile world starts to integrate with the Web world in delivering new value-added services, the area of semantics ubiquitous mobile communication inevitably gains a larger importance and potential.

Ubiquitous computing, also referred to as pervasive computing, is the seamless integration of devices into the users every day life. Applications should vanish into the background to make the user and his tasks the central focus rather than computing devices and technical issues (Weiser, 1991). When applying to mobile communication scenarios, ubiquitous computing can be viewed as when user moves around and changes circumstances, he can always be connected and well served without being aware of the technical issues under the scene. To achieve the goal, information from all the involving participants, such as user, network, service provider etc., needs to be collected, shared and interoperable with each other, known by one or more operational agents but agnostic to the user. Such information is diverse in their language, format and lack of semantic meaning for autonomous processing by computer or operational agent. The Semantic Web can be a rescue with its vision to achieve global information sharing and integration.

An example of combination of the two fields can be a service enabler that could be used by other services and thus make their construction simpler. For example, such an enabler could access distributed information on user's location, availability and friends and inform other services about which groups of friends are available and located at the same place. Services that assist with scheduling business meetings or parties, or the ones that are targeted at selling products for groups of users can be among the services that need such information from the enabler. To achieve the output, the enabler would take external structured data represented in a formal way (e.g., in RDF): for instance, information about user's availability from PIDF files, information on who is a friend of whom from FOAF profiles, information about location from data stemming from such standards as IETF RFC4119, RFC4589. Then the enabler would combine the gathered information, apply certain rules to deduce the result and pass it to other mobile services via an interface.

GROUNDING FOR SEMANTICS IN UBIQUITOUS MOBILE COMMUNICATIONS

In this section, we describe existing developments relevant for the combined field of ubiquitous mobile communications and the Semantic Web.

Ubiquitous Mobile Communication Existing Developments

A lot of work has been undertaken to implement the vision of ubiquitous mobile communications by investigating the underlying technologies. Examples include, but not limited to, user-related context collection, such as sensor network, Bluetooth, GPS, user-related context modelling and transmission, such as CC/PP, UAProf, 3GPP, MPEG-21, multimedia content description, such as MPEG-7, service discovery and service context modelling, such as UPnP, Jini, Bluetooth SDP, agent technologies, such as FIPA, ACL. Network mobility, security and QoS management, such as Hierarchical mobile IP, IPSec, DiffServ etc. The combination of these technologies with Semantic Web has become an inevitable trend in a ubiquitous mobile communication environment. Some of such joint developments are presented later in Section: Relevant research fields.

Semantic Web Existing Developments

Existing developments of the Semantic Web include languages (i.e. core formalisms to specify domain knowledge or services), methodologies and tools. In this section we outline the major developments in these areas and indicate their role and contributions in the area of Semantic Web enabled mobile platforms.

Languages and Formalisms

RDF(S)

RDF (Lassila & Swick, 1999; Manola & Miller, 2004) became a W3C recommendation in 1999. It is a general-purpose language for representing resources on the web in terms of named properties and values (McBride, 2004). With RDF it is not possible to define the relationships between properties and resources. For this purpose, RDF Schema (Brickley & Guha, 2004) has been specified. It became a W3C recommendation in 2004 and is basically an extension of RDF. More specifically, it is a formal description language for eligible RDF expressions. In particular, a schema defines the kinds of properties available for resources (e.g., title, author, subject, size, colour, etc.) and the kind of resource classes being described (e.g., books, Web pages, people, companies, etc.). RDF Schema is a simple ontology and a simple ontology definition language. RDF and RDF Schema are usually denoted RDF(S).

RDF(S) bases on some syntactical principles of XML (e.g. URIs) and has been equipped with an XML syntax as well. The most basic Semantic Web language which provides the syntactical basis for all other Semantic Web languages is RDF(S). RDF(S) is not provided completely with a formal logical semantics, thus reasoning is only on partially supported.

Topic Maps

Topic Maps are a data modelling language and became an ISO standard (ISO/IEC 13250) in 2000. A Topic Map offers a means to create an index of information which resides outside of that information. It describes the information in documents and databases by linking into them using URIs. A Topic Map consists of topics, associations (relationships between topics), occurrences (information resources relevant to a topic). Topics and occurrences can by typed. Types in Topic Maps are themselves topics and thus there is no real difference between a topic and a type.

There exists SGML, XML and RDF language support for Topic Maps. However, they are very simple and do not have a formal semantics and thus no sophisticated inference support. Nevertheless, because of their simplicity, they are often used in industry applications.

OWL

OWL (Dean & Schreiber, 2004) became a W3C recommendation in 2004. OWL is mainly based on OIL and DAML+OIL, which are obsolete Semantic Web languages and therefore not mentioned further here. OWL is equipped by an RDF syntax and includes three sub languages:

OWL-Lite roughly consists of RDF(S) plus equality and 0/1-cardinality. It is intended for classification hierarchies and simple constraints. OWL-Lite corresponds semantically to the formal Description Logic *SHIF(D)* and cannot express the whole RDF vocabulary.

OWL-DL contains the language constructs of OWL-Lite. OWL-DL corresponds semantically to the Description Logic *SHOIN*(D). Although strictly more expressive than OWL-Lite, it still provides computational completeness and decidability.

OWL Full does not correspond to a formal logic anymore as it builds upon the complete RDF(S) vocabulary which also lacks a correspondence to a formal logic. The language incorporates maximum expressive power and syntactic freedom, but offers no computational guarantees.

Semantic Web Languages in Progress

In this subsection, we consider Semantic Web languages which have been submitted to the W3C and thus have communities promoting them. At least some of them can be expected to become W3C recommendations. Examples of such languages are:

- *Languages based on the Logic Programming Knowledge Representation paradigm*: The trend to the aforementioned paradigm exists already since the year 2000 when the development of RuleML[4] has started. RuleML is a set of languages revolving around the Logic Programming paradigm and being equipped with an RDF syntax. Other examples of Semantic Web Languages with Logic Programming semantics are WRL[5], a set of three layered rule languages of increasing expressivity, and SWRL[6], a language which combines OWL and RuleML but is computationally intractable. Furthermore, a W3C working group[7] has been formed for establishing standards for Semantic Web rule languages.

- *Semantic Web Service Modelling Languages*: Semantic Web Services will play an important role in the Semantic Web as they combine Web Services with semantics. Examples for Semantic Web Services Languages are WSML[8] and SWSL[9]. The languages serve for the specification of ontologies describing Semantic Web Services. E.g., WSML is used to describe WSMO[10] and SWSL is used to describe SWSO[11].

Ontologies and Tools

Apart from the languages to describe data, specific ontologies related to the mobile communication domains and appropriate ontology management tools are necessary to implement the vision of ubiquitous mobile communications.

In a nutshell, the state of the art in development of the ontologies addressed by ubiquitous mobile communications comprises:

- Ad-hoc small-size schemata on certain general purpose topics are specified in ontology languages.
- Detailed, XML-based standards on certain narrow telecommunications topics.

Certain standardisation schemata and activities to be considered for the development of the ontology framework for mobile communications are listed later in Section: Mobile Ontology. Additional efforts coming from the Semantic Web community are listed in Table 1.

Typically, existing ontology management tools are adopted and explored in the semantic telecommunications projects. Stemming from the SPICE project (Zhdanova et al., 2006), examples of relevant and popular ontology management tools used, as well as encountered problems in their exploitation and an expected resolution times, are provided in Boxes 1 and 2.

As the ontology data are processed within the ubiquitous mobile communication applications, exploration and reuse of further ontology management technology is on the roadmap of the research and development field.

Major Challenges and Approaches

In this section, we define major challenges that are under development or need to be developed in the joint area. From ubiquitous mobile communication point of view, the challenges can be viewed from three perspective, i.e., from the user, the network operator, and the service provider. From user's point of view, there is an expectation of autonomous, non-stop service being provided with satisfying quality whatever terminal he/she uses, whenever he/she needs and wherever he/she goes without having to set the configuration. From network perspective, there is a challenge of ensuring the service delivery by providing a smooth handover, a guaranteed QoS and security level etc. when delivery circumstance changes, for example, from one type of network to another. From service provider point of view, the anticipation is to provide only one version of service, which, however, can be used by any device via any networks. To face these challenges, when designing a ubiquitous service delivery based on ontologies, the information from and about the user, network condition, the content/service being provided, together with any context information (to support personalization and service push to the user), need to be described in an unambiguous and interoperable manner in order to perform effective service delivery. Therefore, common ontologies describing such domain knowledge, as well as best practices in their reuse are required. Generally, challenges faced from ontology point of view include:

- **Heterogeneity:** Resolving inconsistencies, format differences (syntactic and semantic

Table 1. Ontologies related to the mobile ontology

Ontology Name	Producer	URL	Description	Development Status
MeNow	Chris Schmidt	http://crschmidt.net/foaf/menow/menow.rdf	The motivation for the MeNow schema is to be able to describe a variety of aspects of the current status of someone, either online or off, in a way that the data can be easily aggregated or retrieved. This schema allows the definition of a variety of terms that would be common in many applications: describing the current book you are reading, music you are listening to, mood you are in, and more.	
Pervasive SO – Describing User Profile and Preferences	Harry Chen, UMBC	http://pervasive.semanticweb.org/doc/ont-guide/part1/	Pervasive Computing Standard Ontology (PERVASIVE-SO) is a set of RDF/OWL ontology documents. Each ontology document is identified by a unique XML namespace and defines the ontologies of a specific domain. In a pervasive computing environment, computer systems often need to access the profiles and the preferences of a user in order to provide services and information that are tailored to the user. The profile of a user includes typical contact information (telephone numbers, email addresses, name, etc.) and information that describe other computing entities that can act on the behalf of the user (e.g., the personal agent of a user). The preference of a user is a description of the environment state that the user desires the computer systems to honor or achieve whenever it is possible.	Frozen in 2004
Platform for Privacy Preferences	Brian McBride, HP	http://www.w3.org/TR/p3p-rdfschema/	The Platform for Privacy Preferences Project (P3P) enables Web sites to express their privacy practices in a standard format that can be retrieved automatically and interpreted easily by user agents. P3P user agents will allow users to be informed of site practices (in both machine- and human-readable formats) and to automate decision-making based on these practices when appropriate. Thus users need not read the privacy policies at every site they visit.	Frozen in 2002
Gadget	Morten Frederiksen	http://www.wasab.dk/morten/2004/10/gadget	Definitions of various terms related to (typically) electronic gadgets such as GPS receivers, cameras and mobile phones.	Frozen in 2004
ConOnto: Context Ontology	Mohamed Khedr	http://www.site.uottawa.ca/~mkhedr/contexto.html	ConOnto describes the different aspects of context-aware systems. ConOnto includes location, time, activities, software and hardware profiles. ConOnto also includes meta-information that describes negotiation and fuzzy ontologies to be used in systems that will negotiate and infer about context information.	
Ambient Networks: General, Cost, QoS Ontology	Anders Karlsson, TeliaSonera	http://kiwi.intra.sonera.fi/an_costs.owl		Created in 2005

Box 1.

Name……………………….. Protege
Website…………………….. http://protege.stanford.edu/
White page………………… n/a
Main characteristics……….. Ontology editor

Open problems	Relevance					Term		
	1 Very low	2 Low	3 Normal	4 High	5 Very high	0-3 short	3-6 medium	6-12 long
Needs improvement of usability features, robustness		X					X	

Box 2.

Name………………………. Jena
Website…………………….. http://jena.sourceforge.net/
White page………………… n/a
Main characteristics……….. A Semantic Web Framework for Java. Ontology API and implementation, supports RDF(S), OWL, performs basic ontology management and reasoning (similar idea as Xerces for XML)

Open problems	Relevance					Term		
	1 Very low	2 Low	3 Normal	4 High	5 Very high	0-3 short	3-6 medium	6-12 long
Scalability: works slowly on large volumes of data					X		X	

differences in formalization), business process mediation.

- **Versioning** when merging or combining ontologies representing different knowledge domains, ontology and instance data evolution and maintenance during updates.
- **Scalability:** Scalable repositories for ontology instance data (currently popular ontology management toolkits such as Jena and Sesame do not always meet industrial standards).
- **Ontology and instance data visualization**, user interfaces.
- **User/community generated content:** formalization, acquisition and employment (when building innovative mobile services in Web 2.0 style of social applications alike to YouTube, Flickr, LinkedIn, Google Base, etc.).
- **Semantically described mobile and web services** (no yet widely accepted "standard" solutions).
- **Service composition and discovery**, user-driven creation of new services by composing service enablers, mash-ups.
- **Integration with the non-semantic web services and formats**, which use traditional technologies such as WSDL, SOAP and UDDI, XML.
- **Integration with legacy applications**, which are not terminologically a service, but can be upgraded to be a service, e.g. MPEG codec.

Relevant Research Fields

The following research areas are related to, impact and will be potentially impacted by the described involvement of the Semantic Web in Mobile Communications.

Multimedia

Today, the amount of digital multimedia information is growing over the World Wide Web, in broadcast data streams and in personal and professional databases. One of the major challenges for ubiquitous mobile communication is to enable any mobile devices, e.g. mobile phone, PDA, to access, exchange and consume a rich set of multimedia content seamlessly over dynamic and heterogeneous networks. The need of semantic description of the multimedia information becomes apparent. The MPEG-7 and MPEG-21 are the dominant efforts for multimedia content and service description framework.

MPEG-7 is known as the multimedia description standard and offers several tools, i.e. Description Schemes, to annotate multimedia content at different levels. The main parts are: Description Definition Language (DDL), Visual, Audio and Multimedia Description Schemes (MDS). The DDL is a language that allows the creation of new Description Schemes and, possibly, Descriptors. It also allows the extension and modification of existing Description Schemes. The DDL is based on XML Schema Language, but with MPEG-7 extensions specifically for audiovisual description. The Visual description tools provide structures to describe basic visual features, such as color, texture, shape, motion and localization etc. The Audio description tools provides structures for the description of audio features that are common across many applications, such as spectral, parametric, and temporal features, and that are application-specific features, such as audio indexing, recognition and signature. MPEG-7 Multimedia Description Schemes provides the description tools for generic media entities, such as vector, time and more complex media entities. The latter can be grouped into 5 different classes according to their functionality: Content description, Content management, Content organization, Navigation and access and User interaction.

MPEG-21, the 21st century multimedia framework, goes further and provides tools to describe the environment to enable transparent multimedia creation, delivery and consumption between heterogeneous environments. The main parts are Digital Item Declaration (DID), Digital Item Identification (DII), Intellectual Property Managements and Protection (IPMP), Rights Expression Language (REL), Rights Data Dictionary (RDD) and Digital Item Adaptation (DIA). The Digital Item Declaration (DID) specification contains three normative sections, a model to describe a set of abstract terms and concepts to form a useful model for defining Digital Items, a representation to describe the syntax and semantics of each of the Digital Item Declaration elements, and a Schema comprising the entire grammar of the Digital Item Declaration representation in XML. The DII specification provides mechanisms to uniquely identify Digital Items, Intellectual Property related to the Digital Items such as abstractions, Description Schemes and types of Digital Items. IPMP is an extended efforts based on MPEG-4 to develop new systems and tools with enhanced interoperability. The REL, together with the RDD that supports the REL and provides extensive semantics, provides a universal method for specifying rights and conditions associated with the distribution and use of digital items and thus facilitates the creation of an open DRM architecture. DIA provides tools to describe the Digital Item usage environment including: Usage characteristics, such as user info, usage history, User preferences and physical characteristics such as disabilities, Device characteristics such as display, memory and battery, Network characteristics, such as error characteristics and bandwidth, and Natural environment characteristics such as noise and illumination. This is to facilitate transparent access to distributed digital items by shielding users from the technical complexity, such as network and terminal installation, management and implementation issues.

Web Service (SOA)

Service-Oriented Architecture (SOA), especially a Web service-based SOA, has the potential in speeding up the application development process and the agility in responding to the change of business needs. This is due to the loose coupling of client from service and the set of standard protocols and technologies used by the Web service, such as XML, WSDL, SOAP, and UDDI. The inherent features of SOA, i.e. reusability, interoperability, scalability and flexibility, can virtually meet the requirement of a supportive framework for ubiquitous mobile communication.

Coming along with Semantic Web is Semantic Web service, where Web service is described with added computer-processable semantics, and thus a number of services can be concatenated autonomously to compose a new service for a more complex task. This will benefit service provision in a ubiquitous environment where all information from user, network, together with the requested service and any intermediate service are required to make a delivery decision autonomously. In accordance with this advance, the set of standard protocols and technologies for Web services are evolving to reach their semantic counterparts or brand new standards are created. For example, the ontology languages, RDF, RDFS, and OWL are developed to add computer-processable semantics on top of the exiting syntax provided by XML. WSDL specifies a way to describe the abstract functionalities of a Web service and concretely how and where to invoke it. Semantic Annotations for WSDL (SAWSDL)[12] defines mechanisms using which semantic annotations can be added to WSDL components based on an earlier effort, namely WSDL-S[13], which adds semantic expressivity to the service description by extending original WSDL elements. SOAP, as the message exchange protocol and originally XML-based, can be combined with RDF and OWL in order to introduce semantics to assist the flexible service invocation (Zhao, 2004). Similarly, enabling UDDI to store semantic markup and handle semantic enquiries has been investigated in recent years (Luo et al., 2006). Correspondingly, service-oriented architectures require machine-processable semantics to achieve its full potential. In particular, DERI[14], targets this challenge by offering a set of tools and techniques ranging from ontology construction to description language. The later include Web Service Modelling Ontology (WSMO), Web Service Modelling Language (WSML) and a Web Service Execution Environment (WSMX)[15].

Security, Privacy, Trust

In recent years, the security and trust aspects of Web services are standardised by OASIS[16] with a WS-security specification released in 2004. WS-security insets security-related information to Web service messaging that provides for message integrity and confidentiality using security token and digital signatures. The use of Semantic Web technologies enables Web into a genuinely distributed and global content and service provider. Inherent with this are the issues of more widespread security, trust, information quality and privacy.

To achieve a security solution for Semantic Web service, the traditional security solutions can be described as one of the contextual information attached to the service and can be interpreted on the other end at semantic level. This solution has the advantage of not requiring to design a bottom-up Semantic Web service security architecture and thus provides the service provider with flexibility of control. However, embedding security and trust policies into every Web service may not appear to be an attractive solution and can result in tight coupling between services and particular security implementations. An alternative is to design an integrated Semantic Web security framework with security mechanisms available at various layers of the network. This may provide a comprehensive solution when more security and trust issues and challenges arise from the traditional communication domain. For example, when seamless interconnecting heterogeneous networks, particularly when security issue are jointly considered with other issues, such as QoS and mobility management, in the overall communication process.

Human Communication Interface

The design of the Human Communication Interface (HCI) for Web applications has always been of great importance. The Web technology is evolving with Semantic Web, the interaction must also evolve. With the emergence of the Semantic Web and Semantic Web services, which give information well-defined meaning and enable computer work in cooperation with humans, the interactions with Web-delivered information has become possible and thus the complexity of the human communication interface has increased. For example, instead of being an information receiver only, user can interact with information to construct new information and build knowledge. In addition, with the user terminal getting smaller and smaller in size like PDA, Pocket PC, together with the service being more customized to the user's personal need and preference, the human interface design becomes even more challenging than ever.

In ubiquitous mobile communication environment, human computing interfaces form one of the major contextual information of the user as well as one of the major component in the delivery path. Therefore, it is essential to bring interaction design principles with other contextual information into the semantically structured information in order to facilitate this evolution of the Web. On the other hand, with the popularity of Semantic Web and wide acceptance of its technology in integrating knowledge from various heterogeneous parties, embedding Semantic Web technology into the HCI design is envisioned to be a necessity for automating the deployment of rich user interfaces.

Lower Layer of Mobile Communication

One of the important contextual information to be included, and maybe semantically described, when customizing Semantic Web content and service to the user ubiquitously are the network conditions in the delivery path, e.g. bandwidth, QoS, Security, latency, jitter etc. Signalling and/or information exchange is required between the application layer and the underlying network layers for request and response. Enhancements to the communication protocols with semantic capabilities are envisaged to be required in order to assure user's satisfaction in ubiquitous service delivery. In addition, there have been extensive efforts to tackle the network layer integration of QoS, security and mobility management over heterogeneous network in a mobile environment.

Autonomous Computing

The vision of autonomous computing is to enable computing system operate in a fully autonomous manner. The challenges of autonomous computing are robustness, simplicity for the end-user and seamless integration (Hercock, 2002). With the vision of being a global information integrator by making information computer-interpretable, Semantic Web technologies can help realizing the vision of autonomous computing, particularly in a ubiquitous mobile communication environment where constant changes take place and autonomous process are expected.

Grid, Semantic Grid

Grid development is targeted at the problem of efficiently using computing power of distributed resources for achieving a common set of tasks.

Typical Grid toolkits include GLOBUS[17]. Semantic Grid is oriented towards enhancing typical Grid services or processes with ontology-based descriptions (Goble et al., 2004). Different types of Grid resources can be used by ontology–enabled mobile services to reach their goals.

ONTOLOGY FRAMEWORK STUDIES

In this section, we describe ongoing research and development combining the areas of ubiquitous mobile communication and Semantic Web. One can view convergence of Semantic technologies with mobile communication from two sides: inclusion of the Semantic technologies in solutions delivered by mobile communication project from one side; and mobile communication use cases in core Semantic technology projects from the other side. On the one hand, work on development of mobile communication applications, enablers and services with involvement of ontologies has been carried out in the following mobile communication projects: SPICE[18], MobileVCE[19], Mobilife[20], OPUCE[21], Ambient Networks[22], etc. On the other hand, research and development involving mobile aspects has been carried out in the following Semantic Web projects: TripCom[23], SWING[24], ASG[25], SmartWeb: Mobile Access to the Semantic Web[26], etc. In this section, we provide detailed illustrating examples of the ontology work carried out for mobile communication solutions, specifically, mobile ontology (SPICE project) and an ontology solution for multimedia (MobileVCE project).

Mobile Ontology

Mobile environments and the Web converge forming a shared Distributed Communication Sphere (DCS). This causes the appearance of new settings to be supported, e.g., when the user utilizes mobile and fixed devices to interact with systems. Interaction and connectivity of mobile applications with the Internet increase. To ensure interoperation of mobile and Web services, applications and tools (running on heterogeneous various service platforms in such a sphere), developers need to have a shared specification of objects belonging to the sphere and their roles. Certain ontologies have already been developed for the mobile communication domain by employing area with employment of Semantic Web formalisms (Korpipää et al., 2004; Pfoser et al., 2002). However, widespread and global adoption of such ontologies remains a challenge.

Approaching the problem of interoperation between the Web and mobile service technologies, Mobile ontology, a comprehensive "higher-level" ontology for mobile communication domain, is being developed. Currently, definition and implementation of the Mobile ontology is managed as a collaborative effort amongst participants of the EU IST SPICE Integrated Project.

Mobile Ontology Introduction

What's Mobile Ontology for?

Mobile Ontology is being developed as a comprehensive "higher-level" ontology for mobile communication domain. The ontology is a machine readable schema intended for sharing knowledge and exchanging information both across people and across services/applications, and it covers domains related to mobile communications, specifically, addressing persons, terminals, services, networks.The added values of Mobile Ontology are:

- Providing an easy and formal way to reference objects from the mobile communication domain (in particular, to serve as an exchange format between mobile service enablers).
- Providing an opportunity to implement enhanced, ontology-based reasoning.
- Providing a formal representation of the

domain to be used in research and development projects, and for educational purposes.

Mobile Ontology Overview

DCS-related vocabulary terms, grouped in broad categories, are presented in Figure 1.

Mobile Ontology, in particular, its DCS Vocabulary (Zhdanova et al., 2006) definitions are written using RDF and OWL (Manola & Miller, 2004; Dean & Schreiber, 2006) that makes it easy for software (both Web-based and mobile-oriented) to process facts about the terms in the DCS vocabulary, and consequently about the things described in DCS documents. A DCS document/instance data can be combined with other DCS documents to create unified sources of information.

Example

A very basic annotation describing the state of the communication model is:

```
<?xml version=”1.0”?>
<rdf:RDF
xmlns:rdf=”http://www.
w3.org/1999/02/22-rdf-syntax-
ns#”
xmlns:xsd=”http://www.
w3.org/2001/XMLSchema#”
xmlns:rdfs=”http://www.
w3.org/2000/01/rdf-schema#”
xmlns:owl=”http://www.
w3.org/2002/07/owl#”
xmlns:p1=”http://www.owl-ontolo-
gies.com/assert.owl#”
xmlns=”http://www.ist-spice.org/
mobile_ontology/2006/5/26/mo-
bile_ontology.owl#”
xml:base=”http://www.ist-spice.
org/mobile_ontology/2006/5/26/
mobile_ontology.owl”>
…
<Device
rdf:ID=”BlueSonyEriccson”>
<belongsTo>
<Person rdf:ID=”AnnaZ”>
<owns rdf:resource=”#BlueSonyEri
ccson”/>
<rdfs:comment
rdf:datatype=”http://www.
w3.org/2001/XMLSchema#string”
></rdfs:comment>
</Person>
</belongsTo>
</Device>
…
</rdf:RDF>
```

RDF/S and OWL have been chosen as formats to represent the mobile ontology, as they are current recommendation ontology languages of W3C and have a relatively large tool support for implementation of enablers and applications.

Starting from the DCS ontology, the Mobile ontology has developed a new structure and evolved in Mobile ontology Core and Mobile subontologies. Mobile ontology Core comprises the telecommunications domain concepts and properties that occur most commonly and in various subdomains or application types. Mobile subontologies contain further details related to specific topics in telecommunications, and its items are linked to the items of the Mobile ontology Core. The Mobile ontology Core overview is depicted at Figure 2.

Currently the subontologies on the following topics are being represented in the Mobile ontology infrastructure: Profile, Service, Service Context, DCS, Service Roaming, Rules and recommendations, Presence, Location, and Content. The up-to-date ontology versions and the status of the work are represented at the Mobile Ontology website[27].

Figure 1. DCS-related Classes and Properties of Mobile Ontology

DCS Basics

- **CommunicationSphere**
- **AccessMethod**
- **Availability**
- **PhysicalLocation**
- **Modality**
- **InputModality**
- **OutputModality**
- **Cost**
- **Time**
- **AccessParameter**
- **consistsOf**
- **associatedWith**
- **characterizedBy**
- **providesAccessTo**
- **hosts**
- **PhysicalSpace**
- **Outdoors**
- **Indoors**
- **Building**
- **Roomspace**
- **Hall**
- **Floor**
- **Stairway**
- **isInsideOf**
- **hasInside**
- **isOnFloor**
- **isPhysicallyConnectedTo**
- **hasAboveFloor**
- **isAboveFloor**
- **isWithinRangeOf**
- **belongsTo**
- **owns**

Users and Groups, Personalization

- **Person**
- **UserGroup**
- **Mood**
- **UserRule**
- **CurrentActivity**
- **isUserDefaultActivity**
- **isWillingToCommunicate**
- **Recommendation**
- **RecommendedItem**
- **Feature**
- **hasWeight**
- **hasRelevanceScore**
- **hasSupport**
- **hasConfidence**
- **hasPreferred**
- **hasMeanRelevanceScore**
- **hasMeanSupport**
- **hasMeanConfidence**
- **Profile**
- **Subset**
- **hasType**
- **hasDescription**
- **hasSubset**

Devices

- **Terminal**
- **Device**
- **EmbeddedDevice**
- **SensorNode**
- **MultimodalityDevice**
- **ConfigurationElement**
- **DescriptionElement**
- **DataElement**
- **Datastore**
- **Module**
- **TerminalParameter**
- **hasParent**
- **hasModule**
- **hasAdditionalConfiguratio n**
- **hasAdditionalParameters**
- **hasAdditionalDescription**
- **hasData**
- **hasDescription**
- **hasConfiguration**
- **hasSize**
- **hasVersion**
- **hasName**
- **hasEncoding**
- **hasTStamp**
- **hasLocation**
- **hasDisplayName**
- **hasCategory**
- **hasMimeType**
- **hasVendor**
- **isEquippedWith**
- **isAvailableAt**

Services

- **Service**
- **QoS**
- **MultimodalityService**
- **InformationService**
- **KnowledgeService**
- **ContextService**

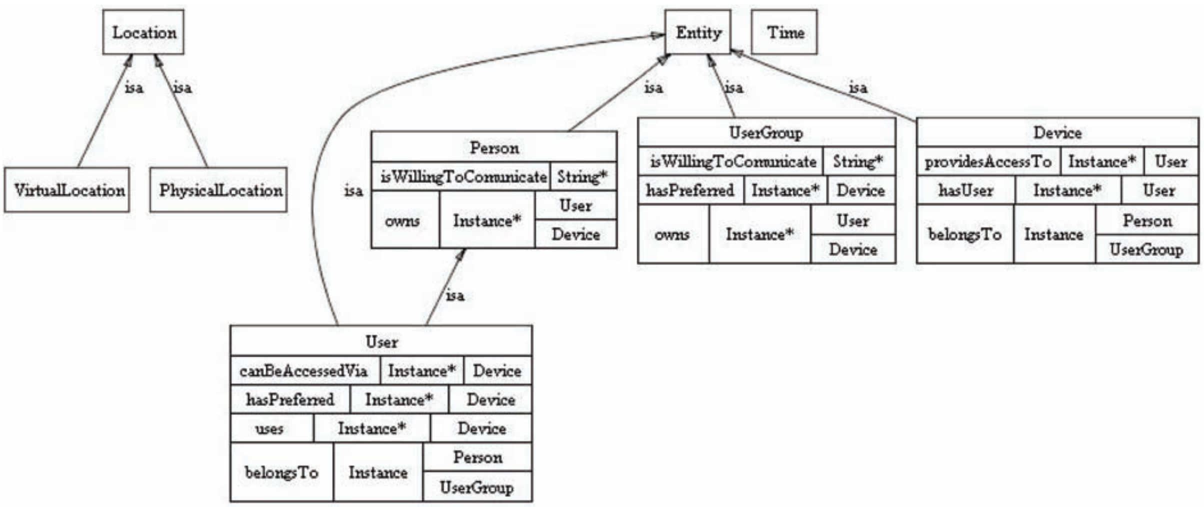
Networks

- **Network**
- **BluetoothNetwork**
- **WiFiNetwork**
- **GSMNetwork**
- **UMTSNetwork**
- **canBeAccessedVia**
- **isConnectedTo**
- **hostedBy**

Reuse of Schemata and Ontologies

Certain schemata covering the domains of the Mobile ontology exist and have already acquired significant communities. Such schemata can be either an output of the standardization bodies or coming in a "bottom-up" manner from companies and individuals and being widely accepted by the

Figure 2. Mobile ontology core visualization



masses. These schemata and ontologies can be specified in different knowledge representation formalisms. We address external sources represented via the most popular formats, namely OWL, RDF/S and XML/S.

Relating and mapping these schemata to the Mobile ontology is mainly beneficial for interoperation of the Mobile ontology community with other mobile communities. Thus:

- Mobile ontology developers and users benefit acquiring additional knowledge in the mobile communication domain captured in the existing OWL, RDF, XML schemas (i.e., reusing the present knowledge).
- Users of the related ontologies and schemas benefit from a straightforward mapping of their schemas to the Mobile ontology that enables a simpler move to/involvement or extension of the Semantic technologies for these communities.

Technically, two different approaches to combine the Mobile ontology with the existing common ontologies and schemata will be considered, depending on whether the data is encoded via an ontology language (such as PDF/S and OWL) or only via XML.

Approach 1: RDF/S or OWL Encoding

The following principles are valid when considering integration of Mobile ontology with ontologies of relevant topics expressed via RDF/S or OWL formalisms:

- When necessary directly reusing the agreed ontologies or their parts when modelling processes;
- Establishing and using the library of mappings of these ontologies with the "higher" level Mobile ontology classes and properties that have similar items as the used external ontology. Such a mapping library would not be re-modelling, but stating relations between items in a machine readable format. Equivalence, for example, can be stated using constructions "owl:sameAs" so that applications and enablers can "understand" that an item from the Mobile ontology and an "imported" agreed upon ontology are the same.

The RDFS and OWL-based standard schemata considered for this approach are listed in Table 2.

Approach 2: XML Encoding

The following principles are valid when considering integration of Mobile ontology with schemata of relevant topics expressed via XML formalisms:

- Re-modelling XML schemata in OWL and providing the new sub-ontologies as relatively independent ontology submodules under the umbrella of the Mobile ontology;
- Creation of the converters lifting up the instance data represented solely in the XML format to RDF.

So the ontology work with the existing XML schemas would focus on ontologizing/considering the knowledge present these schemas, and combining it with the Mobile ontology, and not extending these schemata.

The XML-based standard schemata considered for this approach are listed in Table 3.The following goals addressed by Mobile Ontology are open challenges for the current state of the art:

- The first comprehensive higher level ontology for mobile communication domain that is constructed with involvement/support of major players in mobile communication area, i.e. the ontology (i) responds to the needs of mobile service developers, (ii) is evolving, (iii) representatively captures the domain in an unbiased fashion.
- The most large scale international investigation on the use of Semantic technology

Table 2. OWL and RDFS -based relevant standards

Ontology name	Ontology Web address
UAProf	http://www.openmobilealliance.org/release_program/uap_v2_0.html
FOAF	http://www.foaf-project.org/
vCard	http://www.w3.org/TR/vcard-rdf

Table 3. XML-based Relevant Standards

Schema name	Schema Web address
Presence simple specification	http://www.openmobilealliance.org/release_program/Presence_simple_v1_0.html
Basic Presence Data model	http://www.ietf.org/rfc/rfc3863.txt
Generic Presence Data Model	http://www.rfc-editor.org/rfc/rfc4479.txt
Rich Presence Information	http://www.rfc-editor.org/rfc/rfc4480.txt
Location Types Registry	http://www.ietf.org/rfc/rfc4589.txt
A Presence-based GEOPRIV Location Object Format	http://www.ietf.org/rfc/rfc4119.txt

in mobile communication domain.

Collaboration Aspects

Mobile Ontology construction has been initially implemented within two major deliverables of SPICE EU project. 25 persons from 10 industry and research organizations stemming from 6 European countries have been initially involved in this specific cross issue. Therefore, apart from the definition of the up-to-date ontology infrastructure for Mobile services, the results of this study include observation of collaboration aspects of developers and users in the ontology construction.

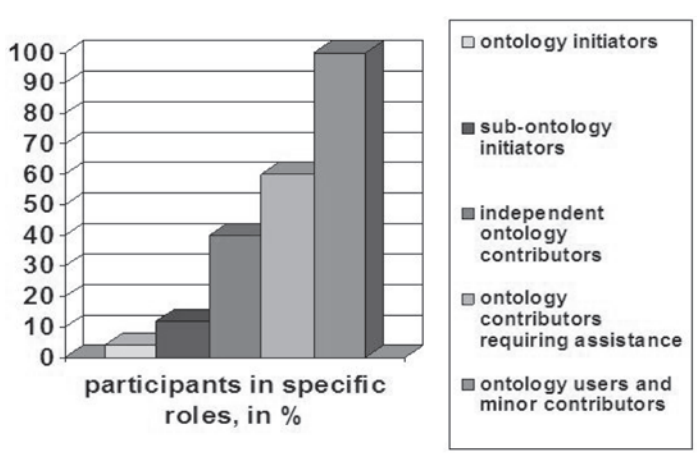
Here we show how involved parties collaborated on the ontology construction, and what personal involvement expectations for a larger scale ontology infrastructure would be. In Figure 3, the extent to which the developers have been typically involved in initial definition of the ontology is demonstrated. The figure shows that most contributors tend to provide minor extensions to the ontology or choose a role of the user, which also confirms the previous research (Zhdanova, 2006).Summarising, the main challenges identified in collaborative ontology construction as they have been observed are as follows:

- **Educational:** People with no/little knowledge on ontologies require *at least an introduction to the field*.
- **Methodology:** As of yet *no widely accepted or best practice solutions* on how to acquire ontologies from people in such a setting.
- **Basic technology:** Current ontology language standards (such as OWL) cause *confusion and awkward modelling solutions*.
- **Tool support:***Better tools for ontology construction process coordination, documentation* would help to avoid ad-hoc solutions and manual work.

Ontology-Based Multimedia

Nowadays, network access technologies, such as Bluetooth, WiFi, WiMAX, are bringing the dream of ubiquitous service provision closer to reality. Ubiquitous service provision is of interest to service providers, telecommunication operators and technology manufactures for their future revenue prospect. However, the barriers to the delivery of ubiquitous services arise from the desire to deliver a wide service mix to users having a wide range of access devices via a multitude of heterogeneous access networks with different preferences and likings. Most of the existing applications and services are created and provided

Figure 3. Collaboration patterns in mobile ontology construction



assuming a traditional pre-set delivery method and homogeneous transport media, which indicates a great importance and necessity for content and service adaptation in order to deliver them in such a ubiquitous environment. Such adaptation must be 'context-aware' and must facilitate user and situation specific content and service provision.

Therefore, the information from all the parties involved in the content and service delivery chain will form a contextual knowledge base that is shared by different operational agents in order to come up with a delivery or adaptation decision autonomously. These agents are generally heterogeneous in nature and distributed across networks. How to describe and represent such a knowledge base is fundamental to the development of a content and service adaptation framework which supports ubiquitous service delivery. This work has formed part of the Ubiquitous Services Core Research Programme of the Virtual Centre of Excellence in Mobile & Personal Communications, Mobile VCE.

Ontology has been recognized as the knowledge representation scheme and OWL as the knowledge representation language. However, to define a set of commonly-agreed vocabularies for the adaptation domain remains as a challenging issue. The ubiquitous content/service adaptation domain involves multiple sub-domains, i.e. the user domain, the content/service domain and the adaptation operation domain. Many efforts have been seen in recent years aiming to reach a description standard including vocabularies in order to achieve maximum acceptance and interoperability among communities. So far, the widely-acknowledged standards include usage environment description standards describing user information, device and network characteristics etc., such as CC/PP, UAProf and MPEG-21, and content description standards such as MPEG-7. Among those, MPEG-7 (ISO/IEC JTC1/SC29/WG11 N3752, 2000) and MPEG-21 DIA (ISO/IEC 21000-7, 2004) provide a good combination to linking content description with user environment description besides their well-established comprehensiveness in describing the respective domains. MPEG-7 offers several tools, i.e. Description Schemes (DS), to annotate multimedia content at different levels. These include Description Definition Language (DDL), Visual Schemes, Audio Schemes and Multimedia Description Schemes etc. MPEG-21 provides tools to describe the environment to enable transparent multimedia creation, delivery and consumption between heterogeneous environments. The most relevant part within MPEG-21 standard for the adaptation domain is Digital Item Adaptation (DIA). It provides tools to describe the user environment including: user characteristics, such as user info, preferences, usage history and physical characteristics, Device characteristics, such as display, memory and battery, Network characteristics, such as error characteristics and bandwidth, and Natural Environment characteristics such as noise and illumination.

Recent research efforts have reflected the recognition of using MPEG-7 and MPEG-21 DIA, together with ontology-based technologies, to construct an ontology to support the content/service adaptation (Soetens et al., 2004; Jannach et al., 2006). Though OWL has been chosen as the description language, Soetens et al. adopted limited usage of MPEG-21 vocabularies due to the immaturity of this standard at the time of writing (Soetens et al., 2004). In (Jannach et al., 2006), though MPEG vocabularies are adopted to form the domain ontology, the representation remains its original format of XML. With the actual adaptation operations being described in OWL-based language, this work realizes the integration of the different representation formats on the technical level using XML-based and logic-based technologies. Therefore, although MPEG-7 and MPEG-21 standards have been acknowledged for their strengths in multimedia domain description and delivery, their strengths can be still greatly enhanced by adding machine-processable semantics via ontology representation languages, such as OWL and RDF(S).

There exist several efforts to construct ontology representations of MPEG-7 and MPEG-21 (Hunter, 2001; Garcia, 2005). Those efforts construct ontology automatically by means of XSLT transformation according to the rules specified in (Garcia, 2005). By automatically converting the XML tree structure, the obtained ontology describes the relationship between the types of the tree element instead of describing the relationships between the semantics embodied by the tree elements. Although this approach expresses the XML-based standards in an OWL or RDF format, it does not add much semantic expressiveness to them. Such approach would be applied in any automatic XML schema to OWL conversion regardless of the semantics of the respective domain. In (Li et al., 2007), it argues that, for an expressive OWL representation of the XML-based standards, manual conversion is necessary. The manual conversion may result in some XML elements being discarded or treated with another XML construct as one OWL concept as a consequence of its semantic interpretation. There are no rules on how to manually convert an XML schema description into OWL ontology. Different from automatic conversion, which merely translates the XML syntax to OWL syntax, manual conversion has to examine the elements and the attributes of the XML schema, study their semantics, and translate them into OWL constructs.

CONCLUSION

State of the art and trends in convergence of the Semantic Web and mobile communication fields are presented in this article. Knowledge representation formalisms, relevant research fields, relevant ontologies are detailed, and the challenges of Semantic technology application to the mobile communications have been discussed. State-of-the-art examples of the work in this area have been outlined, including the development and use of ontology infrastructures that can serve as a semantic basis for applications and enablers within the convergence of the mobile and the Web worlds.

In a nutshell, one may conclude that (i) there exist a large number of ontologies addressing context-awareness and mobile communication issues, (ii) these ontologies are difficult to find and they are not or weakly linked and connected to each other. Factually, most of the time they do not form Semantic Web as the Web is about linking the data (and the users thus obtaining the typical Semantic Web benefits, such as interoperability), which is not the case for the current Semantic Mobile Communications.

FUTURE RESEARCH DIRECTIONS

The questions for the further research and development include: How to make these ontologies collaboratively constructed, linked to each other, easily found, used and evolved? And more specific, follow-up questions thus are:

1. How to involve developers and users in community-driven ontology construction and to what extent one should expect their involvement?
2. Which technical infrastructure is needed to implement this vision?

Ontologies are evolving as the domain is evolving and capturing the whole domain and all the needs by a (small) group of ontology developers alone is ineffective (Zhdanova, 2006). The ontology infrastructure for ubiquitous mobile communications should provide a user-friendly support to the ontology-based context-aware application and service developer with the following methods:

1. Key-word based ontology search (e.g., similar to OntoSelect[28], Swoogle[29])

2. Extraction and segmentation of the required ontology parts (e.g., operated on a level of triples and eventually on demand arranged in a stand-alone schemata)
3. In case no existing relevant ontology parts are found in the ontology platform infrastructure, possibility to plug in freshly developed ontologies and extensions
4. Simple ontology instantiation, with a subsequent simple discovery and use of the instance data
5. Ontology matching and alignment to the existing ontologies in case of duplicated modeling discovery, e.g., in a community-driven manner (Zhdanova & Shvaiko, 2006)
6. Search of relevant data within the instances (e.g., employing technologies such as YARS (Harth et al., 2006)
7. In case the developer made new ontology design/extensions, allow him/her easily plugging in the evolved versions of his/her ontologies/extensions, keeping up with the agreed ontology versioning practices

In conclusion, the starting points for the integration of the fields of Semantic Web and ubiquitous mobile communications exist both on the ontology schemata level as well as on the tools level. However, practices and processes for the common usage of the technologies originating from these two fields are still to be acquired. It is expected, that the new practices and processes are also to influence the future development of applications, services and tools appearing in the unified field.

REFERENCES

Brickley, D., & Guha, R. V. (Eds.). (2004 February). *RDF Vocabulary Description Language 1.0: RDF Schema*. W3C Recommendation.

Dean, M., & Schreiber, G. (Eds.). (2004 February). *OWL Web Ontology Language Reference*. W3C Recommendation.

Fensel, D. (Ed.). (2001). *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*, Springer-Verlag.

Garcia, R., & Celma, O. (2005). Semantic integration and retrieval of multimedia metadata. In *Proceedings of the 5th International Workshop on Knowledge Mark-up and Semantic, Annotation at the 4th International Semantic Web Conference*.

Goble, C. A., De Roure, D., Shadbolt, N., & Fernandes, A. A. A. (2004). Enhancing services and applications with knowledge and semantics. In Foster, I. and Kesselman, C. (Ed.). *The Grid: Blueprint for a New Computing Infrastructure* (2nd. Ed., Ch. 23, pp. 431-458). Morgan Kaufmann.

Harth, A., Umbrich, J., & Decker, S. (2006). MultiCrawler: A pipelined architecture for crawling and indexing Semantic Web data. In *Proceedings of the 5th International Semantic Web Conference*, Athens, GA, USA.

Hercock, R. G. (2002). Autonomous Computing. In *Proceedings of the Workshop on Grand Challenges for Computing Research*, Edinburgh, UK.

Hunter, J. (2001). Adding multimedia to the semantic Web-building and MPEG-7 ontology. In *Proceedings of the 1st International Semantic Web Working Symposium*.

ISO/IEC 21000-7, (2004). *Information Technology—Multimedia Framework (MPEG-21)—Part 7: Digital Item Adaptation* .

ISO/IEC JTC1/SC29/WG11 N3752, (2000). *Overview of the MPEG-7 Standard (version 4.0)*.

Jannach, D., Leopold, K., Timmerer, C., & Hellwagner, H. (2006). A knowledge-based framework for multimedia adaptation. [Special Issue on Innovations in Applied Artificial Intelligence, Springer.]. *The International Journal of Artificial Intelligence*, *24*(2), 109–125.

Korpipää, P., Häkkilä, J., Kela, J., Ronkainen, S., & Känsälä, I. (2004). Utilising context ontology in mobile device application personalization. *ACM International Conference Proceeding Series; Vol. 83 Proceedings of the 3rd international conference on Mobile and ubiquitous multimedia*, pp. 133-140, ACM Press.

Lassila, O., & Swick, R. (1999). *Resource Description Framework (RDF) Model and Syntax Specification*. W3C Recommendation.

Li, N., Attou, A., & Moessner, K. (2007). A MPEG-based ontology for ubiquitous content/service adaptation. In *Proceedings of the 3rd Workshop on Context Awareness for Proactive Systems (CAPS'2007)*, Guildford, UK.

Luo, J., Montrose, B., Kim, A., Khashnobish, A., & Kang, M. (2006). Adding OWL-S support to the existing UDDI infrastructure. In *Proceedings of IEEE International Conference on Web Services (ICWS'06)* pp. 153-162.

Manola, F., & Miller, E. (2004). *RDF Primer*. W3C Recommendation.

Pfoser, D., Pitoura, E., & Tryfona, N. (2002). Metadata modeling in a global computing environment. *GIS'02*, November 8-9, 2002, McLean, Virginia, USA, ACM Press (2002).

Soetens, P., Geyter, M. D., & Decneut, S. (2004). Multi-step media adaptation with semantic Web services. In *Proceedings of International Semantic Web conference.*

Tarkoma, S., Prehofer, C., Zhdanova, A. V., Moessner, K., & Kovacs, E. (2007). SPICE: Evolving IMS to next generation service platforms. In *Proceedings of the 3rd Workshop on Next Generation Service Platforms for Future Mobile Systems (SPMS 2007) at the 2007 International Symposium on Applications and the Internet*, IEEE Computer Society Press.

Weiser, M. (1991). The Computer for the Twenty-First Century . *Scientific American*, 94–10.

Zhao, Y. X. (2004). Combining RDF and OWL with SOAP for Semantic Web Services. In *Proceedings of the 3rd annual Nordic Conference on Web Services (NCWS'04)*.

Zhdanova, A. V. (2006 January). *An Approach to Ontology Construction and its Application to Community Portals*. PhD thesis, University of Innsbruck, Austria.

Zhdanova, A.V., Boussard, M., Cesar, P., Clavier, E., Gessler, S., Hesselman, C., Kernchen, R., Le Berre, O., Melpignano, D., Nani, R., Patrini, L., Räck, C., Strohbach, M., Sutterer, M., van Kranenburg, H., Villalonga, C., Vitale, A. (2006). *Ontology Definition for the DCS and DCS Resource Description, User Rules*, EU IST SPICE IP deliverable (D3.1).

Zhdanova, A. V., & Shvaiko, P. (2006). Community-driven ontology matching. In *Proceedings of the Third European Semantic Web Conference*, 11-14 June 2006, Budva, Montenegro, Springer-Verlag, LNCS 4011, pp. 34-49.

ADDITIONAL READING

Cho, M. J., Kim, H., Kim, S. H., Lee, H. J., Hong, S. C., & Jung, M. J. (2007). *Context knowledge modelling method for sharing and reusing context knowledge in context-aware system*, vol. US2007038438, pp. 10.

De, S., & Moessner, K. (2007). Context gathering in ubiquitous environments: enhanced service discovery. In *Proceedings of the 3rd Workshop on Context Awareness for Proactive Systems (CAPS'2007),* Guildford, UK.

Franz, B., Diego, C., Deborah, L. M., Daniele, N., & Peter, F. P.-S. (Eds.). (2003). *The description logic handbook: theory, implementation, and applications*, Cambridge University Press.

Kobeissy, N., Genet, M. G., & Zeghlache, D. (2007). Mapping XML to OWL for seamless information retrieval in context-aware environments. *Second IEEE International Workshop on Services Integration in Pervasive Environments (SEPS'07) at IEEE International Conference on Pervasive Services (ICPS'2007),* pp. 349-354.

Li, N., & Moessner, K. (2007). The MVCE knowledge-based content and service adaptation management framework. In *Workshop on Applications and Services in Wireless Networks,* Santander, Spain.

Li, N., & Moessner, K. (2007). *Design of content/service adaptation framework*. Mobile VCE project Deliverable (D-U3.3).

Mukherjee, D., Delfosse, E., Jae-Gon, K., & Yong, W. (2005). Optimal adaptation decision-taking for terminal and network quality-of-service. *Multimedia . IEEE Transactions on*, *7*, 454–462.

Noy, N. F., & Musen, M. A. (1999). An algorithm for merging and aligning ontologies: automation and tool support. *In Proceedings of the Workshop on Ontology Management at the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*.

Reif, G., Gall, H., & Jazayeri, M. (2005). WEESA - Web engineering for Semantic Web applications. *In Proceedings of 14th International World Wide Web Conference*, Chiba, Japan.

Smith, J. R., & Schirling, P. (2006). Metadata standards roundup. *IEEE MultiMedia*, *13*, 84–88. doi:10.1109/MMUL.2006.34

Stamou, G., van Ossenbruggen, J., Pan, J. Z., Schreiber, G., & Smith, J. R. (2006). Multimedia annotations on the semantic Web. *IEEE MultiMedia*, *13*, 86–90. doi:10.1109/MMUL.2006.15

Tsinaraki, C. (2007). Interoperability support between MPEG-7/21 and OWL in DS-MIRF. *Knowledge and Data Engineering . IEEE Transactions on*, *19*, 219–232.

Villalonga, C., Strohbach, M., Snoeck, N., Sutterer, M., Belaunde, M., Kovacs, E., et al. (2007). Mobile Ontology: Towards a standardized semantic model for the mobile domain. *In Proceedings of the 1st International Workshop on Telecom Service Oriented Architectures (TSOA 2007) at the 5th International Conference on Service-Oriented Computing*, Vienna, Austria.

Zhdanova, A. V., Du, Y., & Moessner, K. (2007). Mobile experience enhancement by ontology-enabled interoperation in a service platform. *In Proceedings of the 3rd Workshop on Context Awareness for Proactive Systems (CAPS'2007)*, Guildford, UK, ISBN: 978-0-9556240-0-1.

ENDNOTES

[1] vCard: http://www.w3.org/TR/vcard-rdf

[2] CC/PP: http://www.w3.org/Mobile/CCPP/

[3] UAProf: http://www.openmobilealliance.org/release_program/uap_v2_0.html

[4] http://www.ruleml.org
[5] http://www.w3.org/Submission/WRL/
[6] http://www.w3.org/Submission/SWRL/
[7] http://www.w3.org/2005/rules/wg
[8] http://www.w3.org/Submission/WSML/
[9] http://www.w3.org/Submission/SWSF-SWSL/
[10] http://www.w3.org/Submission/WSMO/
[11] http://www.w3.org/Submission/SWSF-SWSO/
[12] http://www.w3.org/2002/ws/sawsdl/
[13] http://www.w3.org/Submission/WSDL-S/
[14] http://www.deri.org
[15] http://www.w3.org/Submission/WSMX/
[16] http://www.oasis-open.org/committees/wss/
[17] The GLOBUS Alliance: http://www.globus.org
[18] SPICE: http://www.ist-spice.org
[19] MobileVCE: http://www.mobilevce.com
[20] Mobilife: http://www.ist-mobilife.org
[21] OPUCE: http://www.opuce.tid.es
[22] Ambient Networks: http://www.ambient-networks.org
[23] TripCom: http://www.tripcom.org
[24] SWING: http://www.swing-project.org
[25] ASG: http://asg-platform.org
[26] SmartWeb: http://www.smartweb-project.de
[27] Mobile Ontology website: http://ontology.ist-spice.org
[28] OntoSelect: http://olp.dfki.de/ontoselect
[29] Swoogle: http://swoogle.umbc.edu

This work was previously published in The Semantic Web for Knowledge and Data Management: Technologies and Practices, edited by Z. Ma; H. Wang, pp. 266-287, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 1.5
A Review of Fuzzy Models for the Semantic Web

Hailong Wang
Northeastern University, China

Zongmin Ma
Northeastern University, China

Li Yan
Northeastern University, China

Jingwei Cheng
Northeastern University, China

ABSTRACT

In the Semantic Web context, information would be retrieved, processed, shared, reused and aligned in the maximum automatic way possible. Our experience with such applications in the Semantic Web has shown that these are rarely a matter of true or false but rather procedures that require degrees of relatedness, similarity, or ranking. Apart from the wealth of applications that are inherently imprecise, information itself is many times imprecise or vague. In order to be able to represent and reason with such type of information in the Semantic Web, different general approaches for extending semantic web languages with the ability to represent imprecision and uncertainty has been explored. In this chapter, we focus our attention on fuzzy extension approaches which are based on fuzzy set theory. We review the existing proposals for extending the theoretical counterpart of the semantic web languages, description logics (DLs), and the languages themselves. The following statements will include the expressive power of the fuzzy DLs formalism and its syntax and semantic, knowledge base, the decidability of the tableaux algorithm and its computational complexity etc. Also the fuzzy extension to OWL is discussed in this chapter.

DOI: 10.4018/978-1-60566-028-8.ch002

INTRODUCTION

The Semantic Web is an extension of the current web in which the web information can be given well-defined semantic meaning, and thus enabling better cooperation between computers and people. From this point of view, we should find some methods which can describe the semantic meaning of the web. Fortunately, "ontology" can do this. The core of the Semantic Web is "ontology" which refers to a set of vocabulary to describe the conceptualization of a particular domain. Over the past few years, several ontology definition languages for the Semantic Web have emerged, including RDF(S), OIL, DAML, DAML+OIL, and OWL. Among them, OWL is the newly released standard recommended by W3C. As the Semantic Web expects, OWL has the reasoning nature because description logics (*DL*s) (Baader, 2003) are essentially the theoretical counterpart of OWL and play a crucial role in this context. *DL*s provide a logical reconstruction of object-centric and frame-based knowledge representation languages. It is a subset of first-order logic that provides sound and decidable reasoning support (Baader, 2003).

It is clear that *DL*s play a key role in the Semantic Web. As with traditional crisp logic, any sentence in OWL, being asserted facts, domain knowledge, or reasoning results, must be either true or false and nothing in between. However, most real world domains contain uncertainty knowledge and incomplete or imprecise information that is true only to a certain degree. Ontologies defined by these languages thus cannot quantify the degree of the overlap or inclusion between two concepts, and cannot support reasoning in which only partial information about a concept or individual in the domain can be obtained. Uncertainty becomes more prevalent when more than on ontologies are involved where it is often the case that a concept defined in on ontology can only find partial matches to one or more concepts in another ontology. To overcome the difficulty arising from the crisp logics, existing ontology languages need to be extended to be able to capture uncertainty knowledge about the concepts, properties and instances in the domain and to support reasoning with partial, imprecise information. Along this direction, researchers in the past have attempted to apply different formalisms such as Fuzzy logic (Zadeh, 1965), Rough set theory and Bayesian probability as well as ad hoc heuristics into ontology definition and reasoning.

In this paper, we review existing proposals to extend semantic web languages with the capability to handle uncertain information to better deal with the situations mentioned above. There are many ways of representing and dealing with uncertainty. In this paper, we restrict our attention to approaches that use fuzzy methods for representing uncertain information. In particular, we will not cover recent proposals for probabilistic extensions of semantic web languages. We will also not discuss non-monotonic and non-standard logics for representing uncertainty. As described above, existing Semantic Web languages are mainly based on logic and do not support representing imprecise and uncertain information. In this paper, we therefore review a number of proposals for extending logical languages with fuzzy extensions in more details. We focused on:

1. Approaches that extend description logics which play as the theoretical counterpart of the semantic web languages.
2. Approaches that directly extend semantic web languages, in particular OWL.

In the first category, we cover fuzzy extensions of description logics which are commonly accepted as being the formal basis of OWL. Even though most approaches only cover logics that are much weaker than OWL, the methods proposed can directly be applied to the corresponding subset of OWL without changes because the description logics play as the theory counterpart of the OWL. When talking about the different approaches, we will survey them according to the expressive power from weaker to stronger. And in the following survey, we should discuss the following issues of the different approaches:

- Expressiveness of the logical language
- The syntax and semantic of the fuzzy extension to description logics

- The components of the knowledge base
- Tableaux algorithm for the description logics
- The decidability and complexity of the tableaux algorithm

Indeed, the balance of expressive power and the computability of the fuzzy extension of description logics is a hot topic of the research. Generally speaking, the more expressive of the description logic, the higher computational complexity of it, so we should consider the balance of the two factors in a real application. At last, in the latter category, we also review a number of proposals for extending the ontology description language OWL.

The paper is structured as follows. We first present some background information on semantic web languages and related formalisms that are the basis for the logical languages used in the different approaches discussed later in the paper. We also provide a brief introduction to fuzzy set theory which the fuzzy description logics are based on. In the mainly part of this chapter, we survey the different approaches to extend the description logics to represent the imprecise and uncertainty information according their expressive power. We also discuss proposals for fuzzy languages OWL for the semantic web. Finally, we conclude with a critical review of the state of the art and an analysis of directions for future research.

PRELIMINARIES AND BACKGROUND

Description Logics

In the last decade a substantial amount of work has been carried out in the context of Description Logics. *DL*s are a logical reconstruction of the so-called frame-based knowledge representation languages, with the aim of providing a simple well-established Tarski-style declarative semantics to capture the meaning of the most popular features of structured representation of knowledge. Nowadays, *DL*s have gained even more popularity due to their application in the context of the Semantic Web (Berners-Lee, 2001).The recent research about description logics can be divided into three categories:

- Introducing the theoretical foundations of description logics, addressing some of the most recent developments in theoretical research in the area;
- Focusing on the implementation of knowledge representation systems based on Descriptions Logics, describing the basic functionality of a *DL* system, surveying the most influential knowledge representation systems based on descriptions, and addressing specialized implementation techniques;
- Addressing the use of description logics and of *DL*-based systems in the design of several applications of practical interest.

In the following statements, we mainly focus on the first category, especially the theoretical formalism of the description logics, with respect to the balance between its expressive power and its computational complexity. Indeed, subsequent results on the tradeoff between the expressiveness of a *DL* language and the complexity of reasoning with it, and more generally, the identification of the sources of complexity in *DL* systems, showed that a careful selection of language constructs was needed and that the reasoning services provided by the system are deeply influenced by the set of constructs provided to the user. We can thus characterize three different approaches to the implementation of reasoning services. The first can be referred to as *limited + complete*, and includes systems that are designed by restricting the set of constructs in such a way that subsumption would be computed efficiently, possibly in polynomial time. The second approach can be denoted as *expressive*

+ *incomplete*, since the idea is to provide both an expressive language and efficient reasoning. The drawback is, however, that reasoning algorithms turn out to be incomplete in these systems. After some of the sources of incompleteness were discovered, often by identifying the constructs—or, more precisely, combinations of constructs—that would require an exponential algorithm to preserve the completeness of reasoning, systems with complete reasoning algorithms were designed. Systems of this sort are therefore characterized as *expressive + complete*; they were not as efficient as those following the other approaches, but they provided a test bed for the implementation of reasoning techniques developed in the theoretical investigations, and they played an important role in stimulating comparison and benchmarking with other systems.

Now, we survey the languages of the description logics according to their expressive power with the beginning of *AL*. Elementary descriptions are atomic concepts and atomic roles (also called concept names and role names). Complex descriptions can be built from them inductively with concept constructors and role constructors. In abstract notation, we use the letters *A* and *B* for atomic concepts, the letter *R* for atomic roles, and the letters *C* and *D* for concept descriptions. Description languages are distinguished by the constructors they provide. In the sequel we shall discuss various languages from the family of *AL*-languages. The language *AL* (= attributive language) has been introduced as a minimal language that is of practical interest. The other languages of this family are extensions of *AL*.

The Basic Description Logic *AL*

Concept descriptions in *AL* are formed according to the following syntax rule:

$C, D ::= \top$ | (universal concept)

$\bot$ | (bottom concept)

A | (atomic concept)

$C \sqcap D$ | (intersection)

$\forall R . C$ (value restriction)

$\exists R . \bot$ | (limited existential quantification)

In *AL*, negation can only be applied to atomic concepts, and only the top concept is allowed in the scope of an existential quantification over a role. In order to define a formal semantics of *AL*-concepts, we consider interpretations I that consist of a non-empty set Δ^I (the domain of the interpretation) and an interpretation function $\bullet^I$, which assigns to every atomic concept A a set $A^I \subseteq \Delta^I$ and to every atomic role R a binary relation $R^I \subseteq \Delta^I \times \Delta^I$. The interpretation function is extended to concept descriptions by the following inductive definitions:

$\top^I = \Delta^I$

$\bot^I = \Phi$

$(\neg A)^I = \Delta^I \setminus A^I$

$(C \sqcap D)^I = C^I \cap D^I$

$(\forall R . C)^I = \{a \in \Delta^I \mid \forall b. (a, b) \in R^I \rightarrow b \in C^I\}$

$(\exists R . \bot)^I = \{a \in \Delta^I \mid \exists b. (a, b) \in R^I\}$

We say that two concepts C, D are equivalent, and write $C \equiv D$, if $C^I = D^I$ for all interpretations I.

The Family of *AL*-languages

We obtain more expressive languages if we add further constructors to *AL*. The *union* of concepts (indicated by the letter U) is written as $C \sqcup D$, and interpreted as

$(C \sqcup D)^I = C^I \cup D^I$:

Full existential quantification (indicated by the letter E) is written as $\exists R.C$, and interpreted as

$$(\exists R.C)^I = \{a \in \Delta^I \mid \exists b.\ (a, b) \in R^I \wedge b \in C^I\}$$

Note that $\exists R.C$ differs from $\exists R.$ in that arbitrary concepts are allowed to occur in the scope of the existential quantifier.

Number restrictions (indicated by the letter N) are written as $\geq nR$ (at-least restriction) and as $\leq nR$ (at-most restriction), where n ranges over the nonnegative integers. They are interpreted as

$$(\geq nR)^I = \{a \in \Delta^I \mid |\{b \mid (a, b) \in R^I\}| \geq n\}$$

and

$$(\leq nR)^I = \{a \in \Delta^I \mid |\{b \mid (a, b) \in R^I\}| \leq n\}$$

respectively, where "|•|" denotes the cardinality of a set. From a semantic view point, the coding of numbers in number restrictions is immaterial. However, for the complexity analysis of inferences it can matter whether a number n is represented in binary (or decimal) notation or by a string of length n, since binary (decimal) notation allows for a more compact representation.

The *negation* of arbitrary concepts (indicated by the letter *C*, for "complement") is written as $\neg C$, and interpreted as

$$(\neg C)^I = \Delta^I \setminus C^I$$

Extending *AL* by any subset of the above constructors yields a particular *AL*-language. We name each *AL*-language by a string of the form

AL[U][E][N][C];

where a letter in the name stands for the presence of the corresponding constructor. For instance, *ALEN* is the extension of *AL* by full existential quantification and number restrictions.

The More Expressive Description Logics

There are several possibilities for extending *AL* in order to obtain a more expressive *DL*. The three most prominent are adding additional concept constructors, adding role constructors, and formulating restrictions on role interpretations. Below, we start with the third possibility, since we need to refer to restrictions on roles when defining certain concept constructors. For these extensions, we also introduce a naming scheme. Basically, each extension is assigned a letter or symbol. For concept constructors, the letters/symbols are written after the starting *AL*, for role constructors, we write the letters/symbols as superscripts, and for restrictions on the interpretation of roles as subscripts. As an example, the *DL* $ALCQ^{-1}_{R+}$, extends *AL* with the concept constructors negation (*C*) and qualified number restrictions (*Q*), the role constructor inverse ($^{-1}$), and the restriction that some roles are transitive($_{R+}$).

Restrictions on role interpretations These restrictions enforce the interpretations of roles to satisfy certain properties, such as functionality and transitivity. We consider these two prominent examples in more detail. Others would be symmetry or connections between different roles.

(i) *Functional roles*. Here one considers a subset N_F of the set of role names N_R, whose elements are called *features*. An interpretation must map features *f* to functional binary relations $f^I \subseteq \Delta^I \times \Delta^I$. *AL* extended with features is denoted by AL_f.

(ii) *Transitive roles*. Here one considers a subset N_{R+} of N_R. Role names $R \in N_{R+}$ are called *transitive roles*. An interpretation must map transitive roles $R \in N_{R+}$ to transitive binary relations $R^{I} \subseteq \Delta^{I} \times \Delta^{I}$. *AL* extended with transitive roles is denoted by AL_{R+}.

All the *DL*s mentioned until now contain the concept constructors intersection and value restriction as a common core. *DL*s that allow for intersection of concepts and existential quantification (but not value restriction) are collected in the *EL*-family. The only constructors available in *EL* are intersection of concepts and existential quantification. Extensions of *EL* are again obtained by adding appropriate letters/symbols. In order to avoid very long names for expressive *DL*s, the abbreviation *S* was introduced for ALC_{R+}, i.e., the *DL* that extends *ALC* by transitive roles. Prominent members of the *S*-family are *SIN* (which extends ALC_{R+} with number restrictions and inverse roles), *SHIF* (which extends ALC_{R+} with role hierarchies, inverse roles, and number restrictions of the form $\leq 1R$), and *SHIQ* (which extends ALC_{R+} with role hierarchies, inverse roles, and qualified number restrictions). Actually, the *DL*s *SIN*, *SHIF*, and *SHIQ* are somewhat less expressive than indicated by their name since the use of roles in number restrictions is restricted: roles that have a transitive sub-role must not occur in number restrictions.

Description Logics with Data Type Representation

A drawback that all *DL*s introduced until now share is that all the knowledge must be represented on the abstract logical level. In many applications, one would like to be able to refer to concrete domains and predefined predicates on these domains when defining concepts. To solve the problem, Baader and Hanschke prompt two extensions (Hanschke, 1992; Haarslev et al., 1999). In the two papers, the definition of *concrete domain* is given and a tableau-based algorithm for deciding consistency of *ALC*(*D*)-ABoxes for admissible *D* was introduced in (Baader & Hanschke, 1991). The algorithm has an additional rule that treats existential predicate restrictions according to their semantics. The main new feature is that, in addition to the usual "abstract" clashes, there may be concrete ones, i.e., one must test whether the given combination of concrete predicate assertions is non-contradictory. This is the reason why we must require that the satisfiability problem for *D* is decidable. As described in (Baader and Hanschke, 1991), the algorithm is not in PSpace. Using techniques similar to the ones employed for *ALC* it can be shown, however, that the algorithm can be modified such that it needs only polynomial space (Lutz, 1999), provided that the satisfiability procedure for *D* is in PSpace. In the presence of acyclic TBoxes, reasoning in *ALC*(*D*) may become NExpTime-hard even for rather simple concrete domains with a polynomial satisfiability problem (Lutz, 2001).

The more expressive description logics *SHOQ*(*D*) which can represent data information is proposed in (Horrocks, 2001). Although *SHOQ*(*D*) is rather expressive, it has a very serious limitation on data types; i.e., it does not support customised data types. It has been pointed out that many potential users will not adopt it unless this limitation is overcome. Pan and Horrocks release a series of papers about data types to solve the problem, in (Pan & Horrocks, 2006; Pan, 2007). In the two papers, they summarize the limitations of OWL datatyping and propose the data type approach. For example, the *SHIQ*(*G*) and *SHOQ*(*G*) *DL*s presented in (Pan & Horrocks, 2006; Pan, 2007) can support user-defined data type and user-defined data type predicates.

Fuzzy Set Theory

Fuzzy data is originally described as fuzzy set (Zadeh, 1965). Let *U* be a universe of discourse, then a fuzzy value on *U* is characterized by a fuzzy set *F* in *U*. A membership function $\mu_F: U \rightarrow [0, 1]$ is defined for the fuzzy set *F*, where $\mu_F(u)$, for each $u \in U$, denotes the degree of membership of *u* in the fuzzy set *F*. Thus the fuzzy set *F* is described as follows.

$F = \{\mu_F(u_1)/u_1, \mu_F(u_2)/u_2, ..., \mu_F(u_n)/u_n\}$

When the $\mu_F(u)$ above is explained to be a measure of the possibility that a variable *X* has the value *u* in this approach, where *X* takes values in *U*, a fuzzy value is described by a possibility distribution π_X. Let π_X and *F* be the possibility distribution representation and the fuzzy set representation for a fuzzy value, respectively. In the fuzzy set theory, each object $u_i \in U$ is assigned a single value between 0 and 1, called the degree of membership, where *U* is a universe of discourse.

Fuzzy set theory is the theory basis for fuzzy extensions to description logics to represent imprecise and uncertain information.

FUZZY EXTENSIONS OF SEMANTIC WEB LANGUAGES

Extensions of Description Logics

Much work has been carried out towards combining fuzzy logic and description logics during the last decade. The initial idea was presented by Yen in (Yen, 1991), where a structural subsumption algorithm was provided in order to perform reasoning. The following statements will illustrate all the fuzzy extensions to *DL*s from weaker to stronger in expressive power.

The Family of *FALC* Languages

ALC is the basic format of the description logics. Reasoning in fuzzy *ALC* was latter presented in (Straccia, 2001), as well as in other approaches (Straccia, 1998), where an additional concept constructor, called membership manipulator was included in the extended language. In all these approaches tableaux decision procedures were presented for performing reasoning services. The operations used to interpret the concept constructors in all these approaches were the same ones as in our context. (Tresp & Molitor, 1998) contains complete algorithms for solving these inference problems in the respective fuzzy extension of *ALC*. Although both algorithms are extensions of the usual tableau-based algorithm for ALC, they differ considerably. For example, the algorithm in (Tresp & Molitor, 1998) introduces numerical variables for the degrees, and produces a linear optimization problem, which must be solved in place of the usual clash test. In contrast, (Straccia, 2001) deals with the membership degrees within his tableau-based algorithm.

The fuzzy description logic *FALC* *Definition.* Let N_I, N_C and N_R be three disjoint sets: N_I is a set of individual names, N_C is a set of fuzzy concept names and N_R is a set of fuzzy role names. Fuzzy *ALC*-concepts are defined as

$C, D ::= \bot | \mathrm{T} | A | \neg C | C \cup D | C \cap D | \exists R.C | \forall R.C$;

Here $A \in N_C$, $R \in N_R$. Fuzzy *ALC* semantics is defined by a fuzzy interpretation $I = \langle \Delta^I, \cdot^I \rangle$, Here Δ^I is a nonempty set and $\cdot^I$ is an function which maps every $a \in N_I$ to an element $a^I \in \Delta^I$, maps every $A \in N_C$ into a function $A^I: \Delta^I \rightarrow [0, 1]$, and maps every $R \in N_R$ into a function $R^I: \Delta^I \times \Delta^I \rightarrow [0, 1]$.Furthermore, for any fuzzy *ALC*-concepts *C* and *D*, $R \in N_R$ and $x \in \Delta^I$, we have:

$\mathrm{T}^I(x) = 1;$

$\bot^I(x) = 0;$

$(\neg C)^I(x) = 1 - C^I(x);$

$(C \cap D)^I(x) = C^I(x) \wedge D^I(x);$

$(C \cup D)^I(x) = C^I(x) \vee D^I(x));$

$(\exists R.C)^I(x) = sup_{\,y \in \Delta I} \{min\,(R^I(x, y), C^I(y))\};$

$(\forall R.C)^I(x) = inf_{\,y \in \Delta I} \{max\,(1 - R^I(x, y), C^I(y))\};$

With the introduction of the fuzzy sets into the classical *ALC*, the form of the knowledge base is changed accordingly:*Definition.* A fuzzy *ALC* knowledge base is composed of a TBox and an ABox:

- A TBox is a finite set of terminology axioms of the form $C \subseteq D$. Any interpretation I satisfies $C \subseteq D$ iff for any $x \in \Delta^I$, $C^I(x) \leq D^I(x)$. I is a model of TBox T iff I satisfies all axioms in T.
- An ABox is a finite set of assertions of the form $< \alpha \bowtie n >$, Here $\bowtie \in \{>, \geq, <, \leq\}$, $n \in [0, 1]$, α (called a fuzzy assertion) is either of the form a: C or (a, b): $R(a,b \in N_I)$. Especially, in order to giving a uniform format of the ABox, we define: when n =1, the form $<\alpha \geq 1>$ is equivalent to $<\alpha = 1>$. Concretely speaking, $<a: C \geq 1>$means that a is determinately an individual of C; $<(a, b)$: $R \geq 1>$ means that (a, b) determinately has the relationship R. Any interpretation I satisfies $<a{:}C \bowtie n>$ iff $C^I(a^I) \bowtie n$ and satisfies $<(a, b)$: $R \bowtie n >$ iff $R^I(a^I, b^I) \bowtie n$. Then I is a model of ABox A iff I satisfies all assertions in A.

Reasoning algorithms for *FALC* and their proofs can be found in (Straccia, 2001). It can be seen from syntax and semantics presented above that the entailment and subsumption relationships may hold to some degree in the interval [0, 1]. Complete algorithms for reasoning in *FALC* have been presented, that is, we have devised algorithms for solving the entailment problem, the subsumption problem as well as the best truth-value bound problem. The complexity result shows that the additional expressive power has no impact from a computational complexity point of view.

The fuzzy description logic ALC_{FM} The *DL* language used was a sub-language of the basic *DL ALC*. The main idea underlying the fuzzy extensions of description logics proposed in (Tresp & Molitor, 1998) is to leave the syntax as it is, but to use fuzzy logic for defining the semantics. Thus, an interpretation now assigns fuzzy sets to concepts and roles, i.e., concept names A are interpreted by membership degree functions of the form A^I: $\Delta^I \rightarrow [0, 1]$, and role names R by membership degree functions of the form R^I: $\Delta^I \times \Delta^I \rightarrow [0, 1]$. The interpretation of the Boolean operators and the quantifiers must then be extended from {0, 1} to the interval [0, 1].

Tresp & Molitor (1998) also propose an extension of the syntax by so-called manipulators, which are unary operators that can be applied to concepts. Examples of manipulators could be "*mostly*", "*more or less*", or "*very*". For example, if *Tall* is a concept (standing for the fuzzy set of all tall persons), then *VeryTall*, which is obtained by applying the manipulator *Very* to the concept *Tall*, is a new concept (standing for the fuzzy set of all *very tall* persons). Intuitively, the manipulators modify the membership degree functions of the concepts they are applied to appropriately. In our example, the membership function for *VeryTall* should have its largest values at larger heights than the membership function for *Tall*. Formally, the semantics of manipulators is defined by a function that maps membership degree functions to membership degree functions. The manipulators considered in (Tresp & Molitor, 1998) are, however, of a very restricted form. Lets us now consider what kind of inference problems are of interest in this context. (Yen, 1991) considers crisp subsumption of fuzzy concepts, i.e., given two concepts C, D defined in the fuzzy *DL*, he is interested in the question whether $C^I(d) \leq D^I(d)$ for all fuzzy interpretations I and $d \in \Delta^I$. Thus, the subsumption relationship itself is not fuzzified. He describes a structural subsumption algorithm for a rather small fuzzy *DL*, which is almost identical to the subsumption algorithm for the corresponding classical *DL*. In contrast, (Tresp & Molitor, 1998) are interested in determining fuzzy subsumption between fuzzy concepts, i.e., given concepts C, D, they want to know to which degree C is a subset of D. In (Tresp & Molitor, 1998), also ABoxes are considered, where the ABox assertions are equipped with a degree.

Fuzzy description logics with hedges and modifiers Reasoning in fuzzy *ALC* was also presented in other approaches (Hölldobler, 2002; 2003; 2004; 2005; 2006), where an additional concept constructor, called membership manipulator was included in the extended language. In all these approaches tableaux decision procedures were presented for performing reasoning services. (Hölldobler, 2002) presents a fuzzy description logic ALC_{FH}, where primitive concepts are modified by means of hedges. ALC_{FH} is strictly more expressive than Fuzzy *ALC* defined in (Straccia, 2001). The paper shows that given a linearly ordered set of hedges primitive concepts can be modified to any desired degree by prefixing them with appropriate chains of hedges. Furthermore, it defines a decision procedure for the unsatisfiability problem in ALC_{FH}, and discusses truth bounds, expressivity as well as complexity issues.

Strictly speaking, the language defined by (Tresp & Molitor, 1998) is more expressive, as we do not consider concept modifiers. From a semantics point of view, the extension to Tresp and Molitor's language is quite straightforward. But, the cost that we have to pay for this increasing expressive power is that, from a computational complexity and algorithms point of view, things changes radically. The fuzzy extension to *ALC* can be used as a basis both for extending existing *DL* based systems and for further research. In this latter case, there are several open points. For instance, it is not clear yet how to reason both in case of fuzzy specialisation of the general form $C \sqsubseteq D$ and in the case cycles are allowed in a fuzzy KB. Another interesting topic for further research concerns the semantics of fuzzy connectives. While for a huge number of proposals given in the literature their impact from a semantics point of view is well understood, the question how they impact from a computational complexity and algorithms point of view remains still open.

According to the above statements, the family of fuzzy *ALC* description logics can be summaried in Table 1.

The More Expressive Fuzzy Description Logics

However, *FALC* offers limited expressive power of complex fuzzy information. Some discussions about reducing *FALC* into classical *ALC* and providing a tableau for *FALC* with General Concept Inclusions (GCIs) were given in (Straccia, 2004a) and (Stoilos, 2006c), respectively. In (Stoilos, 2006c), fuzzy description logics have been proposed as a language to describe structured knowledge with vague concepts. A major theoretical and computational limitation so far is the inability to deal with General Concept Inclusions (GCIs), which is an important feature of classical *DL*s. It addresses this issue and develops a calculus for fuzzy *DL*s with GCIs. (Meghini *et al*. 1998) proposed a preliminary fuzzy *DL*, which lacks reasoning algorithm, as a modeling tool for multimedia document retrieval.

In the following statements, we survey the dif-

Table 1. The family of FALC languages

Fuzzy description logics	Corresponding references	Mainly discussed issues
FALC	(Straccia, 1998; 2001)	Syntax,semantics, properties and reasoning services
ALC_{FM}	(Tresp and Molitor, 1998)	Syntax,semantics and a mothod for computing the degree of subsumption
ALC_{FH}	(Hölldobler, 2002; 2003; 2004; 2005; 2006)	Syntax,semantics and reasoning services
ALC_{FL}	(Dinh-Khac, 2006)	Syntax,semantics and reasoning services

ferent formalisms of the description logics according to the concept constructors or role restrictions that have been constraint on them.

Fuzzy description logics with number restrictions Approaches towards more expressive *DL*s, are presented in (Sánchez, D, 2004), where the *DL* is $ALCQ_F^+$. It includes fuzzy quantifiers, which is a new novel idea for fuzzy *DL*s. Unfortunately, in the approach only the semantics of the extended languages are provided and no reasoning algorithms. But in (Sánchez, D, 2006), it introduces reasoning procedures for $ALCQ_F^+$, the fuzzy description logic with extended qualified quantification. The language allows for the definition of fuzzy quantifiers of the absolute and relative kind by means of piecewise linear functions on N and Q∩[0, 1] respectively. In order to reason about instances, the semantics of quantified expressions is defined based on recently developed measures of the cardinality of fuzzy sets. A procedure is described to calculate the fuzzy satisfiability of a fuzzy assertion, which is a very important reasoning task. The procedure considers several different cases and provides direct solutions for the most frequent types of fuzzy assertions. In addition, (Sánchez, D, 2006) defined *independence* of fuzzy assertions and obtained some results that speed up the calculation of fuzzy satisfiability in some (the most common) cases.

The series of *F-SHOIQ* Because today quite a lot of multimedia systems and applications use knowledge representation formalisms to encode and reason with knowledge that exists within the multimedia documents, (Stoilos, 2005a) presents a more expressive fuzzy *DL* f_{KD}-*SI*. The goal of this direction is to narrow the semantic gab between the content of a multimedia object, as perceived by a human being, and as "viewed" by an information system. (Stoilos, 2005a) has extended the *DL* language *SI* with fuzzy set theory. The combination of transitive and inverse roles can capture knowledge about part-whole relationships and aggregated objects. Furthermore, the incorporation of fuzziness allows the users to encode and reason with vague and imprecise knowledge. Both these properties fit well into the framework of knowledge based multimedia processing where both part-whole relationships, as well as, imprecise and vague knowledge appear in applications like multimedia information retrieval and processing. In the latter of the paper, a tableau algorithm for checking the consistency of ABox is given and has been proved that it will terminates. (Stoilos, 2005b) provides an extension of the above f_{KD}-*SI* to an even more expressive *DL*, namely f_{KD}-*SHIN*. The paper has presented an extension of the very expressive description logic *SHIN* with fuzzy set theory. It shows the semantics as well as detailed reasoning algorithms for the extend languages. A fuzzy tableau for f_{KD}-*SHIN* ABoxes is shown and it proves the flowing lemma.**Lemma:***Let A be an* f_{KD}*-SHIN ABox and R a fuzzy RBox. Then*

i. when started for A and R the tableaux algorithm terminates
ii. A has a fuzzy tableau w.r.t. R if and only if the expansion rules can be applied to A and R such that they yield a complete and clash-free completion forest.

In (Stoilos, 2006b), the syntax and semantics of fuzzy *SHOIQ* were presented and the properties of the semantics of transitivity, qualified cardinality restrictions and reasoning capabilities were investigated. (Stoilos, 2006b) extends the current state-of-the-art on fuzzy extensions to Semantic Web languages by presenting the syntax and semantics of the fuzzy-*SROIQ DL* as well as the abstract, XML syntax and semantics of a fuzzy extension to OWL 1.1. Moreover, it provides reasoning support for a fuzzy version of fuzzy-*SROIQ* by extending well-known reduction techniques of fuzzy *DL*s to classical *DL*s for the additional axioms and constructors of fuzzy-*SROIQ*.

Fuzzy description logics with concrete domain The Semantic Web is expected to process knowledge information and data information in an intelligent and automatic way. But recent research has shown that the OWL DL ontology language is very limited in representing data information. Furthermore, the OWL DL can't process imprecision and uncertainty which widely exists in human knowledge and natural language. In (Straccia, 2004b), a fuzzy extension of *ALC*(*D*) (the *ALC* extended with concrete domains) was presented. The paper presents a fuzzy description logic where the representation of concept membership functions and fuzzy modifiers is allowed, together with a inference procedure based on a mixture of a tableaux and bounded mixed integer programming.

The more expressive *DL*s which can support fuzzy concrete domains is shown in (Straccia, 2005), where the language is *SHOIN*(*D*+). It is the corresponding Description Logic of the ontology description language OWL DL. It shows that the representation and reasoning capabilities of fuzzy *SHOIN*(*D*) go clearly beyond classical *SHOIN*(*D*). Interesting features are: (i) concept constructors are based on t-norm, t-conorm, negation and implication; (ii) concrete domains are fuzzy sets; (iii) fuzzy modifiers are allowed; and (iv) entailment and subsumption relationships may hold to some degree in the unit interval [0, 1]. The fuzzy concrete domain is defined as following: $(\Delta_D, \bullet^D)$ is an interpretation. $\bullet^D$ is an interpretation which assigns each concrete individual to an element in Δ_D; assigns each simple data type role $T \in \mathbf{R}_D$ to a function $T^I : \Delta^I \times \Delta_D \rightarrow [0, 1]$; assigns each n-ary predicate p to the fuzzy relation $p^D: \Delta_D^{\,n} \rightarrow [0,1]$ which means the relationship of data types $v_1,\ldots,v_n$ satisfies predicate p in a degree in [0,1]. Based on the definition of the fuzzy concrete domain, the fuzzy data information in the semantic web can be represented and reasoning.

According to the above statements, some more expressive fuzzy description logics can be summaried in Table 2.

Extensions of OWL

More recently, little work has been carried out towards combining fuzzy logic and the Semantic Web ontology. In (Stoilos, 2005c), the OWL web ontology language was extended with fuzzy set theory, which is called f-OWL, in order to capture, represent and reason with imprecise information. Based on a fuzzy extension to OWL called Fuzzy OWL, (Stoilos, 2006a) developed a reasoning platform, Fuzzy Reasoning Engine (FiRE), which lets Fuzzy OWL capture and reason about imprecise and uncertain knowledge. Several connections between Fuzzy Logic, the Semantic Web, and its components were presented in (Stoilos, 2005c).

For most recent research issues about fuzzy logic, and more generally soft computing, in the description logics, ontologies and the Semantic Web, ones can refer to (Ma, 2006), (Sanchez, 2006a; 2006b).

Table 2. Some important fuzzy description logics

	Representation of fuzzy terminologies and concepts					**Representation of fuzzy data information**		
	f-*TSL*	f-*ALCQ*$_F$+	f-*SI*	f-*SHIN*	f-*SHOIQ*	f-*ALC*(D)	f-*SHOIQ* (D)	f-*SHOIQ* (G)
Syntax & semantic	(Yen, 2001)	(Sánchez, 2004,2006)	(Stoilos, 2005a)	(Stoilos, 2005b)	(Stoilos, 2006b)	(Straccia, 2004b)	(Straccia, 2005)	
Tableau algorithm	(Yen, 2001)	(Sánchez, 2004,2006)	(Stoilos, 2005a)	(Stoilos, 2005b)	(Stoilos, 2006b)	(Straccia, 2004b)		
Decidability	(Yen, 2001)	(Sánchez, 2006)	(Stoilos, 2005a)	(Stoilos, 2005b)	(Stoilos, 2006b)			

DISCUSSION AND CONCLUSIONS

In this chapter, we focus our attention on the recent research achievements on fuzzy extension approaches which are based on fuzzy set theory. We survey existing proposals for extending the theoretical counterpart of the semantic web languages, description logics (*DL*s), and the languages themselves. The above statements include the expressive power of the fuzzy *DL*s formalism and its syntax and semantic, knowledge base, the decidability of the tableaux algorithm and its computational complexity etc. Also the fuzzy extension to OWL is discussed in this chapter. After reviewing all the proposals of fuzzy extensions, we find that it is a paradox in the balance between the expressive power of the *DL* formalism and its computational complexity. It is sure that the computational complexity will become more complex with the more expressive power of the *DL* formalism.

FUTURE RESEARCH DIRECTIONS

As overall conclusion, we can summarize that until recently, research has not paid much attention to fuzzy extensions to *DL*s in the area of the Semantic Web. However, it gains more and more interest and new approaches considering imprecise and uncertainty tend to emerge. Still, many of these approaches are rather half-baked and a lot of things are missing:

Fuzzy extensions to concrete domain to support the fuzzy user-defined data type and fuzzy user-defined data type predicate. It should be pointed out that, however, the *SHIQ*(*G*) and *SHOQ*(*G*) *DL*s presented in (Pan & Horrocks, 2006; Pan, 2007) which can support the user-defined data type and user-defined data type predicate can only deal with crisp knowledge. In the real world, human knowledge and natural language have a big deal of imprecision and uncertainty, as a result, a fuzzy extension version of *DL* language which can process fuzzy data information should be developed. Especially, we should pay more attention on the fuzzy extension of user-defined data type and user-defined data type predicates.

The computational complexity of the fuzzy tableau algorithm and the optimal technologies to the fuzzy tableau algorithm. After prompting a fuzzy tableau algorithm to an expressive *DL* language, although the decidability of the algorithm is investigated, the computational complexity is still an open problem. Furthermore, the optimal technologies can be investigated which can reduce the computational complexity of the algorithm.

Fuzzy extensions to description logics based on vague sets. A single membership degree in the fuzzy sets is inaccurate to represent the imprecision in the membership degrees. As a result, based on vague sets (Gau, 1993), a fuzzy extension of description logic should be presented. Instead of a crisp degree of membership, two degrees of membership (lower and upper degrees of membership) are used in the newly proposed fuzzy description logic version. Its syntax, semantics and inference problems and tableaux also should be investigated.

REFERENCES

Baader, F., Calvanese, D., & McGuinness, D. (2003). *The Description Logic Handbook: Theory, Implementation and Applications.* Cambridge University Press.

Baader, F., & Hanschke P. (1991). A scheme for integrating concrete domains into concept languages. In *Proc. of the 12th Int. Joint Conf. On Artificial Intelligence (IJCAI'91)*, pp. 452–457.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *The Scientific American*, 284 (5), 34- 43.

Dinh-Khac, D., Hölldobler, S. & Tran, D. K. (2006). The fuzzy linguistic description logic

ALC_{FL}. *Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2096-2103.

Gau, W. L. & Buehrer, D. J., (1993) Vague sets. *IEEE Transactions on Systems, Man, and Cybernetics*, 23 (2), 610-614.

Haarslev V., Lutz C., & Möller R. (1999). A description logicwith concrete domains and role-forming predicates. *J. of Logic and Computation*, 9(3), 351–384.

Hanschke P. (1992). Specifying role interaction in concept languages. In *Proc. of the 3rd Int. Conf. on the Principles of Knowledge Representation and Reasoning (KR'92)*, Morgan Kaufmann, Los Altos, pp. 318–329.

Horrocks I., & Sattler, U. (2001). Ontology reasoning in the SHOQ(D) description logic. *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*, pp. 199–204.

Hölldobler, S., Khang, T.D., & Störr, H.P. (2002) A fuzzy description logic with hedges as concept modifiers. In *Proceedings In Tech/VJFuzzy'2002*. pp. 25–34.

Hölldobler, S, Störr, H. P. & Tran, D. K. (2003). The fuzzy description logic ALC_{FH} with hedge algebras as concept modifiers. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 7 (3), 294-305.

Hölldobler, S., Nga, N. H. & Khang, T. D. (2005). The fuzzy description logic ALC_{FLH}. *Proceedings of the 9th IASTED International Conference on Artificial Intelligence and Soft Computing*, pp. 99-104.

Hölldobler, S., Störr, H. P., & Khang, T. D. (2004). The subsumption problem in the fuzzy description logic ALC_{FH}. *Proceedings of the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 243-250.

Pan., J. Z. (2007). A Flexible Ontology Reasoning Architecture for the Semantic Web. *IEEE Transaction on Knowledge and Data Engineering.* 19(2), 246 - 260.

Pan, J. Z. & Horrocks, I. (2006). OWL-Eu: Adding Customised Data types into OWL. *Journal of Web Semantics*, 4(1), 29-49.

Lutz, C. (1999). *Reasoning with concrete domains.* In Proc. of the 16th Int. Joint Conf. on Artificial Intelligence (IJCAI'99), Stockholm, Sweden, pp. 90–95.

Lutz, C. (2001). NEXPTIME-complete description logics with concrete domains. In *Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2001)*, volume 2083 of Lecture Notes in Artificial Intelligence, Springer, pp. 45–60.

Meghini, C., Sebastiani F. & Straccia, U. (1997). Reasoning about the form and content for multimedia objects. *Proceedings of AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video and Audio*, pp. 89-94.

Ma. Z. M. (2006) *Soft Computing in Ontologies and Semantic Web.* Springer-Verlag.

Sanchez, E., (2006a). *Fuzzy Logic and the Semantic Web.* Elsevier.

Sanchez, E. & Yamanoi, T. (2006b). Fuzzy ontologies for the Semantic Web. *Lecture Notes in Artificial Intelligence,* 4027: pp. 691-699.

Sánchez, D., & Tettamanzi, G. (2004). Generalizing quantification in fuzzy description logic. In *Proceedings 8th Fuzzy Days in Dortmund*.

Sánchez, D. & Tettamanzi, G. (2006). Reasoning and quantification in fuzzy description logics. *Lecture Notes in Artificial Intelligence*, 3846, pp. 81-88.

Shadbolt, N., Hall, W. & Berners-Lee, T. (2006).

The Semantic Web revisited. *IEEE Intelligent Systems*, 21 (3), 96-101.

Stoilos, G., Stamou, G., & Tzouvaras, V. (2005a) A fuzzy description logic for multimedia knowledge representation. *Proceedings of the 2005 International Workshop on Multimedia and the Semantic Web.*

Stoilos, G., Stamou, G., & Tzouvaras V. (2005b) The fuzzy description logic f-SHIN. *Proceedings of the International Workshop on Uncertainty Reasoning for the Semantic Web*, pp. 67-76.

Stoilos, G., Stamou, G., & Tzouvaras, V. (2005c) Fuzzy OWL: Uncertainty and the Semantic Web. *Proceedings of the 2005 International Workshop on OWL: Experience and Directions.*

Stoilos, G., Simous, N., & Stamou, G. (2006a), Uncertainty and the Semantic Web. *IEEE Intelligent Systems,* 21 (5), 84-87.

Stoilos, G., Stamou, G. & Pan, J. Z., (2006b) Handling imprecise knowledge with fuzzy description logic. *Proceedings of the 2006 International Workshop on Description Logics.*

Stoilos, G., Straccia, U., & Stamou, G. B. (2006c). General concept inclusions in fuzzy description logics. *Proceedings of the 17th European Conference on Artificial Intelligence*, pp. 457-461.

Straccia, U. (1998). *A* fuzzy description logic. In *Proc. of the 15th Nat. Conf. on Artificial Intelligence (AAAI-98),* pp. 594–599 Madison, USA.

Straccia, U. (2001) Reasoning within fuzzy description logics. *Journal of Artificial Intelligence and Research,* 14(1), 137–166

Straccia, U. (2004a). Transforming fuzzy description logics into classical description logics. *Proceedings of the 9th European Conference on Logics in Artificial Intelligence*, pp. 385-399.

Straccia, U. (2004b). Fuzzy ALC with fuzzy concrete domains. *Proceedings of the 9th European Conference on Logics in Artificial Intelligence*, pp. 385-399.

Straccia, U. (2005). Towards a fuzzy description logic for the semantic web. In *Proceedings of the 2nd European Semantic Web Conference*.

Tresp, C., & Molitor, R. (1998). A description logic for vague knowledge. In *Proc of the 13th European Conf. on Artificial Intelligence (ECAI-98).* Brighton, England.

Yen, J. (1991). Generalising term subsumption languages to fuzzy logic. In *Proc of the 12th Int. Joint Conf on Artificial Intelligence (IJCAI-91)*, pp: 472-477, Sydney, Australia

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3), 338-353.

ADDITIONAL READING

General Logic

Ben-Ari, M. (2001). *Mathematical Logic for Computer Scientists.* Springer-Verlag.

Ebbinghaus, H.-D., Flum, J. & Thomas, W. (2007) *Mathematical Logic.* Springer-Verlag.

Enderton, H. B. (2002). *A mathematical Introduction to Logic*, 2nd edition. Academic Press.

Schoening, U. (1994). *Logic for Computer Scientists.* Birkhaeuser Verlag

Fuzzy Set Theory

Zadeh, L.A. (1978) Fuzzy sets as a basis for a theory of possibility, *Fuzzy Sets Systems*, 1(1), pp. 3-28.

Semantic Web

Antoniou, G. and Harmelen, F. V. (2003). *A Semantic Web Primer*, The MIT Press.

Description Logic

Franconi, E., *Description Logics Course Information*. http://www.cs.man.ac.uk/~franconi/dl/course/

OWL

Horrocks, I. & Patel-Schneider, P. F., van Harmelen, F. (2003). From SHIQ and RDF to OWL: The making of a web ontology language. *Journal of Web Semantics*, 1(1):7-26.

Smith, M. K., Welty, C. & McGuiness, D. L. (2004). *OWL Web Ontology Language Guide*. W3C Recommendation, from http://www.w3.org/TR/2004/REC-owl-ref-20040210/

RDF and RDF Schema

A selection of documents on RDF and RDF Schema (Specification, Use Cases, Recommended Readings, Tools, Related Technologies, etc.), from http://www.w3.org/RDF/

de Bruijn & J., Heymans, S. (2007). Logical Foundations of (e)RDF(S): Complexity and Reasoning. In *Proceedings of the International Semantic Web Conference (ISWC).*

This work was previously published in The Semantic Web for Knowledge and Data Management: Technologies and Practices, edited by Z. Ma; H. Wang, pp. 23-37, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 1.6
Semantic Web Adaptation

Alexander Mikroyannidis
University of Manchester, UK

Babis Theodoulidis
University of Manchester, UK

INTRODUCTION

The rate of growth in the amount of information available in the World Wide Web has not been followed by similar advances in the way this information is organized and exploited. Web adaptation seeks to address this issue by transforming the topology of a Web site to help users in their browsing tasks. In this sense, Web usage mining techniques have been employed for years to study how the Web is used in order to make Web sites more user-friendly.

The Semantic Web is an ambitious initiative aiming to transform the Web to a well-organized source of information. In particular, apart from the unstructured information of today's Web, the Semantic Web will contain machine-processable metadata organized in ontologies. This will enhance the way we search the Web and can even allow for automatic reasoning on Web data with the use of software agents. Semantic Web adaptation brings traditional Web adaptation techniques into the new era of the Semantic Web. The idea is to enable the Semantic Web to be constantly aligned to the users' preferences. In order to achieve this, Web usage mining and text mining methodologies are employed for the semi-automatic construction and evolution of Web ontologies. This usage-driven evolution of Web ontologies, in parallel with Web topologies evolution, can bring the Semantic Web closer to the users' expectations.

DOI: 10.4018/978-1-59904-845-1.ch093

BACKGROUND

Web Usage Mining

Web usage mining has a wide variety of applications. User profiles can be created for use in Web personalization. Information can also be extracted that details how a Web site can be reorganized to better facilitate users' navigation through it. In e-commerce Web sites, the results of Web usage mining can be used to improve sales. Analyzing user access patterns can also help when targeting advertisements to specific groups of users.

Srivastava, Cooley, Deshpande, and Tan (2000) divide Web usage mining into three stages:

i. Preprocessing
ii. Pattern discovery
iii. Pattern analysis

Preprocessing consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery. Usage preprocessing involves the identification of users and their visiting sessions. In order for this to be accomplished, several difficulties need to be overcome. For example, proxy servers hide the actual IP addresses of the machines that are using them, thus making user identification problematic. A user that uses more than one browser, even on the same machine, will appear as multiple users. Tracking repeat visitors can also be complex if a user uses different machines.

Content preprocessing consists of converting the text, image, scripts, or multimedia files into forms that are useful for the Web usage mining process. This often involves the application of content mining techniques, such as classification or clustering. For instance, a classification algorithm could be used to limit the discovered patterns to those that contain page views about a certain subject. Similar to the preprocessing of the site's content, structure preprocessing regards the extraction of the site's structure for use in the mining procedure. The hyperlinks of each Web page build the structure of the Web site. Most Web sites have nowadays an utterly dynamic topology, thus presenting a different structure to different users. This characteristic should be considered during the structure preprocessing phase.

The preprocessing stage is followed by the discovery of traversal patterns from the user access data. Traversal patterns reveal the way a user navigates through the site during each session. Clusters of users can be discovered through clustering of similar traversal patterns. Moreover, association rules can be applied to the pages accessed during a session, independent of their ordering. Examples of association rules that were extracted from an IBM analysis of the Web logs of the Official 1996 Olympics Web site (Elo-Dean & Viveros, 1997) are:

- 45% of the visitors who accessed a page about Indoor Volleyball also accessed a page on Handball.
- 59.7% of the visitors who accessed pages about Badminton and Diving also accessed a page about Table Tennis.

The percentages mentioned in both association rules are called confidence. Confidence can be defined as the number of transactions containing all of the items in a rule, divided by the number of transactions containing the rule antecedents (Cooley, Mobasher, & Srivastava, 1999). Additionally, temporal relationships among data items can be discovered, such as the following (Cooley, Mobasher, & Srivastava, 1997):

- 30% of clients who visited the '/company/products' page had done a search in Yahoo within the past week on keyword w.
- 60% of clients who placed an online order in the '/company/product1' page also placed an online order in the '/company/product4' page within 15 days.

Depending on the purpose of the mining, a traversal pattern may contain backward traversals. Backward traversals consist of references of pages earlier visited. Studying backward traversals can help discover missing hyperlinks, which if added will reduce these traversals, thus making navigation paths shorter and more convenient for the users. A pattern's accesses can also be restricted only to contiguous ones, which can be used for prefetching and caching purposes. A frequent pattern is maximal if does not contain any frequent subpatterns. This can reduce significantly the number of meaningful discovered patterns.

Pattern analysis is the last stage of the Web usage mining process. The patterns that have been produced are reviewed, and useful information is extracted from them. Knowledge query mechanisms, similar to structured query language (SQL), can be used to filter out the patterns. Another approach involves the use of data cubes and OLAP operations. Visualization techniques, such as graphing patterns or assigning colors to different values, can also be utilized to highlight interesting trends in the data. Last but not least, content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

Numerous approaches to Web usage mining have been followed, targeted to a wide range of applications. Chen, Park, and Yu (1998) and Nanopoulos and Manolopoulos (2000) have introduced the concept of using the maximal forward references to break down user sessions into transactions for mining access patterns. Yang, Pan, and Chung (2001) have proposed an efficient hash-based method, HMFS, for discovering the maximal frequent itemsets. Spiliopoulou (1999) has presented an algorithm for building aggregating trees from Web logs, then mining the Web access patterns by MINT mining language. (Cooley et al. (1999) have provided a query language on top of external mining software for association rules and for sequential patterns. Another query language for extracting navigation patterns, called MiDAS, has been proposed by Buchner, Baumgarten, Anand, Mulvenna, and Hughes (1999). Xiao and Dunham (2001) have investigated techniques to discover frequently used contiguous sequences of page references, which they call maximal frequent sequences (MFS). They have also developed an algorithm called online adaptive traversal (OAT) pattern mining, to mine MFS. Xing and Shen (2004) have proposed two algorithms, user access matrix (UAM) and preferred navigation tree (PNT), for mining user preferred navigation patterns.

Web Adaptation

Several Web adaptation systems have been developed over the years, mainly based on Web usage mining techniques. The WebWatcher system (Joachims, Freitag, & Mitchell, 1997) suggests links that may interest a user, based on other users' online behaviour. The system is implemented in the form of a proxy server. Each user is asked, upon entering the site, what kind of information he is seeking. Before he departs, the user is asked whether he has found what he was looking for. His navigation paths are used to deduce suggestions for future visitors that seek the same content. These suggestions are visualized by highlighting existing hyperlinks.

The Avanti project (Fink, Kobsa, & Nill, 1996) tries to predict the visitor's final objective as well as her next step. A model for the visitor is built, based partly on the personal information of the visitor and partly on information extracted from her navigation paths. Visitors are provided with direct links to pages that are probably the ones they are looking for. In addition, hyperlinks that lead to pages of potential interest to each visitor are highlighted. The suggestions are extracted with the use of association rules that are applied to the user's model.

A drawback of both the WebWatcher and the Avanti approaches is that they require the active participation of the users in the adaptation process, by asking them to provide information about themselves. On the other hand, Footprints (Wexelblat & Maes, 1999) relies entirely on the navigation paths of the users. The navigation paths of all visitors are recorded and the most frequent ones are presented to the visitor. Additionally, next to each link, the percentage of people who have followed it is displayed.

Perkowitz and Etzioni (2000) have proposed a conceptual framework for adaptive Web sites, focusing on the semi-automatic creation of index pages from clusters of pages. They have developed two cluster mining algorithms, PageGather and IndexFinder. The first one relies on a statistical approach to discover candidate link sets, while the second finds link sets that are conceptually coherent.

The Semantic Web

Since January 2005, the Web has grown by more than 17 million sites, according to monitoring firm Netcraft. This figure exceeds the growth of 16 million sites seen in 2000 when net fever reached its most intense pitch.[1] Significant progress has been made in technologies for publication and distribution of knowledge and information on the Web. People add private, educational, or corporate content. Growth also comes from the rise in blogging, in which users write regularly updated Web journals on various topics. However, much of the published information is not organized, and it is hard to find answers to questions that require more than a keyword search.

The Semantic Web (Berners-Lee, Fischetti, & Dertouzos, 1999) means to address these problems by expressing Web data in forms that are machine-processable, in order to be more efficiently maintained by software agents, thus enhancing precision of search, as well as logical reasoning. The vision behind this concept can be summarized as "giving information a well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee, Hendler, & Lassila, 2001).

Ontologies are a key enabling technology for the Semantic Web, since they offer a way to give information a common representation and semantics. Daconta, Obrst, and Smith (2003) distinguish three levels of ontologies: top, middle, and lower domain levels. At the top level, the ontological information represented concerns primary semantic distinctions that apply to every ontology. The middle level represents knowledge that spans domains and may not be as general as the knowledge of the upper level. Finally, the lower level represents ontologies at the domain or subdomain level. This is typically knowledge about domain-specific subject areas. While an ontologist can address the upper and to a certain extent the middle level, the domain expert is absolutely required for the construction and maintenance of the lower level.

Mikroyannidis and Theodoulidis (2005) distinguish between the domain ontology and the ontology of a Web site dedicated to a certain domain. A Web site ontology is strongly related to the topology of the site and is comprised of the thematic categories covered by the site's pages. These categories are the concepts of the ontology. Each Web page, depending on its content, is an instance of one or more concepts of the ontology. The concepts can be related to each other through a number of relationship types, representing the associations the concepts have according to the Webmaster's perception. Figure 1 shows the Web site ontology belonging to the University of Manchester School of Informatics (www.informatics.manchester.ac.uk). The ontology has been built considering the organization of the thematic categories as this is defined in the current topology of the site. The hierarchy's top level contains seven classes: School, Undergraduate Programmes, Postgraduate Taught Programmes, Postgraduate Research, Research, News and Intranet. These are the main thematic categories of the site. These categories are then

Figure 1. The School of Informatics Web site ontology

expanded to more specific concepts, which are represented by subclasses.

It should be pointed out that the Web site ontology is quite different from the domain ontology. The latter describes relationships between the concepts of a domain, whereas the first is based on the organization of the information found in a Web site. The ontology of a domain is usually more complex than the ontology of a Web site related to the same domain. However, the maintenance of a Web site ontology requires considerable effort and has to be performed on a regular basis, since the content of a Web site is constantly updated.

SEMANTIC WEB ADAPTATION

The Semantic Web is undoubtedly a remarkable advance in the area of information management. The magnitude of Web data necessitates the use of machine-processable metadata. Nevertheless, the Web users' needs and requirements should not be neglected in the process of building and maintaining the Semantic Web. Semantic Web adaptation utilizes traditional Web usage mining methodologies and extends them in order to address the ontological perspective of the Semantic Web. Apart from the topology of the Web, which is targeted in a conventional Web adaptation system, semantic Web adaptation also aims to the evolution of Semantic Web ontologies.

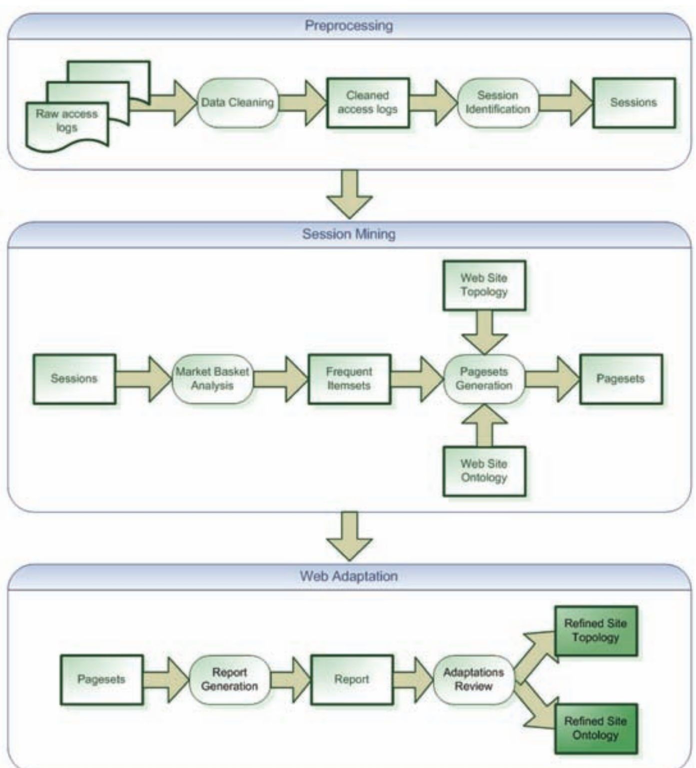
The Heraclitus framework (Mikroyannidis, 2004; Mikroyannidis & Theodoulidis, 2004, 2005) proposes the transformation of the Semantic Web based on Web usage data. Web usage mining and text mining are used for the extraction of knowledge from navigation paths and the adaptation of the physical and semantic structure of the Web. An implementation of the Heraclitus framework has been released as a suite of open source tools (http://heraclitus.sourceforge.net).

Architecture

Figure 2 presents the architecture of the Heraclitus framework. As it can be seen, the inputs of the adaptation process consist of raw access logs, the Web site topology and ontology. The whole procedure aims at the evolution of the topology and ontology of the Web site.

The adaptation starts with a preprocessing stage, during which the data stored in the raw access logs are cleaned and visiting sessions are identified. The sessions are then mined with the use of market basket analysis (Bodon, 2003) to retrieve frequent itemsets. These are then classified based on the Web site topology and ontology in order to produce page sets, that is, sets of pages that are frequently accessed together throughout the same session. During classification, each itemset is assigned its position in the topology and the ontology of the Web site. For the latter, automatic categorization is performed with the use of the support vector machines (SVM) algorithm (Cortes & Vapnik, 1995). SVM has been trained on the thematic categories that are defined in the concepts of the Webmaster's ontology. Each page

Figure 2. Semantic Web adaptation architecture



of the page sets is then assigned by SVM to one or more concepts of the ontology.

Based on the topological and ontological features of the page sets, a report containing proposals for the improvement of the Web site is generated. This report contains proposals for the insertion of shortcut links from source pages to target pages that are frequently accessed together but are currently not linked. It also contains proposals for the change of the appearance of popular hyperlinks. In addition, the report contains proposals for the evolution of the Web site ontology. After the proposed modifications have been revised by the Webmaster, they can be applied to the Web site. The site topology is then refined through the insertion of new shortcut links, as well as changes in the appearance of the existing ones. The ontology is also refined in a number of ways.

Case Study

The Web site of the School of Informatics at the University of Manchester (www.informatics.manchester.ac.uk) was used as a case study for the Heraclitus framework. The topology of the Web site was refined through the insertion of new shortcut links between pages that were not previously linked together, as well as through the highlighting of popular existing links. In addition, the Web site ontology was modified in several ways, based on the outcomes retrieved from the classified page sets.

More specifically, Heraclitus produced two sets of reports: shortcut links reports and highlighted links reports. Page sets of unlinked pages suggest the insertion of shortcut links between these pages, in order to achieve shorter navigation paths. From the page sets of linked pages, changes in the appearance of existing links can be extracted. For example, if an index page and some of its links comprise one or more page sets, then highlighting these links in the index page will provide valuable help to visitors.

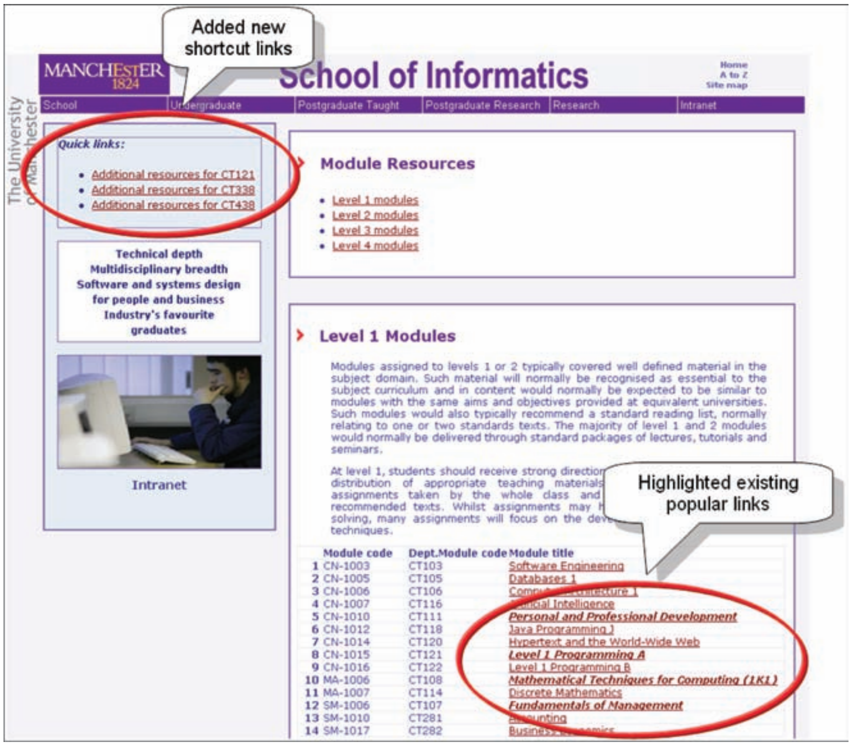
Figure 3 shows an example of a modified Web page, according to the proposed Heraclitus adaptations. The page has been modified to facilitate the navigation of the users during the first semester. Shortcut links to popular courses of the first semester have been inserted in the left side of the page, under the title "Quick links." Moreover, popular links that already existed, such as the hyperlink leading to the page of the "Personal and Professional Development" course, have been highlighted.

The Web site ontology was modified in several ways, based on the outcomes retrieved from the classified page sets. The resulting ontology, after the application of Heraclitus adaptations, is shown in Figure 4. Based on these adaptations, the content organization of the Web site was altered to better satisfy the needs of its visitors. First of all, new associations were discovered between concepts. These associations reflect the interests of the users, as documents belonging to these concepts are frequently accessed together. In particular, new associations were inserted between the following concepts:

- "Research" and "Students"
- "Research" and "Staff"
- "School" and "Students"
- "School" and "Staff"
- "Students" and "Staff"

Reorganization of the concepts' hierarchy was also performed. Further improvements included the creation of new categories, the removal of existing categories, as well as changes to the levels of hierarchy that the concepts belong to. For instance, the "Staff" concept was previously a subconcept of the "School" concept, which resided in the highest level of the ontology. It should be noted that the "Staff" concept has as instances all the Web pages that carry information about the staff members of the school. However, the high frequency with which this concept appeared in

Figure 3. Proposed Heraclitus adaptations in a sample Web page

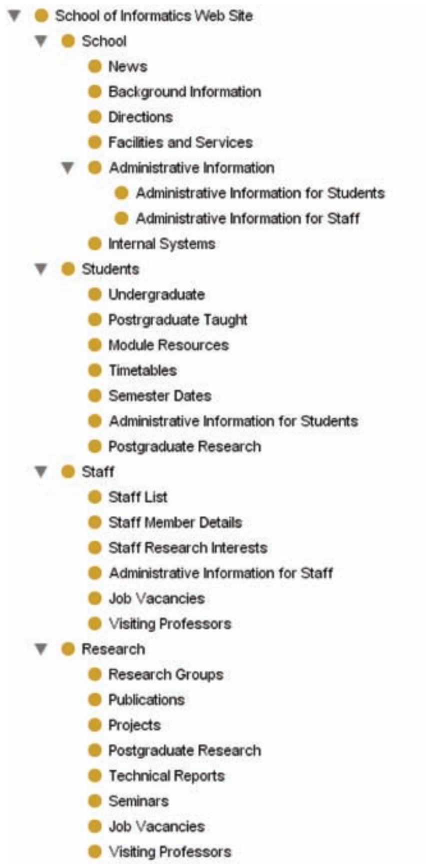


the page sets implies the significance that it has in the interests of the users. It would be thus appropriate to transfer this concept to the top level of the ontology, as shown in Figure 4. Based on the performed classification, the undergraduate and postgraduate programmes were grouped under the more general concept "Students." The "School" concept was also extended to include more subconcepts.

The ontology of the Web site was extended to include multiple instances of concepts or multiple subconcepts. The categorization of the Web pages that was carried out suggested that several pages belong to more than one concept. Moreover, in some cases, Web pages and the corresponding concepts were categorized under different concepts than they previously were in the existing ontology. The Web site ontology should be therefore updated in order to reflect this fact. For example, the "Job vacancies" Web page, which corresponds to the "Job Vacancies" concept, was found to be an instance of both the "Staff" and "Research" concepts. The information contained in this page regards mainly research job posts and is also highly related to the "Staff" concept. This page was previously categorized only under the "School" concept. In the updated ontology, the "Job Vacancies" concept has been placed both under the "Staff" and "Research" concepts. The same modification has been applied to the concepts "Visiting Professors," "Administrative Information for Students," and so forth.

Finally, useful conclusions were deduced about the usage of the Web site. Particularly, the thematic category that was the first in the preferences of the users was, as expected, the "Students" concept. This concept contains all pages that support the school's modules, both undergraduate and postgraduate. This is not surprising, since most of the traffic is generated by the students. Second,

Figure 4. Refined Web site ontology for the School of Informatics



in the users' interests comes the "Staff" concept. The "Research" concept is third, followed by the "School" category. These results can be used to enhance the performance of the server, for example, by the use of additional servers that will host the popular resources, or to promote the problematic concepts by making them more easily accessible.

CONCLUSION

The Semantic Web is undoubtedly a remarkable advance in the area of information management. The magnitude of Web data necessitates the use of machine-processable metadata. Nevertheless, the needs and requirements of the Web users should not be neglected in the process of building and maintaining the Semantic Web.

The Heraclitus framework approaches Semantic Web adaptation from a user-oriented perspective. Transformation of the Web topology and ontology is carried out having as basis the information retrieval tasks of the visitors. Web usage mining and text mining are used for the extraction of knowledge from navigation paths of the visitors. Practicing semantic Web adaptation on a real Web site has provided an insight in navigation difficulties of the users as well as in ways to overcome them.

REFERENCES

Berners-Lee, T., Fischetti, M., & Dertouzos, M. (1999). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. San Francisco: Harper.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *279*(5), 34–43.

Bodon, F. (2003). *A fast APRIORI implementation.* Paper presented at the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03), Melbourne, FL.

Buchner, A. G., Baumgarten, M., Anand, S. S., Mulvenna, M. D., & Hughes, J. G. (1999). *Navigation pattern discovery from Internet data.* Paper presented at the WEBKDD'99, San Diego, CA.

Chen, M. S., Park, J. S., & Yu, P. S. (1998). Efficient data mining for path traversal patterns. *IEEE Transactions on Knowledge and Data Engineering*, *10*(2), 209–221. doi:10.1109/69.683753

Cooley, R., Mobasher, B., & Srivastava, J. (1997). *Web mining: Information and pattern discovery on the World Wide Web.* Paper presented at the IEEE International Conference on Tools with Artificial Intelligence.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining World Wide Web browsing patterns. *Knowledge and Information Systems*, *1*(1), 5–32.

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning*, *20*(3), 273–297.

Daconta, M., Obrst, L., & Smith, K. (2003). *The Semantic Web: A guide to the future of XML, Web services, and knowledge management*. John Wiley and Sons Ltd.

Elo-Dean, S., & Viveros, M. (1997). *Data mining the IBM official 1996 Olympics Web site* (Technical report). IBM T.J. Watson Research Center.

Fink, J., Kobsa, A., & Nill, A. (1996). *User-oriented adaptivity and adaptability in the AVANTI project.* Paper presented at the Designing for the Web: Empirical Conference.

Joachims, T., Freitag, D., & Mitchell, T. (1997). *WebWatcher: A tour guide for the World Wide Web.* Paper presented at the International Joint Conference on Artificial Intelligence, Nagoya, Japan.

Mikroyannidis, A. (2004). *Development of a framework for self adaptive Web sites.* Unpublished MPhil, University of Manchester, Manchester, UK.

Mikroyannidis, A., & Theodoulidis, B. (2004, September 20-24). *A theoretical framework and an implementation architecture for self adaptive Web sites.* Paper presented at the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), Beijing, China.

Mikroyannidis, A., & Theodoulidis, B. (2005, May 29-June 1). *Web usage driven adaptation of the Semantic Web.* Paper presented at the End User Aspects of the Semantic Web Workshop, 2nd European Semantic Web Conference (ESWC 2005), Heraklion, Greece.

Nanopoulos, A., & Manolopoulos, Y. (2000). *Finding generalized path patterns for Web log data mining.* Paper presented at the East-European Conference on Advances in Databases and Information Systems (ADBIS'00), Prague, Czech Republic.

Perkowitz, M., & Etzioni, O. (2000). Adaptive Web sites. *Communications of the ACM*, *43*(8), 152–158. doi:10.1145/345124.345171

Spiliopoulou, M. (1999). The laborious way from data mining to Web mining. *International Journal of Computer Systems, Science and Engineering . Special Issue on Semantics of the Web*, *14*, 113–126.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web usage mining: Discovery and applications of usage patterns from Web data. *SIGKDD Explorations*, 12-23.

Xiao, Y., & Dunham, M. (2001). Efficient mining of traversal patterns. *Data & Knowledge Engineering archive, 39*(2), 191-214.

Xing, D., & Shen, J. (2004). Efficient data mining for Web navigation patterns. *Information and Software Technology*, *46*(1), 55–63. doi:10.1016/S0950-5849(03)00109-5

Yang, D., Pan, C., & Chung, Y. (2001). *An efficient hash-based method for discovering the maximal frequent set.* Paper presented at the 25th Annual International Computer Software and Applications Conference (COMPSAC'01), Chicago, Illinois.

KEY TERMS

Ontology: A representation of a certain domain, through the definition of concepts, relationships between concepts, and instances of concepts.

Semantic Web Adaptation: The process of transforming the topology and ontology of the Web in order to improve its usability.

Web Access Log: A listing of page reference data. Web access logs are created by Web servers in order to keep track of the requests that occur on Web sites by Web users.

Web Adaptation: The process of transforming the topology of the Web in order to align it with the preferences of the users, thus facilitating their browsing.

Web Mining: Mining data related to the Word Wide Web, such as the content of Web pages, intrapage structure, which includes the HTML or XML code of a page, interpage structure that is the linkage structure between Web pages, usage data that describe how Web pages are accessed, and user profiles, including demographic, registration information, or information found in cookies.

Web Site Ontology: An ontology whose concepts are the thematic categories covered by the pages of a Web site. Each Web page, depending on its content, is an instance of one or more concepts of the ontology. The concepts are related to each other through a number of relationship types, representing the associations the concepts have according to the Webmaster's perception.

Web Usage Mining: An application of data mining methodologies to Web access logs in order to discover trends and regularities in navigation patterns of Web users.

ENDNOTE

1 "Web enjoys year of biggest growth" (http://news.bbc.co.uk/2/hi/technology/4325918.stm), BBC News, 10 October 2005.

This work was previously published in Encyclopedia of Information Communication Technology, edited by A. Cartelli; M. Palma, pp. 704-712, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 1.7
Tips for Tracking Web Information Seeking Behavior

Brian Detlor
McMaster University, Canada

Maureen Hupfer
McMaster University, Canada

Umar Ruhi
University of Ottawa, Canada

ABSTRACT

This chapter provides various tips for practitioners and researchers who wish to track end-user Web information seeking behavior. These tips are derived in large part from the authors' own experience of collecting and analyzing individual differences, task, and Web tracking data to investigate people's online information seeking behaviors at a specific municipal community portal site (myhamilton.ca). The tips discussed in this chapter include: (1) the need to account for both task and individual differences in any Web information seeking behavior analysis; (2) how to collect Web metrics through deployment of a unique ID that links individual differences, task, and Web tracking data together; (3) the types of Web log metrics to collect; (4) how to go about collecting and making sense of such metrics; and (5) the importance of addressing privacy concerns at the start of any collection of Web tracking information.

DOI: 10.4018/978-1-59904-974-8.ch013

INTRODUCTION

Upon first consideration, employing Web tracking to better understand end-user experiences with the Web seems to be a simple process of installing the tracking software, collecting the data over a certain period of time, and conducting the analysis. However, our own experience in setting up, collecting, and analyzing Web tracking data has shown us that the process is surprisingly more difficult than originally expected.

To share what we have learned to help others set up and better utilize Web tracking tools, we have reflected upon what we believe are key tips concerning the use of Web tracking in any Web information seeking analysis. Thus, the overall purpose of this chapter is to discuss the practicalities and usefulness of collecting Web tracking data to help measure and assess the performance and usage of a Website or application, particularly with respect to Web information seeking.

Note that the ideas presented in this chapter are grounded in a research project conducted by the authors over the last three years that investigates people's online behaviors at a municipal community portal site called myhamilton.ca (www.myhamilton.ca). The ultimate goal of the project is to understand the relationships among individual user characteristics such as demographics and personality traits, user attitudes toward and perceptions about accomplishing certain tasks (Web services) online, and actual usage behavior. We believe that an understanding of these relationships will provide insight into how characteristics of the individual, the task, and utilization behaviors affect task performance in an online community environment. We also believe that the capture and analysis of Web tracking data is imperative to reaching such an understanding.

The difficulty in utilizing Web tracking data successfully is in knowing how to position its collection and use within the larger confines of Web information seeking analysis. Web tracking is just one tool that needs to be coordinated with other data collection methods to yield a more comprehensive understanding than Web tracking alone could ultimately provide.

The objective of this chapter is to raise awareness of this point and to suggest techniques and approaches for the collection and analysis of Web tracking information that will aid practitioners in their performance measurement initiatives and understanding of how end-users seek information on the Web. Various tips are presented:

- The need to account for both task and individual differences in any Web information seeking analysis assessment
- The benefits of using a unique ID to link individual differences, task, and Web tracking data
- The types of Web metrics to collect
- How to gather and make sense of the Web metric information that is collected in Web logfiles
- The importance of addressing privacy concerns right up-front in the collection of Web tracking information

We begin by providing background on the need to take both task and individual differences into consideration when investigating end-user Web information seeking behavior. To do this, we provide a general model that describes how task and individual differences affect information seeking behavior. Next, methods to conduct a Web information seeking analysis that allows for the collection of both task and individual differences data are presented. Importantly, these methods include the collection of Web tracking data via the use of Web logs. Using a selective subset of variables from the general model presented earlier, our own myhamilton.ca project serves as a point of illustration. We also provide details with respect to the types of Web metrics to collect and what needs to be done to make sense of these data. Finally, the importance of addressing privacy in any Web information seeking analysis is highlighted.

To help clarify things, find below the following definitions of terms:

- *Information seeking behavior* refers to how people seek information in different contexts (Fisher, Erdelez & McKechnie, 2005).
- *Web information seeking behavior* refers to information seeking behaviors that occur over the Web. Choo, Detlor & Turnbull (2000) identify four main modes

of information seeking on the Web ranging from wayward browsing to goal-directed search (undirected viewing, conditioned viewing, informal search, and formal search) where each mode is characterized by predominant information seeking moves or activities (undirected viewing: starting and chaining; conditioned viewing: browsing and differentiating; informal search: differentiating, monitoring, and extracting; and formal search: monitoring and extracting).

- *Individual differences* are the demographic and psychological characteristics of people that distinguish one person from another.
- *Task* in this chapter refers to the information seeking task an individual user experiences that instills a need for information and motivates the user to satisfy this information need through some sort of information seeking behavior. Task is the context surrounding a person's information need.
- *Web tracking* refers to the automated collection of Web information seeking behavioral data.
- *Web metrics* pertains to the measures by which to assess a person's Web information seeking behavior or to assess and monitor activity on a Website. Examples of commonly used Web metrics include page views, page transitions, and session times.

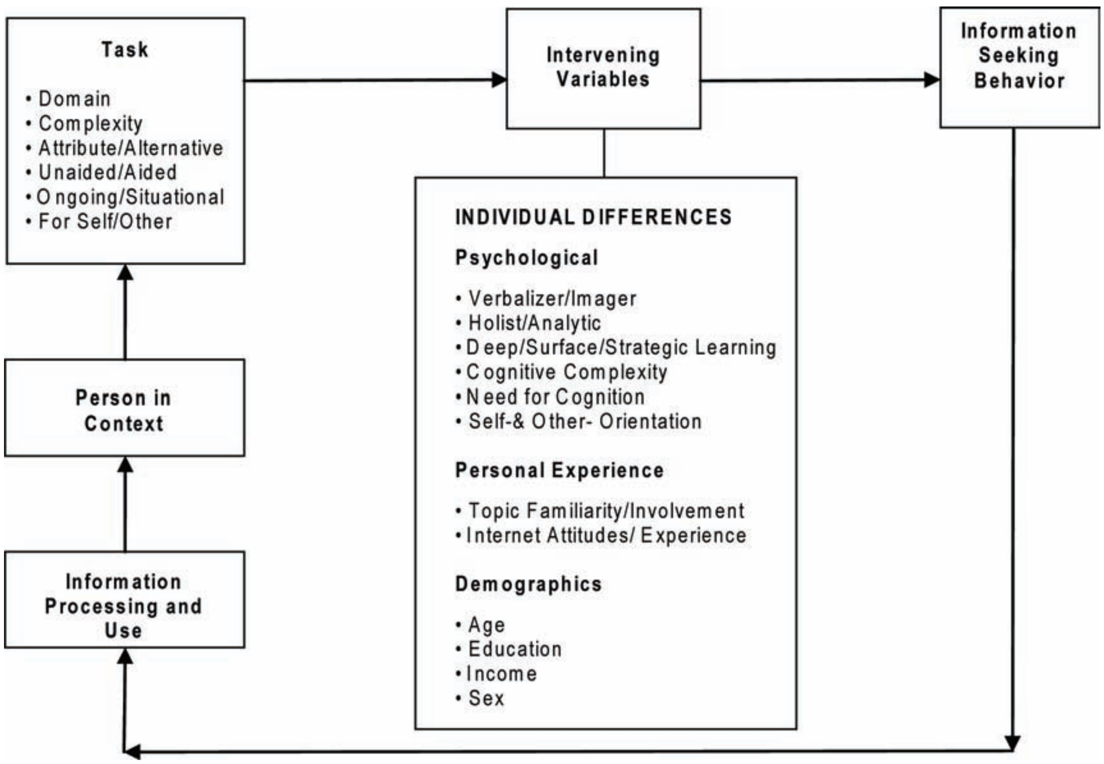
INDIVIDUAL DIFFERENCES, TASKS, AND INFORMATION SEEKING BEHAVIOR

Research concerning online information seeking in both information science and marketing has shown that information seeking strategy depends on the type of information seeking task or its context (e.g., Bhatnagar & Ghose, 2004; Moe 2003; Toms & Trifts, 2006; Wildemuth & Hughes, 2005). Scholars in both fields, as well as those in psychology, also have begun to examine the role of individual differences in online behavior (e.g., Bhatnagar & Ghose, 2004; Das, Echambadi, McCardle, & Luckett, 2003; Dillon & Watson, 1996; Ford, Miller & Moss 2001, 2005a, 2005b; Gugerty, Treadaway & Rubinstein, 2006; Heinström, 2005; Ho, 2005; Martin, Sherrard & Wentzel, 2005; Tuten & Bosnjak 2001). Unfortunately, the study of individual differences in information seeking has tended to take a haphazard approach that has failed to link findings with broader theoretical frameworks concerning information seeking behavior and has neglected to study the effects of individual differences in conjunction with specific seeking contexts (Saracevic, 1991). The work of Ford et al. (2001, 2005a, 2005b) is a notable exception both for its use of Wilson's model of information behavior (Wilson & Walsh, 1996) as a basis for investigation and for its examination of how information seeking complexity and individual differences in cognitive style interact to result in differing information seeking strategies (Ford et al., 2005b).

To situate individual differences within the Web information seeking context, we propose our own model of information seeking behavior that utilizes Wilson's (1999) model as a theoretical foundation (see Figure 1). According to this model, task (analogous to Wilson's "context of the information need" construct) leads to information seeking behavior that is mediated by individual differences variables. The purpose of the model is to illustrate how task and individual differences fit into and influence the end-user information seeking process, and to stress the importance of the need to take both task and individual differences into account when planning any type of Web information seeking analysis assessment.

As Figure 1 shows, with respect to **task**, there are a variety of characteristics about a task that can influence an end-user's information seeking behavior. For instance, prior research has found

Figure 1. How task and individual differences affect information seeking behavior



substantial differences in information seeking patterns across tasks and between product categories or information domains (e.g., Bhatnagar & Ghose, 2004; Trifts & Toms, 2006; Wildemuth & Hughes, 2005). These differences, in part, can be explained by the complexity of the information seeking task, the extent to which the task is clearly structured, whether one is seeking information for oneself or for someone else (Hupfer & Detlor, 2006), and semantic differences between search domains (Byström & Järvelin, 1995; Vakkari, 1999, pp. 825). Ford et al. (2003) suggest that complex tasks require a conceptual broadening of useable terminology to reflect broader search concepts. Conversely, simple tasks would be ones in which all essential concepts necessary to complete the information seeking task are fully specified in the task instructions. In a consumer decision making context, this conceptual broadening may be closely related to how well a consumer is able to mentally formulate the parameters of an information search. For example, consumers who are well aware of their current product needs may be easily able to articulate this need in the form of a search query (e.g., buying a particular DVD), but a decision made with less specificity (e.g., planning a vacation with no particular destination in mind) may require broadening of the search parameters to learn more about the various alternatives available.

Even within a specified information domain, task complexity is influenced by such factors as the number of alternatives available, the number of dimensions of information on which the alternatives vary, and time pressure (Payne, Bettman & Johnson, 1993). Greater complexity in an information seeking task often leads to more heuristic-based processing of information. Decision strategies that require processing information by attribute as opposed to alternative are thought to be easier to undertake.

In the context of online information seeking, tasks that require people to search by attribute as opposed to alternative may be cognitively less complex and require less time at the general search tool level. Therefore, even within a specified domain, differences in information seeking effort allocation may occur, depending upon a person's information seeking orientation (Huneke, Cole & Levin, 2004). That is, whether an individual is engaging in attribute-based or alternative-based processing of information affects the allocation of information seeking effort between general search engines versus specific Websites (Toms & Trifts 2006). Those who are engaged in alternative-based processing are more focused on finding an appropriate source of information and thus allocate a greater amount of their information seeking effort at a general search engine as opposed to in-site search.

Other characteristics of the task shown in Figure 1 that can affect information seeking behavior are whether the information seeking task is aided/unaided or ongoing/situational. The former refers to the extent to which Web information seeking is assisted by interactive decision aids. For example, in the domain of online shopping, Haubl and Trifts (2000) found that interactive tools that assisted consumers in their initial screening of alternatives substantially reduced the amount of information seeking undertaken and improved decision making. Despite the initial learning that is required, in the long run the use of interactive decision aids should reduce task complexity such that users will be able to devote less effort to obtain the required information than they would expend if unassisted. The latter refers to whether information is needed on an ongoing basis, such as when an individual has an interest in a product category or topic but does not intend to make a decision immediately, or whether information is needed for present use, such as a pre-purchase situation in which a decision is imminent.

As Figure 1 shows, **individual differences** play an important role in terms of mediating the effect of task on information seeking behavior in terms of the information seeking strategy or process chosen, as well as its effectiveness. These differences may include an individual's familiarity or level of involvement with the information topic (Moorthy, Ratchford & Talukdar,1997), experience with the Internet (Bhatnar & Ghose, 2004), perceptions of Web-based information seeking (Ford & Miller, 1996), and enduring psychological traits (e.g., Ford et al., 2001; 2005a; 2005b).

For example, Bhatnar and Ghose (2004) found that users with greater experience with the Internet and more education utilized the Internet more frequently. Other demographic characteristics, including age and sex, also have been associated with differences in Web information seeking patterns. Ford et al. (2001) established that information retrieval effectiveness was associated with males while retrieval failure was associated with females. Women felt that they were not in control of their information seeking; they were unable to avoid irrelevant material and stay on target. Men, however, were confident that they were in control and could bypass extraneous content. Educational research conducted with children also has found sex differences in information seeking such that boys searched differently from girls and were able to acquire more target-specific and target-related information. Boys filtered information at an early stage but girls were linear and more thorough navigators (Roy & Chi, 2003).

As Figure 1 illustrates, there are several psychological differences that may influence information seeking behavior in terms of an individual's propensity to engage in elaborate, effortful processing versus effort minimization and reliance on heuristics. Explained below, these include: verbalizer /imager and holistic/analytic cognitive styles; deep, surface and strategic learning approaches; cognitive complexity; Need for Cognition; and Self- and Other-Orientation.

Cognitive Style

Individuals differ in the strategies that they use to seek and process information, and they tend to favor certain strategies, or cognitive styles, on a consistent basis. Among these styles, verbalizer/imager and holist/analytic are the two dominant dimensions (Riding & Cheema, 1991). The verbalizer/imager dimension refers to a preference for and facility with tasks and information that are presented in a verbal versus visual format; verbal and spatial ability are closely related measures (Ekstrom, French & Harman, 1976). Analytic individuals perceive components of complex stimuli as discrete elements and are better able to analyze and impose structure than those who are holist, with their tendency to perceive stimuli in a holistic or global manner. Where Internet searching is concerned, Wang, Hawk and Tenopir (2000) found that holist searchers experienced more difficulty and confusion than analytic users. Ford et al. (2001) found that poor retrieval was linked to a verbalizer cognitive style, as well as perceptions that the Internet's graphic elements were of little value. Similarly, Gugerty et al. (2006) demonstrated that superior spatial, rather than verbal, ability was associated with more favorable computer and Internet attitudes, and also had an indirect effect on information seeking performance. Ford et al. (2001) found no relationships between holist or analytic cognitive styles and retrieval effectiveness, but did find relationships among holists, imagers and Boolean searching and among analytics, verbalizers and Best-match searching (2005a; 2005b). It also appears that cognitive style effects are more important for novice than for experienced Internet searchers (Palmquist & Kim, 2000). Such evidence suggests that individual differences in cognitive style affect not only the information seeking process but also its effectiveness.

Learning Style

Those with a surface approach describe learning as knowledge reproduction achieved through rote learning and memorization. They are passive uncritical learners who devote relatively little effort to information seeking (Ford, 1986; Entwistle & Tait, 1995). Deep learners, on the other hand, view learning as a process that creates knowledge through the synthesis and assimilation of new information. They seek a broad range of information sources using a variety of information seeking strategies. Strategic learners are able to choose either deep or surface learning approaches as appropriate to the task at hand. Analysis of self-reported information seeking behavior has found that a surface approach to learning was associated with a fast surfing information seeking strategy in which users experienced problems with critical analysis and had difficulty judging the relevance of retrieved documents (Heinström, 2005). They also demonstrated confirmatory bias and preferred to access information using only a few documents. In contrast, a deep diving approach characterized those with either deep or strategic learning styles; these individuals were effortful information seekers who sought high quality documents. In addition, Ford et al. (2005a) have found that individual items in the surface learning style (fear of failure and poor time management) were linked to poor retrieval. As with cognitive style, it appears that learning style affects both the information seeking process and its outcomes.

Cognitive Complexity

Those who are cognitively simple tend to see the world in binary terms such as black and white or right and wrong, while those who are cognitively complex are able to see shades of grey and recognize that the validity of a given viewpoint may vary with circumstances. Higher levels of cognitive complexity appear to be associated with poor retrieval (Ford et al., 2001).

Need for Cognition

Individuals with a high Need for Cognition (NFC) enjoy thinking and have a greater tendency to elaborate upon, structure and evaluate information (Cacioppo, Petty & Kao, 1984). They engage in more effortful decision making than those who are low NFC and arrive at better information seeking outcomes (Bailey, 1997). High NFC users also have more favorable attitudes toward Websites with complex verbal and simple visual elements (Martin et al., 2005). NFC is positively correlated with Web information usage (Tuten & Bosnjak 2001) and has a direct impact on self-reported information seeking behavior (Das et al., 2003). Finally, investigation of information seeking at online grocery stores has found that high NFC shoppers, compared with low NFC consumers, investigated more URLs and spent more time reading (Ho, 2005).

Self- and Other-Orientation

These characteristics describe differences in an individual's propensity to be concerned with oneself versus others by tapping gender-related traits that pertain to an independent (Self-Orientation) versus interdependent (Other-Orientation) self-concept orientation (Hupfer, 2001). Self- and Other-Orientation predict Internet use frequency and preferences that male-female indicators often fail to explain. The two scales interact to predict how often individuals seek information online (Hupfer & Detlor, 2006) both for themselves (self-relevant information) and for those close to them (other-relevant information). Other-Orientation also is positively related to usage rates for Internet applications with relationship implications, such as greeting cards (Hupfer & Detlor, 2007a). Furthermore, the two scales interact to predict the importance to an individual of Website characteristics that imply an information-rich environment versus navigational aids that ease processing and maximize efficiency (Hupfer & Detlor, 2007b).

METHODS FOR COLLECTING INDIVIDUAL DIFFERENCES, TASK, AND WEB TRACKING DATA

Recognizing the importance of both task and individual differences in Web information seeking behavior, attention now turns to the methods that allow for the collection of task, individual differences, and information seeking behavior data in a Web information seeking analysis. These methods invariably involve the collection of Web tracking activity via the use of Web logs, but tracking alone is insufficient for a thorough information seeking analysis. Web tracking captures information seeking behavior with Web logfiles, but other data collection instruments, such as questionnaires and interviews, are required to collect individual differences and other task-related data.

In closed environments, like laboratories, researchers can control research participants' information seeking tasks by giving them explicit descriptions or instructions for their tasks and can ask them to complete surveys to collect individual differences data. Closed environments also allow researchers to require that participants use specific software in the lab where their Web activity will be tracked. Other advantages of the closed environment include the opportunity to modify the information seeking tools that are used (e.g., browsers, interfaces), control the available functionality and even provide interactive decision aids. However, closed environments are not without their drawbacks. Processing large numbers of participants through laboratory sessions requires considerable time and resources. In addition, requiring subjects to conduct contrived searches in an artificial setting or scenario may compromise the validity and generalizability of any research results.

Richer Web information seeking analyses are more likely to be found in open environments where end-users can conduct real-life information seeking tasks that are of relevance and importance to them and that take place within natural environments and settings (e.g., the workplace, the home). In open environments, researchers are able to observe natural behavior patterns and collect data that affords greater validity and generalizability. However, conducting Web information seeking investigations in open environments presents its own challenges for the proper collection and analysis of task and individual differences data. For example, dynamic IP addressing prevents the linking of a person's Web tracking data with any individual differences data that is collected through user profiles or questionnaires. Knowing what task prompted a user to turn to the Web to seek information also is problematic in terms of understanding the type of task and its attributes.

There are various ways to go about capturing task, individual differences, and Web behavior data, but a critical component is the ability to link all three types of data together for a specific individual. It probably is easiest to create this linkage in a closed environment. However, if researchers and practitioners want to take advantage of the benefits afforded by Web information seeking analyses conducted in open environments, they must devise a means of connecting these various data sources. We did this, quite successfully, in our own research project at a municipal/community portal site called myhamilton.ca. The project involved the use of two surveys (one pertaining to task; the other pertaining to individual differences) and the collection of Web tracking data to yield a robust understanding of Web user information seeking behavior. Note that a selective subset of task and individual differences variables from the general model presented above were used in our myhamilton.ca research project. Importantly, a unique identification feature linked participants' actual portal activity to demographic, personality and attitude data. To do this, we had to work closely with the portal development group to ensure that the study's data collection instruments (i.e., Web tracking and online surveys) were incorporated directly within the portal's design.

We believe that the use of a unique ID to link data collected in the user surveys to the Web tracking metrics collected in the Web logs is a key strength of our research project. By linking these data sources and triangulating results, we are able to arrive at a rich understanding of end-user online behavior. For example, regressions or path analyses are being used to determine how well individual differences predict task self-reports and actual usage behavior. Further, cluster or discriminant analysis techniques are being used to establish the characteristics of low, medium and high usage groups.

As mentioned above, two types of Web surveys were administered to people who consented to participate in the project. The first of these was an individual differences (user characteristics) survey that collected basic demographic information, personality traits, and technology background on each participant. Items were based on those found in a recent investigation by the authors (Hupfer & Detlor, 2006), the Georgia Institute of Technology's annual GVU WWW User Surveys (cc.gatech.edu/gvu/user_surveys), and Ford and Miller's (1996) scales that measure perceptions of Web-based information seeking (see Appendix A for a list of the individual differences questions used in this project).

The second type of survey was a brief user task evaluation. Eliciting attitudes towards conducting a specific online activity and perceptions of the task's importance, these surveys were administered after participants had completed a given activity on the portal interface (see Appendix B for the actual questions used to measure participants' attitudes toward and perceptions of tasks). These data were collected during short time windows pertaining to a few specific portal activities, such as paying a fine or purchasing a dog license.

In terms of Web tracking, like the two surveys, metrics were collected only from portal users who had agreed to participate in the study. The portal was designed to facilitate metric collection through third party applications hosted on the portal's back-end Web servers. Specific detail concerning the type of Web metrics that were collected and analyzed are discussed in the following section.

WEB METRICS TO COLLECT AND TECHNIQUES FOR ANALYZING THEM

In general, the various sources of Web logs can be classified as either server-level or client-level data sources (Srivastava et al., 2000). The primary metrics in our study were based on server-level data generated through a custom programmed server-side plug-in, and first-party cookies stored on the client-side. These metrics included the following: page attributes such as page views, page transitions, and HTTP referrer information; temporal attributes such as history time stamps, and session times; and visitor attributes such as user identification tags, and remote host information. Consequently, a composite of these primary Web log metrics provided us with the desired analytics output related to information seeking behaviors of end users. Table 1 shows the interrelationships among the various Web log metrics that we used and their sources, as well as the associations among the Web log metrics and their resulting composite analytics.

The advantage of using a customized server-side plug-in rather than other available logging methods such as Web server logs, client-side tracking utilities, and page tagging scripts is that the overall process affords more control for updates and modifications, and also entails less time and effort to clean the data and prepare it for further

Table 1. Summary of Web log sources, metrics, and composite analytics

Web Log Metrics	Sources of Web Log Metrics			Composite Analytics		
	Server Object	Session Cookies	Persistent Cookies	Visitor Footprints	Navigation Tracks	Information Seeking Trails
Page Views	•	•		•	•	•
Page Transitions	•				•	•
HTTP Referrer Information	•				•	•
History Time Stamps	•	•	•		•	•
Session Time	•	•			•	•
User Identification Tags	•		•	•		•
Remote Host Information	•			•		•

analysis (e-consultancy, 2003). Compared to Web server logs, custom logfiles suffer from fewer inaccuracies and redundancies (e-consultancy, 2003; Murata & Saito, 2006). With respect to control, a server-side plug-in allows greater autonomy than a client-side remote tracking utility where there is increased dependency on client platform capabilities and end user intentions (Winett, 1998). Similarly, vis-à-vis page tagging scripts where the deployed solution is typically outsourced to an application service provider (Beasley, 2002; e-consultancy, 2003), a server-side plug-in offers more control over the development and maintenance of an application interface.

In our study, the server-side plug-in was designed to poll several collections and properties of the server-side objects including request, response, and session objects in order to retrieve values pertaining to attributes of visited site pages, times spent on each page and in each user session, and unique visitor identification values stored in client-side cookies.

In addition to using a customized server-side plug-in for the collection of Web metrics data, our study utilized first-party cookies to track visitors. This method is regarded as more reliable than using third-party cookies used by hosted analytics vendors as recent Internet statistics show that 12%-17% of Internet users block third-party cookies while only 2%-5% block first-party cookies (WebTrends, 2005). By storing automatically generated unique identifiers on users' workstations, persistent cookies allow the identification of unique site visitors which can prove to be extremely valuable in determining the reach and audience penetration of a Website.

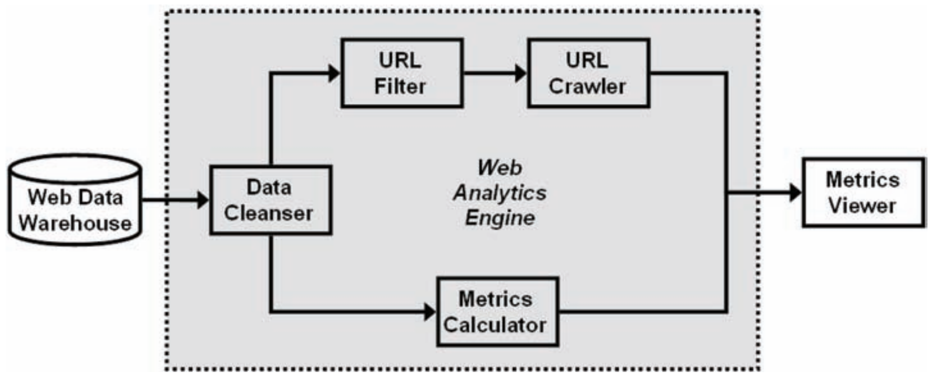
To generate the composite metrics described in Table 1, we created and utilized our own *Web analytics toolkit*. Overall, the collection of singular Web log metrics through server objects and cookies, as shown in Table 1, facilitates the formation of a Web data warehouse, which is regarded as the first step in devising a Web analytics toolkit (Sen et al., 2006). The formation of a Web data warehouse enables simple decision support services through channel traffic reports.

The second step in developing a Web analytics toolkit is to aid sophisticated visitor behavior tracking (Sen et al., 2006) which can only be enabled through additional structured statistical procedures and logic querying methods. In our research study, we did this by following the three phases suggested for Web usage mining studies, namely: 1) *pre-processing*, 2) *pattern discovery*, and 3) *pattern analysis* (Srivastava et al., 2000). The pre-processing phase cleanses, sorts and formats the raw data into organized segments of information (e.g., establishing sequence of activities through sorting first by cookie-based user identifiers and then by server session identifiers). This information feeds into the pattern discovery phase which converts raw logs into data abstractions that are pivotal to the analysis of usage patterns (e.g., deriving session length information from time stamps and page views). The pattern analysis phase calculates descriptive statistics and usage metrics that can help to identify different user clusters based on their patterns of information seeking. Overall, these three phases enable the transformation of dimensions of Web activities that are measurable into those that are meaningful within the context of the analysis of users' information seeking behaviors.

Unlike other Web analytics studies that undertake the development of a self-contained prototypical Web usage mining system as part of the overall research project (Srivastava et al., 2000; Wu et al., 1998; Zaiane et al., 1998), the Web analytics engine used in our study was based on a selection of self-programmed application macros, third party tools, and customized scripts.

Figure 2 shows a high level schematic of the Web analytics engine depicting the various functional modules that were used to operationalize the three phases of Web usage mining. While the data cleanser module enabled the pre-processing

Figure 2. Components of the Web analytics engine



phase, the URL filter and crawler, as well as the metrics calculator, enabled the pattern discovery and pattern analysis phases.

In terms of custom applications, server side scripts using SQL (structured query language) were deployed to extract data from tables in the data warehouse and to export into a format suitable for spreadsheets and statistical analysis applications. In the data cleanser module, Excel macros were utilized to cleanse the data and organize it into meaningful segments that were to be used in metrics calculations. The URL filter allowed us to form clusters of Webpages based on their frequency and mode of access. For instance, Website landing pages were identified by noting external referrer Websites, and search tools were recognized by query information such as keywords contained in URLs. Ultimately, a dictionary of landing pages was compiled to facilitate page lookups during the analysis of online user activities.

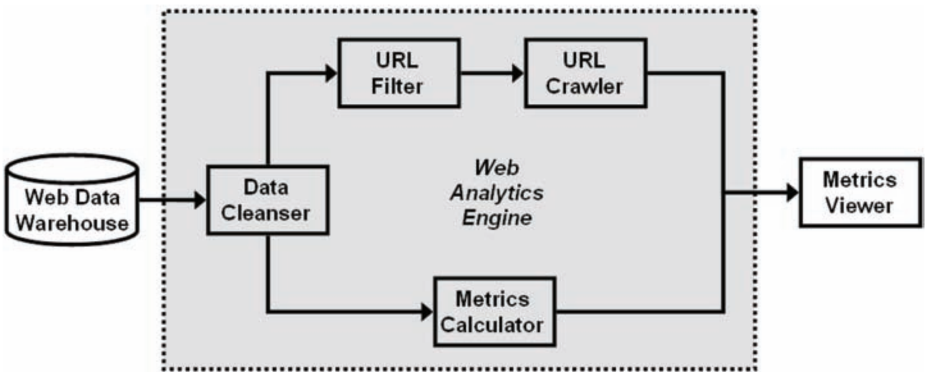
The information derived from the URL filter was further refined by mapping the URLs to the title of the Webpages. These titles were obtained by running the list of URLs through an Internet-based Web crawler utility which parses Webpages for various types of metadata. In this case, the only metadata that was of interest was the title tag pertaining to the Webpages referenced by the URLs. Finally, the metrics calculator module comprised spreadsheet functions and macros in Excel and analysis widgets in the SPSS statistical application. The metrics calculator was used primarily to report descriptive statistics and produce cluster analysis results that could be viewed in text, table or graphical formats through the metrics viewer module.

The Web analytics engine allowed us to perform a composite analysis of Web log metrics from server-side and client-side sources. Specifically, as shown in Table 1, the Web analytics engine allowed us to define a hierarchical view of user activities based on visitor footprints, navigation tracks, and information seeking trails.

At the lowest level of user information seeking pattern identification, a *visitor footprint* represents "a single clickstream record created by the interaction of the visitor with a page on a Website" (Sen et al., 2006). In establishing visitor footprints, various clickstream metrics such as unique cookie based user identification, server based session identification, time stamp information, referrer page URLs and destination page URLs can be compiled into simple records of online user activities. In our study, the data cleanser and URL crawler modules in the analytics engine were utilized to establish these visitor footprints. Figure 3 shows an extract of nine footprints pertaining to a sample user, User-34.

An aggregation of visitor footprints enables the formation of *navigation tracks* which provide

Figure 3. Extract of visitor footprints



a chronological history of a user's activities on a Website (Sen et al., 2006). In our study, the configuration of each visitor navigation track comprised the entry point to the Website, the path of Webpages that were traversed in each user session, along with the average time spent per page, content page requests per session, search queries per session, and the exit point from the Website. Custom programmed spreadsheet macros in the metrics calculator module of the analytics engine processed the visitor footprint information to reveal these navigation tracks. Figure 4 shows the extract of navigation tracks pertaining to the visitor footprints shown in Figure 3.

Based on information from visitor footprints and navigation tracks, *information seeking trails* characterize the deepest level of pattern discovery employed in our study. Information seeking trails can be discerned by using clustering algorithms which can group similar user beliefs, attitudes and behaviors (Sen et al., 2006). In our study, the information seeking trails were analyzed through composite analytics that acted as surrogate measures to identify scanning, searching and browsing

Figure 4. Extract of visitor navigation tracks

User Cookie ID	Server Session ID	Webpage Alias	Time Stamp
30570681-8208-40e7-8742-a4b5203e2c56	ugxqppfw51j3t4b0b51w2cqo	Portal Homepage	10/02/2006 8:08
30570681-8208-40e7-8742-a4b5203e2c56	ugxqppfw51j3t4b0b51w2cqo	City & Government Subpage 1	10/02/2006 8:11
30570681-8208-40e7-8742-a4b5203e2c56	ugxqppfw51j3t4b0b51w2cqo	City & Government Subpage 2	10/02/2006 8:15
30570681-8208-40e7-8742-a4b5203e2c56	fztodda1r1c2wq55wx2v2h55	Personalized Portal Homepage	28/09/2006 14:50
30570681-8208-40e7-8742-a4b5203e2c56	fztodda1r1c2wq55wx2v2h55	Education & Careers Subpage 1	28/09/2006 14:50
30570681-8208-40e7-8742-a4b5203e2c56	fztodda1r1c2wq55wx2v2h55	Search	28/09/2006 14:54
30570681-8208-40e7-8742-a4b5203e2c56	fztodda1r1c2wq55wx2v2h55	Search Results	29/09/2006 14:54
30570681-8208-40e7-8742-a4b5203e2c56	fztodda1r1c2wq55wx2v2h55	Education & Careers Subpage 2	28/09/2006 14:56
30570681-8208-40e7-8742-a4b5203e2c56	fztodda1r1c2wq55wx2v2h55	Portal Homepage	28/09/2006 15:01

User ID	Session ID	Webpage ID	Time on Page
User-34	Session-3-A	PH	180
User-34	Session-3-A	CG1	246
User-34	Session-3-A	CG2	
User-34	Session-4-A	PPH	62
User-34	Session-4-A	EC1	242
User-34	Session-4-A	SP	61
User-34	Session-4-A	SR	123
User-34	Session-4-A	EC2	301
User-34	Session-4-A	PH	

Figure 5. Extract of surrogate metrics and frequencies of archetypical information seeking modes

User ID	Track ID	Page Path	Entry Point	Exit Point	Session Length	Avg Time / Page	Page Requests	Search Queries
User-34	1	PH,CG1,CG2	PH	CG2	426	142	3	0
User-34	2	PPH,EC1,SP,SR,EC2,PH	PPH	PH	789	131.5	4	1

moves on Websites. These information seeking moves were further assembled and classified into modes of information seeking such as undirected viewing, conditioned viewing, informal search, and formal search (Aguilar, 1988; Choo et al., 2000). Episodes of information seeking were analyzed using recursive procedures in spreadsheet macros that hinged on multiple passes through visitor footprints and navigation tracks. Figure 5 shows an extract of the additional surrogate metrics that were used to formulate the archetypical information seeking episodes.

Undirected viewing episodes can be identified by observing the most common entry points in visitor navigation tracks and calculating the average number of page requests and the average time spent on Webpages. Since users engaged in undirected viewing modes demonstrate broad scanning web moves (Choo et al., 2000), these modes typically consist of navigation tracks that start at main landing pages such as the Website homepages and sitemaps, and exhibit high average number of page requests and low average time spent per Webpage.

Conditioned viewing episodes can be tallied by identifying similarities in entry and exit points between visitor navigation tracks for the same user. This can be accomplished by calculating the dot product of page paths in visitor navigation paths. Since users engaged in conditioned viewing episodes are interested in selected topics and specific types of information (Choo et al., 2000), these episodes also will typically exhibit a lower ratio of search queries to page requests.

Informal search episodes show signs of unstructured search efforts (Choo et al., 2000) and can be identified by using surrogate measures such as observing navigation tracks that consist of references to external search engines, high average number of sessions with multiple search queries, and low average time spent per Webpage.

Formal search episodes are emblematic of users who make a purposeful and planned effort to acquire specific information (Choo et al., 2000). Using information available in visitor navigation tracks and other surrogate measures of information seeking modes, these episodes can be expected to exhibit a high average number of sessions that started with a search page, a low average number of page requests, and high average time spent per page.

The next step in our research study will involve further examination of the four archetypical episodes of information seeking outlined above. Specifically, we plan on linking the Web analytics information we collected in terms of visitor footprints, navigation tracks and information seeking trails to the data we collected through the individual differences questionnaire and task attitudes survey. Doing so will allow us to better understand the associations between Web behavior and the specific characteristics of users or groups of users.

ADDRESSING PRIVACY IN WEB INFORMATION SEEKING ANALYSIS

A final tip for any person wishing to conduct Web information seeking analysis in real-world settings (i.e., non-laboratory environments) is the need to address privacy concerns during the project's initial stages and to design Web data collection and analysis methods with privacy in mind. Ethics boards at academic institutions have always cautioned researchers about the need for anonymity and/or confidentiality. Academics, industry analysts and privacy advocates also have raised concerns about the vast amount of data that is collected using passive devices, such as adware, cookies, spyware and Web viruses, to record online behavior (Marshall & Swartwout, 2006). Such calls for privacy protection are warranted.

For example, the need to preserve privacy with Web log analysis became a highly public debate in the summer of 2006 after America-On-Line (AOL) posted query log data (approximately 20 million search inquiries obtained from over 650,000 users over a three month period) to a publicly accessible Website. These data, which were intended for academic use, assigned each user a unique ID and included the date and time of each query as well as addresses of Websites that were visited after searching was concluded. The AOL team that released the data intended to provide researchers with the opportunity to analyze search patterns and strategies over time without having to disclose any personally identifiable information.

Unfortunately, it rapidly became apparent that the specification of search parameters could in some cases permit the identification of individuals (Barbaro & Zeller, 2006; Hansell 2006a). These data were quickly removed from the site, but not before the data had been downloaded and circulated (Barbaro & Zeller, 2006). The ensuing media furor informed readers about how much data is stored by major search engines (Zeller, Jr. 2006a), how advertisers used search history for segmentation and targeting purposes (Hansell, 2006b) and also advised them as to how they could protect their identity online by using proxy servers and by deleting browser cookies (Biersdorfer, 2006). In late August 2006, AOL dismissed a researcher and project manager, and their chief technology officer resigned (Zeller, 2006b). AOL also announced plans for the implementation of new technologies that would protect privacy and restrict access. Academic opinion concerning the data's use has been divided; the very real privacy concerns have been acknowledged, but at the same time, the availability of a very large and current data set has immeasurable value for those who investigate personalization and information retrieval (Hafner, 2006).

With our own myhamilton.ca research project, we were very aware of the need to address privacy issues at the outset. In our initial discussions with the portal project team, we agreed that each user would be assigned a unique ID that would allow us to link tracking logs with survey data. Participant identities would not be released to us and the portal staff would handle the distribution of incentives to those who participated in our research. While these measures ensured anonymity, we later were confronted with the ramifications of municipal privacy legislation requirements. These concerns were further exacerbated by a portal privacy breach that resulted in the inadvertent broadcasting of people's personal dog licensing information and the attendant negative press in the local media. The ensuing privacy impact assessment conducted on our research project required documentation of the flow and storage of both the survey and tracking data. This documentation included the names of all tables where data was stored, a description of the programs that accessed and updated the data, and the process by which the data would be sent to our research team. The city also planned to delete all data once our research team confirmed its safe receipt at the conclusion of the data collection period.

The city's privacy assessment identified one particularly important outstanding issue in that the city had no means to control or prevent the possible misuse of the survey's demographic information and Web tracking information, both of which resided in the city's databases. This situation arose because participants' unique IDs were stored within the city's databases and thereby created the potential for unauthorized linking of this information. Addressing this concern required extensive and costly program code and database changes. Consequently, rather than storing participants' unique IDs within the user profile table in the city's database, unique IDs were embedded in session cookies that were transmitted between a user's computer and the City of Hamilton's servers. Using session cookies that contained unique IDs allowed us to track user behavior individually but eliminated the need to store a unique ID in a database resident on one of the city's servers that possibly could have been used as a foreign key to access and link together a person's private information.

Elimination of unique IDs within the user profile table also caused us to rethink the way we handled the distribution of participant incentives. Some mechanism was required to identify which participants had participated in the study so that we could contact them and distribute gift certificates. To facilitate this, a field was created in the user profile table to simply act as a flag that would indicate whether or not a particular user had agreed to participate in our research project. This flag allowed us to identify those people who participated in the study but it did not allow us to associate a person's research data with their personally identifying (contact) information.

These changes to our data collection methods required an additional review of our protocol by the McMaster research ethics board and the amendment of our call for participation and consent form in order to clarify privacy implications. Specifically, potential participants were informed that privacy had been built into the methods by which data would be collected and stored. In terms of databases, the unique ID would be stored only in three raw data tables that contained the Web tracking and survey data. No link would exist between these tables and any other contained in the myhamilton.ca database, and no personally identifiable information would be sent to the McMaster research team. All of these revisions, the additional ethics review, and further testing of our data collection instruments delayed the project's launch for over six months. A better understanding of privacy legislation and its impact for our research would have allowed us to avoid both the delay and the expense.

CONCLUSION

To provide insight into methods for conducting a Web information seeking analysis, we presented several tips. First, we raised awareness of the importance of individual differences and tasks in understanding information seeking behavior. We encourage practitioners and researchers to include the collection and analysis of task, individual differences, and behavioral data in any Web information seeking analysis design. The second tip spoke to devising methods that collect task, individual differences, and Web tracking data and provide a means to link these data sets together. We provided our own research project at myhamilton.ca as an example for others to follow in this regard. The two surveys (see Appendices A and B) may help others structure similar research instruments to collect data on tasks and individual differences. The third and fourth tips provided guidelines on which Web metrics to collect and how to go about analyzing them. Though many alternative Web metrics and methods of analysis exist, we anticipate that our description of the Web metrics we collected and our illustration of their analysis will assist others in their own investigations. Finally, we discussed the need to address privacy concerns right up-front in the collection of Web tracking information so as to avoid lengthy and costly delays in conducting a Web information seeking analysis. By paying attention to these five tips, both academic researchers and practitioners can ensure that their Website performance measurement initiatives run smoothly.

REFERENCES

Aguilar, F. J. (1988). *General managers in action*. New York: Oxford University Press.

Bailey, J. R. (1997). Need for cognition and response mode in the active construction of an information domain. *Journal of Economic Psychology*, *18*(1), 69–85. doi:10.1016/S0167-4870(96)00047-5

Barbaro, M., & Zeller, T., Jr. (2006). A face is exposed for AOL searcher No. 4417749. *The New York Times*, August 9, 1.

Beasley, C. (2002). *It's a hit! Gauging success through traffic analysis*, Retrieved May 7, 2007 from http://www.sitepoint.com/article/success-traffic-analysis

Bhatnagar, A., & Ghose, S. (2004). Online information search termination patterns across product categories and consumer demographics. *Journal of Retailing*, *80*(3), 221–228. doi:10.1016/j.jretai.2003.10.001

Biersdorffer, J. D. (2006). How to digitally hide (somewhat) in plain sight. *The New York Times*, August 12, 9.

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management*, *31*(2), 191–213. doi:10.1016/0306-4573(94)00041-Z

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306–307. doi:10.1207/s15327752jpa4803_13

Choo, C. W., Detlor, B., & Turnbull, D. (2000). *Web work: Information seeking and knowledge work on the world wide web*. Netherlands: Kluwer Academic Publishers.

Das, S., Echambadi, R., McCardle, M., & Luckett, M. (2003). The effect of interpersonal trust, need for cognition, and social loneliness on shopping, information seeking and surfing on the Web. *Marketing Letters*, *14*(3), 185–202. doi:10.1023/A:1027448801656

Dillon, A., & Watson, C. (1996). User analysis HCI. The historical lessons from individual differences research. *International Journal of Human-Computer Studies*, *45*(6), 619–638. doi:10.1006/ijhc.1996.0071

e-consultancy. (2003). *Web measurement and analytics*. London, UK: e-consultancy.

Ekstrom, R.B, French, J.W., & Harman, H.H. (1976). *Manual for kit of factor-referenced cognitive tests*, 109-113, 173-177.

Entwistle, H., & Tait, N. J. (1995). *The revised approaches to studying inventory*. Edinburgh: Centre for Research on Learning and Instruction, University of Edinburgh.

Fisher, K. E., Erdelez, S., & McKechnie, L. E. F. (Eds.). (2005). *Theories of information behavior*. Medford, NJ: Information Today.

Ford, N. (1986). Psychological determinants of information needs: A small-scale study of higher education students. *Journal of Librarianship*, *18*(1), 47–61. doi:10.1177/096100068601800103

Ford, N., & Miller, D. (1996). Gender differences in internet perception and use. In *Papers from the 3rd Electronic Library and Visual Information Research (ELVIRA) conference, April 30* (pp. 87-202). London: ASLIB.

Ford, N., Miller, D., & Moss, N. (2001). The role of individual differences in internet searching: An empirical study. *Journal of the American Society for Information Science and Technology*, *52*(12), 1049–1066. doi:10.1002/asi.1165

Ford, N., Miller, D., & Moss, N. (2003). Web search strategies and approaches to studying. *Journal of the American Society for Information Science and Technology, 54*(6), 473–489. doi:10.1002/asi.10233

Ford, N., Miller, D., & Moss, N. (2005a). Web search strategies and human individual differences: Cognitive and demographic factors, internet attitudes, and approaches. *Journal of the American Society for Information Science and Technology*, *56*(7), 741–756. doi:10.1002/asi.20168

Ford, N., Miller, D., & Moss, N. (2005b). Web search strategies and human individual differences: A combined analysis. *Journal of the American Society for Information Science and Technology*, *56*(7), 757–764. doi:10.1002/asi.20173

Gugerty, L., Treadaway C., & Rubinstein, J.S. (2006). Individual differences in internet search outcomes and processes. *CHI 2006*, April, 815-820.

Hafner, K. (2006). Researchers yearn to use AOL logs, but they hesitate. *The New York Times*, August 23, 1.

Hansell, S. (2006a). AOL removes search data on vast group of web users. *The New York Times*, August 8, 4.

Hansell, S. (2006b). Advertisers trace paths users leave on internet. *The New York Times*, August 15, 1.

Häubl, G., & Trifts, V. (2000). Consumer decision making in online shopping environments: The effect of interactive decision aids. *Marketing Science*, *19*(1), 4–21. doi:10.1287/mksc.19.1.4.15178

Heinström, J. (2005). Fast surfing, broad scanning and deep diving: The influence of personality and study approach on students' information-seeking behavior. *The Journal of Documentation*, *61*(2), 228–247. doi:10.1108/00220410510585205

Ho, S. Y. (2005). An exploratory study of using a user remote tracker to examine web users' personality traits. *ICEC*, *05*(August), 15–17.

Huneke, M. E., Cole, C., & Levin, I. P. (2004). How varying levels of knowledge and motivation affect search and confidence during consideration and choice. *Marketing Letters*, *15*(2&3), 67–79. doi:10.1023/B:MARK.0000047385.01483.19

Hupfer, M. E. (2001). *Self-Concept orientation and response to agentic and communal advertising messages*. Unpublished Doctoral Thesis, University of Alberta, Edmonton.

Hupfer, M. E., & Detlor, B. (2006). Gender and web information seeking: A self-concept orientation model. *Journal of the American Society for Information Science and Technology*, *57*(8), 1105–1115. doi:10.1002/asi.20379

Hupfer, M. E., & Detlor, B. (2007a). Beyond gender differences: Self-concept orientation and relationship-building applications on the Internet. *Journal of Business Research*, *60*(6), 613–619. doi:10.1016/j.jbusres.2006.06.010

Hupfer, M. E., & Detlor, B. (2007b). Sex, gender and self-concept: Predicting web shopping site design preferences. In Detlor, B., Hassanein, K., & Head M. (Eds.), *Proceedings of the 8th World Congress on the Management of E-Business (WCMeB)*, July 11-13, Toronto, Canada. Los Alamitos, California: IEEE Computer Society.

Marshall, K. P., & Swartwout, N. (2006). Marketing and internet professionals' fiduciary responsibility: A perspective on spyware. *Journal of Internet Commerce*, *5*(3), 109–128. doi:10.1300/J179v05n03_05

Martin, B. A. S., Sherrard, M. J., & Wentzel, D. (2005). The role of sensation seeking and need for cognition on web-site evaluations: A resource-matching perspective. *Psychology and Marketing*, *22*(2), 109–126. doi:10.1002/mar.20050

Moe, W. W. (2003). Buying, searching or browsing: differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*, *13*(1&2), 29–40.

Moorthy, S., Ratchford, B. T., & Talukdar, D. (1997). Consumer information search revisited: Theory and empirical analysis. *The Journal of Consumer Research*, *23*(4), 263–277. doi:10.1086/209482

Murata, T., & Saito, K. (2006). Extracting users' interests from web log data. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'06).*

Palmquist, R. A., & Kim, K. S. (2000). Cognitive style and on-line database search experience as predictors of web search performance. *Journal of the American Society for Information Science American Society for Information Science*, *51*(6), 558–566. doi:10.1002/(SICI)1097-4571(2000)51:6<558::AID-ASI7>3.0.CO;2-9

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993), *The adaptive decision maker*. Cambridge UK: Cambridge University Press.

Riding, R. J., & Cheema, I. (1991). Cognitive styles – an overview and integration. *Educational Psychology*, *11*, 193–215. doi:10.1080/0144341910110301

Roy, M., & Chi, M. T. H. (2003). Gender differences in patterns of searching the web. *Journal of Educational Computing Research*, *29*(3), 335–348. doi:10.2190/7BR8-VXA0-07A7-8AVN

Saracevic, T. (1991). Individual differences in organizing, searching and retrieving information. *Proceedings of the American Society for Information Science*, *28*, 82–86.

Sen, A., Dacin, P. A., & Pattichis, C. (2006). Current Trends in Web Data Analysis. *Communications of the ACM*, *49*(11), 85–91. doi:10.1145/1167838.1167842

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations*, *1*(2), 12–23. doi:10.1145/846183.846188

Toms, E., & Trifts, V. (2006). When limited cognitive resources meet unlimited information sources: Factors influencing consumers' allocation of web-based search effort. *Proceedings of the 7th World Congress on the Management of eBusiness*, July, Halifax, Canada.

Tuten, T. L., & Bosnjak, M. (2001). Understanding differences in web usage: The role of need for cognition and the five-factor model of personality. *Social Behavior and Personality*, *29*(4), 391–398. doi:10.2224/sbp.2001.29.4.391

Vakkari, P. (1999). Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval. *Information Processing & Management*, *35*(6), 819–837. doi:10.1016/S0306-4573(99)00028-X

Wang, P., Hawk, W. B., & Tenopir, C. (2000). Users' interaction with world wide web resources: An exploratory study using a holistic approach. *Information Processing & Management*, *36*, 229–251. doi:10.1016/S0306-4573(99)00059-X

WebTrends. (2005). *Best practices for accurate web analytics: Avoiding third-party cookie rejection and deletion*. Portland, OR: WebTrends

Wildemuth, B., & Hughes, A. (2005). Perspectives on the tasks in which information behaviors are embedded. In Fisher, K.E., Erdelez, S. & McKechnie, L. (Eds.), *Theories of information behavior*. Medford, NJ: Information Today, for the American Society for Information Science & Technology, 275-279.

Wilson, T. D. (1999). Models in information behaviour research. *The Journal of Documentation*, *55*(3), 249–270. doi:10.1108/EUM0000000007145

Wilson, T. D., & Walsh, C. (1996), *Information behavior: An interdisciplinary perspective*. Sheffield: University of Sheffield Department of Information Studies.

Winett, B. (1998). *Tracking your visitors*. Retrieved May 10, 2007, from WebMonkey Tracking Tutorial, http://www.webmonkey.com/e-business/tracking/tutorials/tutorial2.html.

Wu, K.-l., Yu, P. S., & Ballman, A. (1998). Speedtracer: A web usage mining and analysis tool. *IBM Systems Journal*, *37*(1), 89–105.

Zaiane, O. R., Xin, M., & Han, J. (1998). Discovering web access patterns and trends by applying olap and data mining technology on web logs. *Proceedings of the IEEE Forum on Reasearch and Technology Advances in Digital Libraries*, Santa Barbara, CA.

Zeller, T., Jr. (2006a). Privacy vs. viewing the internet user as a commodity. *The New York Times*, August 12, 1.

Zeller, T., Jr. (2006b). AOL acts on release of data. *The New York Times*, August 22, 1.

KEY TERMS

Individual Differences: The demographic and psychological characteristics of people that distinguish one person from another.

Information Seeking Behavior: Refers to how people seek information in different contexts (Fisher, Erdelez & McKechnie, 2005).

Task: In this chapter, refers to the information seeking task an individual user experiences that instills a need for information and motivates the user to satisfy this information need through some sort of information seeking behavior. Task is the context surrounding a person's information need.

Web Information Seeking Behavior: Refers to information seeking behaviors that occur over the Web. Choo, Detlor & Turnbull (2000) identify four main modes of information seeking on the Web ranging from wayward browsing to goal-directed search (undirected viewing, conditioned viewing, informal search, and formal search) where each mode is characterized by predominant information seeking moves or activities (undirected viewing: starting and chaining; conditioned viewing: browsing and differentiating; informal search: differentiating, monitoring, and extracting; and formal search: monitoring and extracting).

Web Metrics: Pertains to the measures by which to assess a person's Web information seeking behavior or to assess and monitor activity on a Website. Examples of commonly used Web metrics include page views, page transitions, and session times.

Web Tracking: Refers to the automated collection of Web information seeking behavioral data.

APPENDIX A: INDIVIDUAL DIFFERENCES QUESTIONNAIRE

Note that some of these questions and their responses pertain to the City of Hamilton and a Canadian context. Modifications would be necessary if administered to a different sample population.

AGE

How old are you?

- Under 18
- 18-24
- 25-29
- 30-34
- 35-39
- 40-44
- 44-49
- 50-54
- 55-59
- 60-64
- 65+
- Prefer not to say

SEX

Are you?

- Male
- Female
- Prefer not to say

RACE

How would you classify yourself?

- Aboriginal
- Asian
- Black
- East Indian
- Hispanic/Latino
- Middle Eastern
- Multi-Racial

West Indian
White
Other
Prefer not to say

KIND OF AREA YOU LIVE IN

Which of the following best describes the area you live in?

Urban
Suburban
Rural
Prefer not to say

Location in Hamilton

Which geographical area of Hamilton best categorizes the area you live in?

Hamilton East
Hamilton Central
Hamilton West
Dundas
Ancaster
Flamborough
Stoney Creek
Mountain West
Mountain East
Mountain Central
Other
Prefer not to say

Language

At a general conversational level, which of the following languages do you speak? Indicate all that apply.

English
French
Arabic
Bengali
Cantonese
German
Hindi/Urdu

Indonesian
Italian
Japanese
Mandarin
Polish
Portuguese
Russian
Spanish
Other
Prefer not to say

Marital Status

What is your current marital status?

Single (never married)
Co-habiting
Married
Separated/Divorced
Widow/Widower
Prefer not to say

Education

What is your highest education level obtained?

Primary School
Secondary (High) School
Vocational/Technical/Community College
University Undergraduate (bachelors degree)
University Graduate (masters/professional/doctoral degree)
Prefer not to say

Occupation

Your current primary employment status is

Employed
Unemployed
Student
Retired
Other
Prefer not to say

Occupation (cont'd)

Which of the following categories best describes the **industry** you primarily work in (regardless of your actual position)?

Agriculture
Arts, Entertainment, and Recreation
Broadcasting
Construction
Fishing & Hunting
Education
College, University, and Adult Education
Primary/Secondary (K-12) Education
Other Education Industry
Finance and Insurance
Forestry,
Government and Public Administration
Health Care and Social Assistance
Homemaker
Hotel and Food Service
Information Industry
Information Services and Data Processing
Other Information Industry
Legal Services
Manufacturing
Computer and Electronics Manufacturing
Other Manufacturing
Military
Mining
Publishing
Software
Telecommunications
Transportation and Warehousing
Real Estate, Rental and Leasing
Religious
Retail
Scientific or Technical Services
Utilities
Wholesale
Other Industry
Prefer not to say

Which of the following best describes your role in industry?

Administrative Staff
Consultant
Junior Management
Middle Management
Researcher
Self-employed/Partner
Skilled Laborer
Student
Support Staff
Temporary Employee
Trained Professional
Upper Management
Other
Prefer not to say

The organization you work for is in which of the following:

Public sector (e.g., government)
Not-for-profit sector
Other
Private sector (e.g. most businesses and individuals)
Don't know
Prefer not to say

Household Composition

What is the number of adults 18 and over living in your household?

0
1
2
3+
Prefer not not say

What is the number of children age 4 and under living in your household?:

0
1
2
3+
Prefer not not say

What is the number of children ages 5 to 9 living in your household?:

0
1
2
3+
Prefer not not say

What is the number of children ages 10 to 13 living in your household?:

0
1
2
3+
Prefer not not say

What is the number of children ages 14 to 17 living in your household?:

0
1
2
3+
Prefer not not say

Household Income

What is your household income before taxes?

Less than $20,000
$20,000-$39,999
$40,000-$59,000
$60,000-$79,999
$80,000-$99,999
$100,000-$119,999
$120,000-$139,999
$140,000-$159,999
$160,000-$179,999
$180,000-$199,999
More than $200,000
Prefer not to say

Years on the Internet

How long have you been using the Internet (including using email, gopher, ftp, etc.)?

Less than 6 months
6 to 12 months
1 to 3 years
4 to 6 years
7 years or more
Prefer not to say

Community Building

Complete the following sentence in the way that comes closest to your own views: 'Since getting on the Internet, I have...'

Become MORE connected with people like me
Become LESS connected with people like me
Become EQUALLY connected with people like me
Don't know
Prefer not to say

Self- versus Other-Orientation*(taken fromHupfer, 2001)*

Rate each item below according to how well you think these statements describe you.

(Utilize a 9-point scale: "never true of me" 1 2 3 4 5 6 7 8 9 "always true of me"; 0 = "Prefer not to say"):

I am a nurturing person
I am a self-sufficient person
I am understanding
I make my own choices
I am a compassionate person
I am my own person
I am self-reliant
I am sympathetic
I am sensitive to the needs of others
I am an independent person

Computer Proficiency

Indicate the extent to which you agree with the following statements

(Utilize a 7-point scale of “strongly disagree 1 2 3 4 5 6 7 strongly agree; 0 = “Prefer not to say”):

I am highly competent at...

... creating and editing documents in a word processor
... creating and maintaining electronic spreadsheets of data
... creating and maintaining data tables & records in a database progra
... sending and receiving email messages
... searching for information utilizing a Web search engine (like ‘Google’)

Web Skill Test

Which of the following have you done? (Check all that apply)

- Ordered a product / service from a business, government or educational entity by filling out a form on the web
- Made a purchase online for more than $100
- Created a Webpage
- Customized a Webpage for yourself (e.g. MyYahoo, CNN Custom News)
- Changed your browser’s “startup” or “home” page
- Changed your “cookie” preferences
- Participated in an online chat or discussion (not including email)
- Listened to a radio broadcast online
- Made a telephone call online
- Used a nationwide online directory to find an address or telephone number
- Taken a seminar or class about the Web or Internet
- Bought a book to learn more about the Web or Internet
- Did Internet banking
- Prefer not to say

Technology Comfort

How comfortable do you feel using computers, in general?

Very comfortable
Somewhat comfortable
Neither comfortable nor uncomfortable
Somewhat uncomfortable
Very uncomfortable
Prefer not to say

How comfortable do you feel using the Internet?

Very comfortable
Somewhat comfortable
Neither comfortable nor uncomfortable
Somewhat uncomfortable
Very uncomfortable
Prefer not to say

How satisfied are you with your current skills for using the Internet?

Very satisfied - I can do everything that I want to do
Somewhat satisfied - I can do most things I want to do
Neither satisfied nor unsatisfied
Unsatisfied - I can't do many things I would like to do
Very unsatisfied - I can't do most things I would like to do
Prefer not to say

Frequency of Accessing the Web from Different Locations

From home (including a home office)

Daily
Weekly
Monthly
Less than once a month
Never
Prefer not to say

From work:

Daily
Weekly
Monthly
Less than once a month
Never
Prefer not to say

From school:

Daily
Weekly
Monthly
Less than once a month
Never
Prefer not to say

From public terminals (e.g., a library terminal, public kiosk):

Daily
Weekly
Monthly
Less than once a month
Never
Prefer not to say

From other places:

Daily
Weekly
Monthly
Less than once a month
Never
Prefer not to say

Connection to the Internet

Which of the following connection speeds do you primarily use to connect to the Internet? (Round to the closest value if necessary.) If you access the Internet at home via a commercial provider, choose the speed from you to your Internet provider:

Regular dial-up (through your phone company)
DSL low-speed (through your phone company)
DSL high-speed (through your phone company)
Cable (through your cable provider)
Do not know
Prefer not to say

Number of personal computers

How many personal computers are in your household (including laptops, but not including electronic organizers)?:

0
1
2
3 or more
Prefer not not say

Perceptions of Web-based Information Seeking*(taken fromFord & Miller, 1996)*

Rate each item below according to how well you think these statements describe you.

(Utilize a 5-point scale: “strongly agree” 1 2 3 4 5 “strongly disagree”; 0 = “Prefer not to say”)

- I usually only look at things on the Internet that have been suggested to me.
- Despite its complexity, I generally manage to find my way around the Internet fairly effectively.
- I rarely find anything useful on the Internet.
- I usually manage to keep ‘on target’ and avoid too much irrelevant material when using the Internet.
- I’m prepared to plough through quite a lot of irrelevant information in case there’s something useful I might otherwise miss on the Internet.
- If I had to choose only one, I’d prefer keyword searching to browsing (hypertext) on the Internet.
- The Internet is too unstructured for my liking.
- I personally think that the graphical elements of the World Wide Web (i.e., pictures, icons, graphics, etc. as opposed to just text) make me much more likely to use the Internet than if it were just text-based.
- When I use the Internet, I feel as though I’m not as ‘in control’ as I would like.
- My advice to someone like me would be: The best way to learn to use the Internet is to explore everything broadly to get a comparative ‘feel’ of the various aspects/tools before getting down to mastering one in any depth.
- I tend to get lost when using the Internet.
- It’s best to use the Internet only when you have a well-defined plan (rather than just browsing around).

APPENDIX B: THE TASK SURVEY

Perception of task importance

For you personally,

1. How important is it to do this task online?
   “not important at all” 1 2 3 4 5 6 7 8 9 “very important”
2. How useful is it to do this task online?
   “not useful at all” 1 2 3 4 5 6 7 8 9 “very useful”
3. How critical is it do this task online?
   “not critical at all” 1 2 3 4 5 6 7 8 9 “very critical”

Attitude towards a particular task

4. I have positive feelings towards doing this task online.
   “strongly disagree” 1 2 3 4 5 6 7 8 9 “strongly agree”
5. The thought of doing this task online is appealing to me.
   “strongly disagree” 1 2 3 4 5 6 7 8 9 “strongly agree”
6. It is a good idea to be able to do that task online.
   “strongly disagree” 1 2 3 4 5 6 7 8 9 “strongly agree”

This work was previously published in Handbook of Research on Web Log Analysis, edited by B. Jansen; A. Spink; I. Taksa, pp. 256-283, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 1.8

A Proposed Template for the Evaluation of Web Design Strategies

Dimitrios Xanthidis
Ciber, University College London, UK, DEI College, Greece, & NYC College, Greece

David Nicholas
Ciber, University College London, UK

Paris Argyrides
University College London, UK

ABSTRACT

This chapter is the result of a two years effort to design a template aiming at standardizing, as much as such a task is feasible, the evaluation of Web sites. It is the product of a few publications in international conferences and journals. A thorough review of the international literature on the subject led the authors to conclude there is a very large number of opinions, thoughts and criteria from different professionals involved, directly or indirectly, with the process of designing a good Web site. To make matters even more complicated there are a number of different terms used by various scholars, scientists and professionals around the world that often refer to similar, if not the same, attributes of a Web site. However, it seems that all these differences could boil down to a systematic approach, here called evaluation template, of 53 points that the design strategies of the Web sites should be checked against. This template was tested on a significant number (232) of Web sites of Greek companies and proved it can be used to evaluate the quality of Web sites not only by technology experts but by non-experts alike. The evaluation template, suggested here, is by no means the solution to the problem of standardizing the process of evaluating a Web site but looking at other work done on the subject worldwide it is a step ahead.

DOI: 10.4018/978-1-60566-100-1.ch014

INTRODUCTION

Despite the fact that in many developing countries internet access and e-commerce was not introduced until very recently (Xanthidis and Nicholas, 2004), the world is, clearly, moving towards the digital platform (Figure 1 – Greece in the red circle) with a rapidly increasing number of companies in these countries hosting Web sites. The question is whether these Web sites meet the international standards concerning a site's functionality. A preliminary research of large on- or off-line universities' libraries and governmental Web sites revealed that there are a number of different professions involved in the process of designing a good Web site, e.g. managers, marketing people, information technology experts, lawyers, ethnologists, all having different opinions regarding the functionality and appearance of a good Web site. However, it also proved it is possible to design a template that incorporates the main points of all these different views.

A critical parameter affecting the way people interact with a Web site is its structure, overall design and layout. A poorly designed Web site might lead to accessibility problems as well as reduced interest to navigate through it causing reluctance to visit the site again. Even though professional firms design many Web sites, there is, still, a substantial amount of work done by people with limited knowledge on how a well-designed Web site should be regardless of the fact these technology experts have the know-how to build any e-commerce solution, from a simple to a very sophisticated one.

The main problem in most cases where an evaluation of a Web site is required is the lack of certain systematic methodology to follow. This is exactly what this chapter is all about. A simple yet comprehensive, straightforward yet seen from four different angles, approach of how a Web site could be evaluated based on a number of criteria gathered from many different sources, scholars, scientists and academics, worldwide. Noone can claim to have found the solution to a problem so complex that it involves a variety of different professions not limited to the information and telecommunications technology. However, the discussion that this template has be subject to in three international conferences leads the authors that it is, indeed, a small step ahead toward the solution to the problem.

BACKGROUND

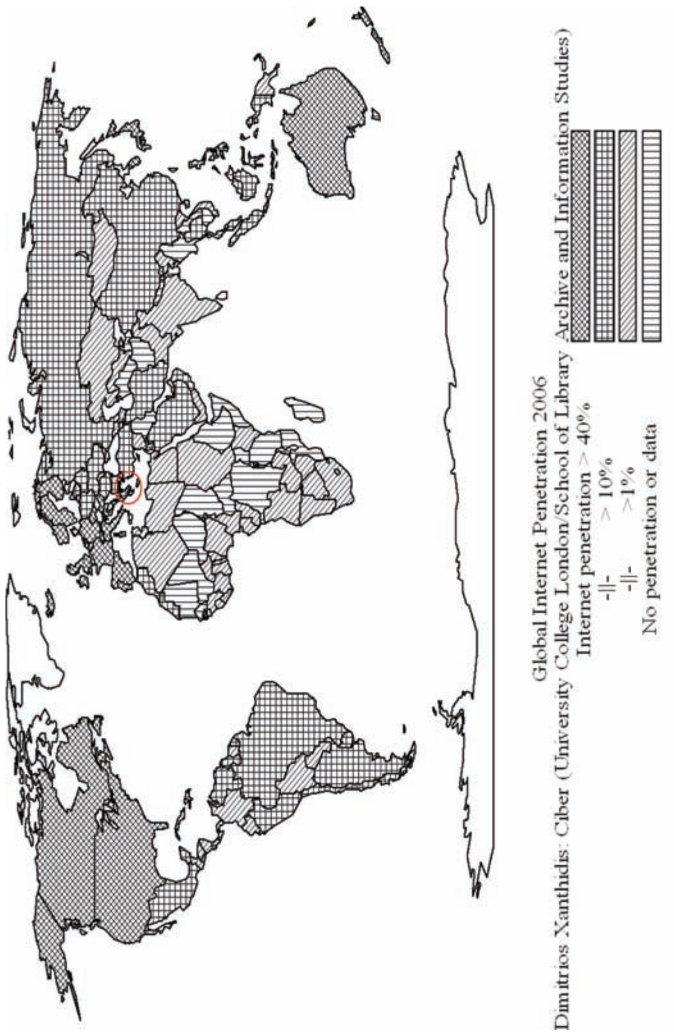
In a highly competitive environment, such as the Web, with billions of sites online and thousands added every day (D-Lib Magazine, 2003) designers/developers consider important that Web sites are attractive and inspire trust to the users so as to cause them to revisit. Therefore, once decisions have been made as of what features and functionalities the e-commerce solution should include the technology experts come into play in order to realize the solution into a good Web site. Based on extensive review of the international literature four categories of features and functionality are considered key for the successful implementation of an e-commerce strategy: user-interface design, globalization and customization, accessibility and availability, security and privacy. This section identifies, describes and analyzes the issues that comprise each category.

Stickiness

In the internet, billions of Web sites exist and even thousands are added every day (Trends…, 2003; Faster…, 2002). Designers and developers consider it important that their Web sites be attractive and inspire trust in users. This is reflected by the amount of time repeatedly spent visiting a site; a practice known as stickiness, a combination of "content, usability and personalization" and other issues, each one to be evaluated and measured on its own merit (Stickiness…, 2000).

Arguably the most important element of a Web site's attractiveness is its graphical user interface (GUI). The user interface should contain a mixture of graphics and text that could make it "appropriate" and "appealing" to any visitor

Figure 1. Global Internet penetration 2006



(Sutcliffe, 2002). Furthermore, Nielsen (1999) supports the assumption that the actual Web site content is among its most important attributes. This can be accomplished by avoiding annoying and distracting elements i.e. banners, marquees, graphics that "overshadow text", also the "overuse of animations", or unusual designs which do not follow certain accepted patterns of Web site design, etc (California, 2002; Interface…, 2002).

Another important guideline to follow when implementing a Web site is to avoid using scrolling mechanisms (Interface…, 2002). The reason is it has been proven that the user, usually, ignores hyperlinks difficult to be seen (Zhang *et al.*, 2000; Iowa, 1999). Several techniques are available to avoid such a negative aspect but two are the most popular among Web developers. The implementation of floating hyperlinks, i.e. hyperlinks that are programmed in such a way as to "follow" (floating on screen) the active/visible part of the particular Web page, is one of them. Another way is to limit the content of a Web page to that which can fit in one screen shot.

Regarding hyperlinks the Web developers should also be concerned with their placement and style. They should be easily accessible at a glance, and readable with their font preferences and their style not causing stress to the user when trying to distinguish them from other content. In case icons are used instead of text as hyperlinks, these should be, intuitively identifiable at least by information and communications experts, business people and marketing professional, if not everyone (Zhang *et al.*, 2000; California, 2002). These are simple human-computer interaction guidelines that apply in the case of Web sites as well.

One of the essential requirements for every Web site is to avoid having undefined objects as the targets of any of its hyperlinks, meaning that none of the hyperlinks should lead to a missing or dead link. Also the content of the target object of the hyperlinks should be relevant to the Web site's content (Helm, 2001; Internet…, 2002; California, 2002). The opposite could seriously damage the image of the company the Web site belongs as it would cause the visitor's lack of trust to the company severely reducing the chances to succeed in the e-commerce arena.

A very useful feature of every Web site is the implementaion of a site map of some sort. The reason for that is every user would like to know their location at any depth in the Web site they are surfing. For this purpose it would be useful if such mechanisms as site tree diagrams, composed of text or icon hyperlinks, were included displaying the user's location (Roy *et al.*, 2001; California, 2002). Not doing so will most likely cause time loss and tiredness from the part of visitor and, consequently, reluctance to revisit the site.

Another quite handy element of a Web site implemented all too often recently is the internal search engine. It is strongly recommended by technology experts that any Web site that includes articles or any type of information of significant size and/or value should either have this feature implemented or at least contain hyperlinks to the complete text if available (Nielsen, 1999). This trend was followed in the past by online newsgroups, journals and the like (Zhang *et al.*, 2000) but the newly developed Web sites, especially those that belong to large organization, incorporate this feature as well. Its advantage is that it radically reduces the time needed to find the things that one is looking for in the site adding positive points to the company's overall e-commerce strategy.

A critical element, often overlooked, of the quality of a Web site is the quality of information it provides to its visitors. The information displayed should be extracted from the most accredited and accurate source (Zhang *et al.*, 2000). Additionally, it should be updated regularly to eliminate the possibility of any outdated content, product specifications or "past-due elements" (Kateratanakul and Siau, 1999). Otherwise, failure to comply with the aforementioned criteria will cause a direct impact on the consumer's behavior towards the corporation i.e. misbelief of the information presented (Katerattanakul and Siau, 1999).

Finally, a very easily implemented feature with surprising positive influences on the visitors' behavior is any mechanism that would allow for visitor feedback. A Web site should allow visitors to comment on its strengths and weaknesses (Zhang *et al.*, 2000) as well as provide a two-way communication channel (Katerattanakul and Siau, 1999), by means of online surveys, e-mail link, feedback forms (Interface…, 2002; Evaluating…, 2004). It would help the visitor feel that it's ideas, comments, remarks are important and taken seriously under consideration.

Customization and Globalization

A Web site's success depends not only on features such as the aforementioned, but also on its global (or local) perspective, meaning its attractiveness and usefulness to populations in different geographic regions, a feature often referred to as globalization. The reason is people in different parts of the globe behave in different ways to various stimulations triggered by a Web site because they live in different cultures, practice different religions and communicate using various languages and symbols (Hanrahan and Kwok, 2001). Furthermore, it would be quite helpful to clarify, through the Web site, legal and/or customs particularities that might affect a visitor's engagement in e-commerce activity when using the Web site of a company based in another country than the one the individual resides. This factor could cause positive first impressions to some populations or culture shock to others. Therefore, the strategy followed in designing the site should address whole populations (Rutherford, 2000).

In order to achieve the above a Web designer/developer has to take under consideration certain realities. First, it is a false assumption that English is the internet's dominant language. Indeed several studies and surveys reveal other languages are used more by the internet users (Xanthidis and Nicholas, 2004; Communicate…, 2000). A company that wishes to attract multilingual audiences should include them in their e-commerce strategy as alternative target audience in addition to the English speaking internet users (Hanrahan and Kwok, 2001). Then, the color used for the background is another issue that affects users' positive or negative reactions. Recent surveys revealed that different colors have different connotations for different regional populations (Zhang *et al.*, 2000; Hanrahan and Kwok, 2001). One such example is the case of China where red is the dominant color (Anderson and Fell, 2003) and has several meanings including the symbolism that the person whose name is written in red is dead, or about to die, or on the other hand expresses such happy moments of life as birthdays, weddings etc. and in general is considered good luck.

A third key point to seriously consider when developing a Web site with global prospects is that users should also be informed about the legal framework related to taxation and import/export procedures in case products are to move in or out of a country (Rutherford, 2002). The reason for that is it is not rare to have sanctions imposed to a country from the part of the international community or have the a ban on certain products from the part of a country's government. Finally, online ordering forms should include a "universal address" format i.e. instead of using the U.S. term "zip codes" use the universal "postal codes" (Housley, 2004), and allow for the selection of various countries and their corresponding addresses.

On the other hand, the developer must address each individual's preferences as well. This is called customization or "adaptive interfaces" (Ardissono *et al.*, 2002) and refers to a Web site's built-in facility to identify a user's preferences even before any interaction takes place between the user and the site and present information in a way that is tailored by the users' preferences (Svet, 2003) and knowledge. Some scholars, scientists, professionals and generally experts in the field use the term personalization instead of customization. Most of the times they mean the same functionalities and just occasionaly they distinguish the two referring to the former when focusing in Web site content and to the latter when focusing in Web site interface and design. These variations in the way a Web site appears or as of the content presented to different user profiles can be created using forms, queries, cookies or other mechanisms and stored in databases.

A few added features could yield some more positive points towards achieving personalization and/or customization. For example, the different payment options accepted i.e. credit cards, money orders, various types of checks, etc., should be listed somewhere and a detailed description of the specific procedures followed both by the visitor and prospect customer of the company and by the representatives should be provided (Hanrahan and Kwok, 2001; Housley, 2004). Also, a currency converter should be implemented to facilitate quick conversion between currencies (Housley, 2004; Hanrahan and Kwok, 2001). These are very easily implemented features and quite useful helping visitors' decision towards engage in e-commerce transaction but are, once again, quite often overlooked.

Requirements, Accessibility

The introduction of new types of electronic devices such as personal digital assistants (PDAs), new generation mobile phones, also called cellular phones, with embedded internet capabilities, etc., in addition to the personal computers and laptops (notebooks) all having different abilities of presenting content to their user has lead the companies to find ways to make their Web sites available to different platforms and operating systems. The advent of mobile commerce in several digital economies worldwide has just stressed this need even further.

The main problem to be tackled is that a Web site display is primarily hardware/software dependent. Imagine the enormous problems that arise when a Web site designed in 1200x1600 in 32-bit color is viewed in the 120x240, 256-color display of a PDA or a mobile phone. It would, simply, be unreadable. In order to address this issue professionals suggest the design of a Web site that is available and its content prenentable by any type of electronic device. Currently, it appears that there are no tools available that actually convert a Web site's layout for different device usage but there are tools that help create accessible code [Macromedia…, 2004; Bohman, 2003; Sullivan, 2004). The recommended way of dealing with platform diversity is to separate layout design and content by having different layout templates, or "style sheets" for the same content, depending on the specific device (Nielsen, 1999; Accessibility…, 2004).

Closely related to the type of electronic device, the operating system and the browser platform, but mainly to the internet connection speed available is the time required for a Web page of a Web site to be loaded. This is a crucial factor when evaluating the Web site. Slow response speed, i.e. the time required to load a page inside a Web site, of more than 7-10 seconds, could be annoying (California, 2002) and discourage internet users to revisit.

No doubt the biggest challenge for every Web site designer/developer is to implement it in such a way as to make it accessible for persons with disabilities. A Web site should be designed not only with the media through which it is viewed in mind but its viewers as well. People with disabilities such as limited vision, hearing or mobility could find it difficult to navigate in a site filled with "graphically intense" content i.e. fairly large amounts of different colors and graphics. Too much color could be a problem for a person who is "visually impaired" person since it may make the content hard to read (Universal…, 2004). For that reason it is suggested to host a second or even third version of the Web site, customized for persons with disabilities (McManis *et al.*, 2001).

There are a number of guidelines to follow for versions of Web sites for viewers with various types of disabilities. The first is that if tables are used in which text is contained, then a logical grouping must be followed since most screen readers and magnification software may not read the text correctly. Text in cells must be separated into paragraphs to assist reading software (Making…, 2002; Pyatt, 2004). Additionally, include the functionality of providing textual description of images and "non-textual elements" in the case where browsers are configured not to display images, or the person who is using the browser uses a screen reader that cannot "read" images (Making…, 2002).

In the most likely event that a Web site includes "motion and animation" this should be at a frequency of less than 2Hz or more than 55Hz, as within this frequency range an animation may "trigger epileptic seizures" to individuals with related health problems (Making…, 2002; 30; Iowa, 1999). Finally, it would be quite useful to ensure that the site contains a mechanism that may trace the utilization of a screen reader or any other related tool and/or has also the ability to identify a change in language and adjust accordingly (Making…, 2002).

Some other issues related to accessibility and availability are the following:

- Text that is available through JavaScript i.e. pop-up windows, should also be available for users with JavaScript disabled and for users utilizing screen readers since screen readers may not read text contained in JavaScript (Pyatt, 2004).
- Frequently Asked Questions (FAQ) should be implemented together with a Help Topics feature. Accessibility covers matters that deal with user problems, not just hardware. Online guides, help topics and support reduce user stress to solve potential problems and increase the ease of navigating the Web site (Roy et al., 2001).

The main problem with the implementation of the features mentioned just above is the overall costs associated with it both in terms of time but most important in terms of money. To make matters worse the returns of investment are far from sufficient enough as the internet users' population that would be addressed by such sophisticated sites is a very small part of the overall online population. This is the main reason even large multinational corporations are reluctant in realizing such types of Web sites despite the various directives towards this goal from government organizations or nations like the European Commission and the U.S. department of Justice just to name a couple.

Security, Privacy, Legal and Ethical Issues Involved in Internet Marketing

Security, privacy, legalities and ethics are probably the most discussed technologies issues, nowadays. Several studies were conducted to clarify how they affect large corporations' successful or failed strategies to attract digital consumers (Privacy…, 2002). Currently, the most effective way to tackle this problem is to apply available mechanisms, in the form of software packages, aiming to protecting and securing valuable and sensitive data and restricting access to vulnerable systems (Benjamin *et al.,* 1998).

Unfortunately, it is a proven fact that there is no bulletproof mechanism to ensure complete defense against the various types of threats e.g. spyware, viruses including Trojan horses among other malicious programs/scripts, adware, even cookies used improperly by unauthorized people, etc. Then, it is on the developer's judgment to decide which of the available mechanisms should be used and how to ensure the success of internet strategies without compromising their visitors' privacy and keeping security, legal and ethical issues properly addressed. This is a goal very difficult to achieve. On one hand the more visitors' data a company keeps recorded the more effectively it will present a customized Web site tailored to their needs. On the other hand this causes many compromises in terms of the visitors' personal privacy and quite possibly of their personal computer security from malicious software. The following guidelines address the majority of these issues except all those that are not directly connected to the Web e.g. television/radio marketing.

First, in order to ensure the visitor's privacy Web sites should not collect sensitive information from users' PCs without their consent. In those cases where it is decided as necessary to collect such information the visitors should be clearly informed as of the use of the information. Personal data is most often collected through the use of cookies, spyware, and other related mechanisms. It is not always possible to directly confirm existence of such mechanisms like spyware and computer viruses. However, there are a number of anti-spyware and anti-virus software available which can be used by internet users to quite successfully block any attempts to compromise their personal computers' security and their own privacy.

Second, aiming to further protect visitors' personal financial data e-commerce Web sites at the transaction level (Xanthidis and Nicholas, 2007) should implement security protocols and services, like SSL (Secure Socket Layer), SET (Secure Electronic Transaction) to name the most widely used, to ensure safe transactions over the

Internet (Verisign, 2002). Furthermore, in those cases where e-mails are exchanged between the company's representatives and their Web site's visitors it could help if the latters' e-mail addresses were protected or masked via some type of scripts, forms, buttons, etc. to help defend against spam bots' attempts to identify e-mail addresses while crawling the Web.

Looking from the opposite site that of the owner of the Web site things are a little more complicated in those cases where the site is somehow connected with the company's intranet. In that case, the company's information and communication technology experts must ensure that no unauthorized person is allowed to enter Web pages restricted to authorized access. This could be achieved by implementing log-in procedures to authenticate a visit to such pages.

Finally, in every Web site like in every other software a message should be present to inform the reader that the material is an official document bound by relevant international copyright laws (Fishman, 1994).

METHODOLOGY

The aim of this work was to develop a systematic methodology for the evaluation of Web sites which would be as simple as possible yet comprehensive enough to include all those elements discussed in the previous section with straightforward answers while the evaluation takes place. The authors, then, came up with an evaluation template which comprises of 53 questions divided, not equally, into four categories called dimensions each one tackling the issues discussed earlier. It was also decided to use only dichotomous questions and avoid all other types, i.e. multiple answer, semantic differential, likert scale valued, ranked ordered. The rationale behind this decision is the idea to have a template as binary in nature as possible, one that any individual could use, even not having a significant technology background, limiting subjective judgement as much as possible. Also, the authors wanted to design in a way to make the evaluation of any Web site, no matter how large or complicated, quite feasible as of the time required. The results can be seen in the next four tables (see Table 1, Table 2, Table 3, Table 4).

In order to test the template the authors selected for evaluation a quite significant in size sample of the Web sites of 232 medium-large companies from the 15 sectors of the Greek economy, public and private, either "local" (Greek) or international with subsidiaries in the country (see Appendix). It took between 15' and 20' to evaluate each site and the evaluation period started on 11/5/2005 and ended on 30/6/2005. The answers to the evaluation questions were quite straightforward, binary in nature, a fact that simplified the process. In several cases during the evaluation, i.e. appropriate and appealing, we found it was somewhat subjective to determine/measure the Web site but in broad lines our decision was in general agreement with what any person would value as an appropriate and appealing Web site. This holds true despite the fact that people coming from different backgrounds may view this issue under another perspective.

One aspect we found to be problematic was the measurement of the *time availability* of a Web site. This would be possible by monitoring the site's online status on a 24/7 basis or alternatively by examining a detailed specification and possible the logs files of the hosting server. Human resource and time constraints of this part of the study prohibited us from doing the first. Furthermore, the fact that the server and network specifications and/or their log files were not available to us did not permit to follow the second path.

ANALYSIS AND RESULTS

Web Site Design/Stickiness

Initially, concerning Web site design issues, the study showed that developers in Greece follow

Table 1. Evaluation template: Dimension I: Stickiness

1.	Lack of tendency to use scrolling mechanisms.		Yes = 1	No = 0
2.	Hyperlink placement/style			
	i.	Hyperlinks easily accessible (at a glance)?	Yes = 1	No = 0
	ii.	Presence of floating hyperlinks (embedded in bars)?		No = 0
	iii.	Font properties (name, size, bold/no bold, color) of the text hyperlinks distinguishing them from the rest of the text?	Yes = 1	No = 0
	iv.	Icons used in graphical type hyperlinks intuitively identifiable, i.e. do they represent the target object or are they misleading?	Yes = 1	No = 0
3.	Hyperlink target/content			
	i.	Tendency NOT to have dead hyperlinks in the site (use home page)?	Yes = 1	No = 0
	ii.	Hyperlinks lead to relevant pages?	Yes = 1	No = 0
4.	Site maps			
	i.	Presence of any type of site map, i.e. site tree diagram, drop-down menus, etc.?	Yes = 1	No = 0
	ii.	Mapping mechanisms informative as to the actual depth in which the user navigates?	Yes = 1	No = 0
5.	Web site user interface attractiveness			
	i.	Appropriate and appealing?	Yes = 1	No = 0
	ii.	Lack of distracting and annoying elements?	Yes = 1	No = 0
6.	Information quality and completeness			
	i.	Any "read more" hyperlinks available clarifying possibly broad, unclear or unknown topics to the reader?	Yes = 1	No = 0
	ii.	Is the information provided in the Web site signed and, thus, credible?	Yes = 1	No = 0
	iii.	Is the information provided updated on a reasonably expected timeliness?	Yes = 1	No = 0
	iv.	Any internal search engine available?	Yes = 1	No = 0
7.	Visitor's feedback enabled and online help available			
	i.	E-mail links available to the visitors?	Yes = 1	No = 0
	ii.	Online surveys available?	Yes = 1	No = 0
	iii.	Feedback forms available?	Yes = 1	No = 0
	iv.	On line help available (e.g. FAQs, etc.)?	Yes = 1	No = 0

the internationally accepted standards of what a nice looking Web site should be like (Figure 2). Indeed, the majority of developers (142/232; 61.21%) tend to avoid the use of scrolling mechanisms which proved to have a negative impact in the Web site's attractiveness. In all but 3 cases (229/232; 98.7%) the hyperlinks were found to lead to relevant pages and in all but 6 (226/232; 97.4%) the hyperlinks did not lead to a dead end. If icons are used to represent the hyperlinks, the selections were found to be intuitively identifiable in 216/232 cases (93.1%). When text was the basis for the hyperlinks, the font properties were quite helpful in distinguishing it from the rest of the objects on the Web sites in 204/232 of the cases (87.9%). In general, the hyperlinks could be found easily at a glance in 167/232 of the cases (71.9%).

Concerning the presence of any sort of distracting and annoying elements, findings were quite positive again as in 212/232 cases (91.3%) no such elements were found and in general the user interface of the Web sites was found to be appropriate and appealing in 215/232 cases (92.6%). The only problem related to site interface was that in the majority of the Web sites floating hyperlinks

Table 2. Evaluation template: Dimension II: Customization and globalization

1.	Languages supported (cocacola.com is worldwide)			
	i. English		Yes = 1	No = 0
	ii. Spanish		Yes = 1	No = 0
	iii. Chinese		Yes = 1	No = 0
	iv. French		Yes = 1	No = 0
	v. German		Yes = 1	No = 0
	vi. Other (Greek, Turkish, Arabic, Hebrew, Japanese, etc.)		Yes = 1	No = 0
2.	Colors used: Is Web site color related with the cultural background (Western, Asian, etc) of the targeted population?		Yes = 1	No = 0
3.	Issues related to globalization			
	i. Briefing/information provided concerning import/export and taxation issues?		Yes = 1	No = 0
	ii. Any restrictions applicable for a commodity to be exported/imported to/from certain countries?		Yes = 1	No = 0
	iii. Any list of countries to which import/export restrictions apply?		Yes = 1	No = 0
	iv. Any information provided about available shipping/ delivery options?		Yes = 1	No = 0
4.	Level of customization the Web site achieves			
	i. Level 0: No customization		Yes = 1	No = 0
	ii. Level 1: Content → display information based on previous user interaction and preferences stored in log files.		Yes = 1	No = 0
	iii. Level 2: Suggestive → display information on relevant or competitive commodities/services.		Yes = 1	No = 0
	iv. Level 3: Informative → display further clarifications on issues not in the sphere of the user's knowledge.		Yes = 1	No = 0
	v. Level 4: Design format → lets the user permanently decide the layout of the Web site as it appears in his/her browser.		Yes = 1	No = 0
	vi. Level 5: Language and Culture → identifies the user's of language and culture preference based on the IP address of the user's system or on the user's selection of a region/country from a map/list available.		Yes = 1	No = 0
5.	Payment – shipping/billing options			
	i. List of different payment options available?		Yes = 1	No = 0
	ii. Detailed description of each payment option available?		Yes = 1	No = 0
	iii. Currency converter available?		Yes = 1	No = 0
	iv. Use of the universal "postal code" instead of the regional "zip code"?		Yes = 1	No = 0

were not present (152/232; 65.5%).

The problems started when more technical details were evaluated (Figure 3). One of the central components of any Web site, the site map, was only found in 85/232 (36.6%) of the Web sites. More technical mapping mechanisms that could provide information as of the depth in which the visitor navigates were even more seldom utilized (10/232; 4.3%). As far as online help made available to visitors is concerned, unfortunately, such a feature was only available in one site (1/232; 0.4%). The same can be said of the availability – lack, rather – of online surveys with only 2 sites implementing such feedback mechanisms for the visitors (2/232; 0.8%).

The most surprising element of the findings, however, was that only 156/232 sites (67.2%) incorporated the very easy to implement and almost obvious to include e-mail link for the visitors and, also, less than half the Web sites (106/232; 45.6%) made a feedback form available to visitors. The information was updated quite regularly in 227/232

Table 3. Evaluation template: Dimension III: Accessibility, availability, hardware/software requirements

1.	Is the Web site accessible (Platform Compatibility)		Yes = 1	No = 0
2.	Is the Web site optimized for users with a mental or physical handicap?		Yes = 1	No = 0
3.	Hardware/ Software/ Network requirements			
	i.	Time required loading the Web site's home page?	<10" = 1	No = 0
	ii.	Web site displayed properly, i.e. no horizontal scrolling mechanisms, no twisting of objects, etc., in different display resolutions?	Yes = 1	No = 0
	iii.	Option to download and install "third party" components required to view the Web site, e.g. activeX, flash players, different fonts, etc.	Yes = 1	No = 0

Table 4. Evaluation template: Dimension IV: Security, privacy

1.	Security			
	i.	Authentication required to login into possible intranet part of the site?	Yes = 1	No = 0
	ii.	If transactional or interactive what protocol are they using (None, SSL, SET, Other)?	Yes = 1	No = 0
	iii.	What is the cipher strength?	Yes = 1	<128 = 0
	iv.	On-line anti-virus scanner available?	Yes = 1	No = 0
	v.	Web site expires after a pre-defined amount of idle time?	Yes = 1	No = 0
2.	Privacy			
	i.	Avoid the use of tracking/identifying mechanisms i.e. cookies, spyware, etc, without the consent of the user?	Yes = 1	No = 0
	ii.	Privacy statement?	Yes = 1	No = 0
	iii.	Masked e-mail addresses through scripts, forms, buttons, etc?	Yes = 1	No = 0

cases (97.8%), however, it is signed – and, hence, credible – in only 32/232 cases (13.8%). There is the projection of providing more information in 134/232 sites (57.8%) through the utilization of "read more" hyperlinks and 88/232 of those (38.0%) included an internal search engine.

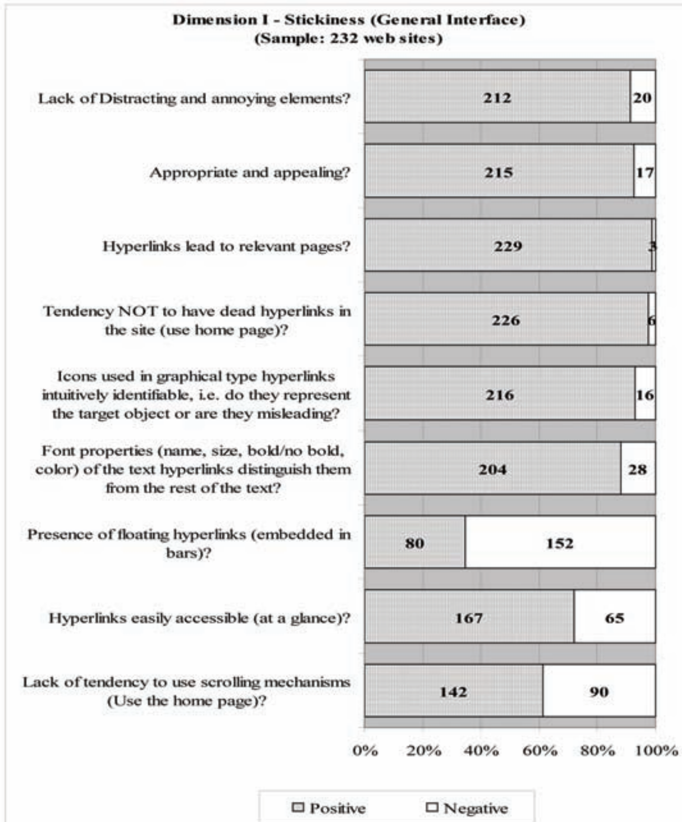
Customization and Globalization

It was quite obvious that except Greek (211/232; 90.95%), the only other language seriously considered and used when developing Web sites was English (178/232; 76.72%). All the rest, i.e. Spanish (7/232; 3.01%), French (7/232; 3.01%), German (8/232; 3.44%) and Chinese (3/232; 1.29%), included into our study, were seldom used mainly in specific cases of companies which probably had some business or other relations to certain groups of people in parts of the world.

In terms of the colors used, they were found to be appropriate (231/232; 99.57%) in connection with the populations of visitors targeted (based on the assumption that the language in which the text was written reflected the mother tongue and culture of the visitor targeted). However, it was also realized that this last conclusion was not based on enough evidence and it would be very interesting to see how things would change if, say, a large proportion of the targeted population were Chinese-speaking people.

Finally, only the executives of a handful of companies seem to take things seriously as only in 10/232 cases (4.31%) was there some kind of briefing about import/export and taxation matters, and only in 11/232 Web sites was information provided on shipping/delivery options (4.34%).

Figure 2. Web site evaluation from a technical viewpoint. Dimension 1 - Stickiness



No discussion can be made about providing lists of countries or commodities for which certain imports/exports restrictions apply (Figure 4).

Next, concerning customization features offered by the Web sites, the study showed that 89/232 (38.36%) Web sites provide no customization features whatsoever. Furthermore, no Web site was found displaying information based on previous user interaction and preferences stored in log files or in the form of cookies. A relatively significant (and rather unexpected considering the previous results) number (108/232; 46.55%) displayed information on relevant or competitive commodities/services placing those sites in the suggestive level of customization. A small percentage (32/232; 13.79%) displayed further clarifications on issues not in the sphere of the knowledge of the user (informative level). Finally, there were no Web sites found to allow changes in the format of the design, i.e. let the user permanently decide on the layout of the Web site as it appears on his/her client browser, neither did any attempt to identify the user's language and culture preferences based on the IP address of the user's system or on the user's selection of a region/country from a map/list available.

Last, as far as available payment and shipping/billing options, the general picture of the results of the study proved disappointing (Figure 5). Only 21/232 companies' Web sites (9.05%) provided a list of available payment options and just 3 of them (1.29%) described the steps to be followed for each option in some detail. The currency converter feature, often found in many international business-oriented Web sites, was present only in 3/232 (1.29%). Universal terminology for pay-

Figure 3. Web site evaluation from a technical viewpoint. Dimension 2 – Mapping and communication mechanisms – Informative structures

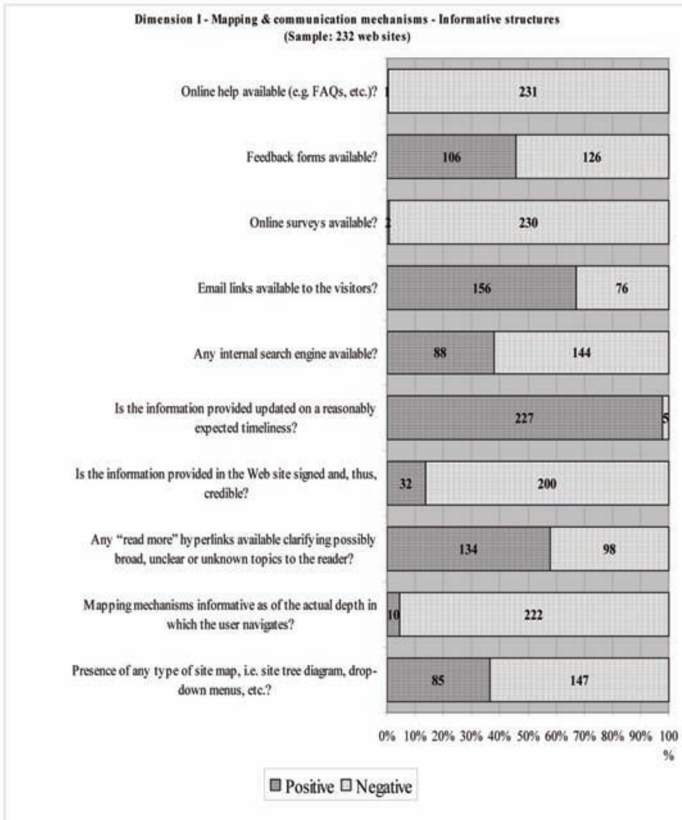
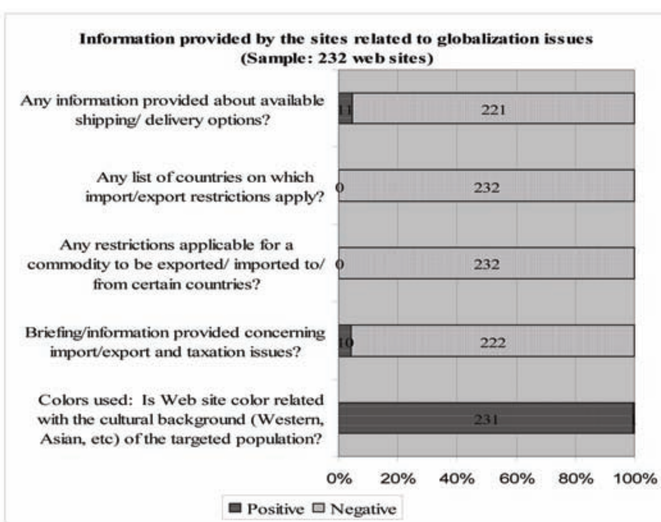


Figure 4. Information provided by the sites related to globalization issues



ment and shipping/delivery, e.g. the universal term "postal code" used instead of the regional term "zip code", appeared in 34/232 cases (14.66%).

Accessibility, Hard/ Software Requirements

The results of the evaluation concerning accessibility and hard/software requirements were radically different (Figure 6). In most cases (162/232; 69.83%) the sites were accessible from the common different platforms, i.e. Windows, Linux. The results were also positive when the time to load the home page was evaluated (less than 10'' was considered a reasonable time) with 214/232 (92.24%) succeeding, as well as the proper display of the page contents without object distortion or any other display anomalies under different display resolutions (226/232; 97.41%). On the negative side, unfortunately as expected, only in 1 case were the Web sites designed with the physically disabled people in mind. Also disappointing was the fact that in case third party tools were needed to run the Web pages of a Web site only in 15/232 cases (6.47%) were the options to download the respective tools given to the visitors.

Security, Privacy

The results of evaluating the Web sites with security and privacy in mind were completely disappointing (Figure 7).

On the one hand, security concerns should be tackled but very little was found be undertaken in this direction. Only 20/232 (8.62%) sites had some kind of authentication process running when a visitor requested to access to the companies'/organizations' intranet. Even less were the sites (16/232; 6.90%) protected by a security protocol like SSL, SET, etc. The cipher strength for the sites protected by such protocols was less than 128 (very low indeed) in 218/232 cases (6.03%). No on-line anti-virus scanner was available and no expiration time after a specific amount of idle time (as suggested for security purposes) was present and/or activated.

On the other hand, concerning the visitors' privacy, the study showed a tendency to avoid tracking/identifying mechanisms, like cookies, spy ware, etc. (231/232 cases; 9.57%) but there is serious doubt this was a conscious decision on the part of the companies' executives or lack of technical expertise required to implement it. Additionally, only 31/232 sites were found (13.36%) with a privacy statement (an utterly simple feature to realize), and only 56/232 (24.14%) cases of Web sites with masked e-mail addresses through scripts, forms, buttons, etc. for privacy-related reasons.

IMPLICATIONS

It is quite possible after all, although very difficult, to put together all bits and pieces that make up the puzzle of what suggests a good Web site, be it a simple one or a very sophisticated. The number of the elements to be evaluated is significant, of course, but limited enough to make the whole idea of evaluating a Web site feasible as of time and effort required. Furthermore, the nature of the questions of the template is such that allows even non-experts of the field of e-commerce to use it for their evaluation of their Web sites.

The template introduced in this chapter could be used as a "marking scheme" for the assessment of all different types of Web sites by information and communication technology (ICT) experts in a systematic way. It could also be used by business and other professionals as a suggestion, a rule of thumb indeed, of how could or should their companies' Web sites be implemented to meet their online strategies. The main reason the template can deliver both the aforementioned goals is that the questions to be answered carry on them the "correct" approaches of how to address each of the issues related to the development of a Web site. Therefore, it is quite straightforward that

Figure 5. More globalization issues faced

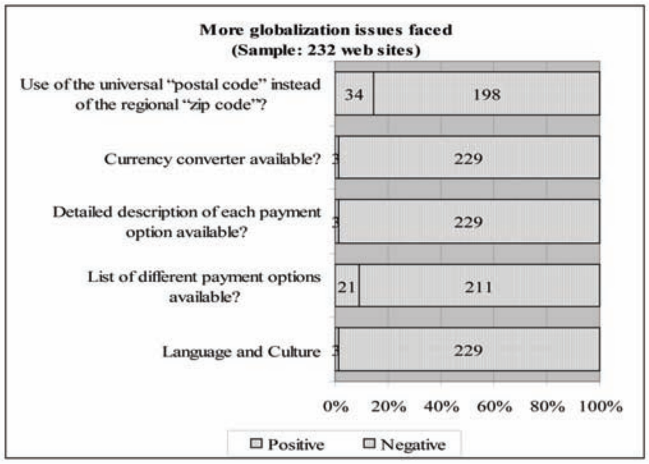
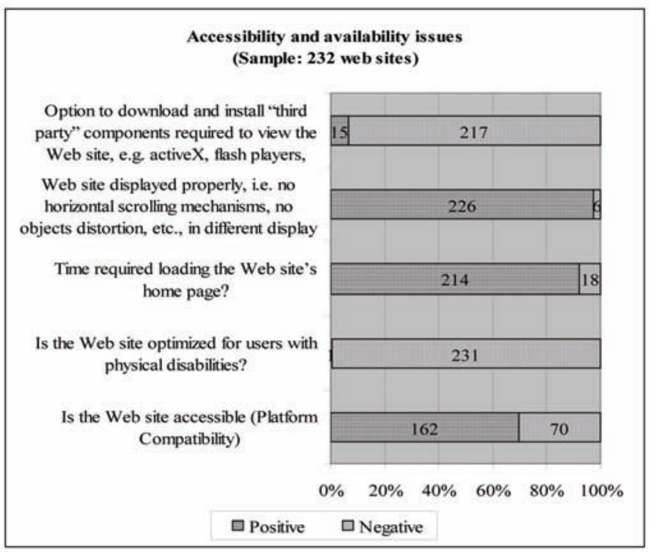


Figure 6. Accessibility and availability issues

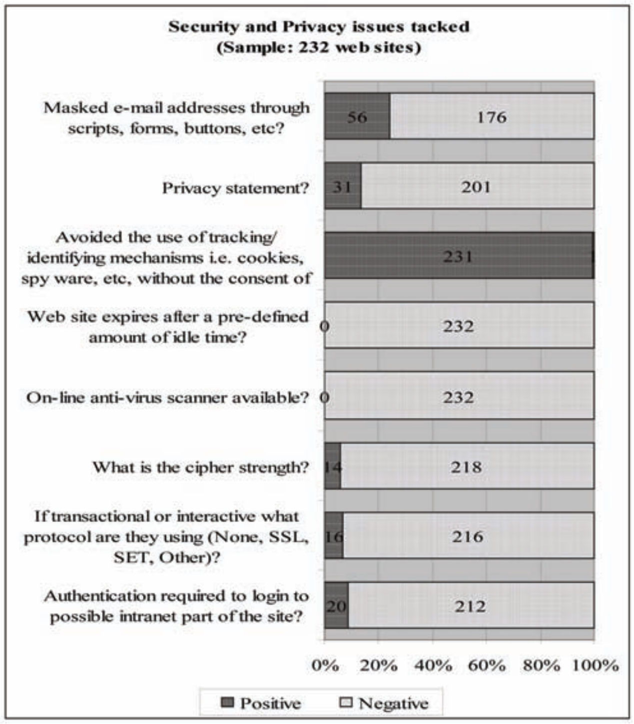


every positive answer adds one positive point to the overall design of the site with any negative answer pointing to its weaknesses.

LIMITATIONS: FURTHER RESEARCH

The authors used a number of Greek Web sites to test this systematic methodology that the suggested evaluation template represents. Despite the poor performance of the rather large local information society in e-commerce activity it was possible to use this template to draw useful conclusions for the quality of the Web sites and their effect in e-commerce growth in the country. The authors selected the Greek Web sites for reasons of physical proximity to the companies (Greek residence). However, it would be quite interesting to test the same methodology on the Web sites of companies in the developed digital economies and those of

Figure 7. Security and privacy issues tacked



the underdeveloped countries as well just to see if it is applicable in those cases as well. Also, the methodology was tested on medium-large companies' sites. It would help to see if and how much applicable it would be in the case of small companies as well.

Apparently, like in every other field of information and communications technology this effort is not over. This is only a measurement template. The next step, parallel to updating the template with new ideas and insight from professionals all over the world, is to work towards clarifying the metric against which the various results of measurement should be checked. There could be certain numerical results that distinguish a poor designed Web site from a better one and then from a very good one. Probably such a distinction is more complicated as it depends on certain variables including but not limited to the type of the Web site, e.g. information level up to full e-commerce solution, the progress towards the digital economy in the country of the company that owns the site, and others.

CONCLUSION

It should be underlined once again that the task of designing a comprehensive template/formula that could be used to evaluate Web sites looking from different viewpoints based on the various professions involved is extremely difficult, but proven feasible. The pieces of the puzzle are many and as the technology changes they change with it. However, there are a few facts that helped achieve our goal of designing the proposed evaluation template.

First, despite the quite significant, some could consider it rather large, number of 53 elements of a Web site to be evaluated they can be organized into categories each one addressing completely different viewpoints. This is the reason the authors called them dimension as it is like looking into the same structure from different angles. Second, it is proven that with just a little difficulty it is possible to answer the questions of the template in a straightforward binary way that leaves very little or no room to different and subjective interpretations of the quality of features implementing while addressing each specific issue. Third, it is proven possible to have a numerical result of the evaluation, although, by itself this cannot lead to safe conclusions as of the quality of the Web site overall but only to relevant suggestions. Finally, the time required to use this template to evaluate any Web site is very reasonable, less than 20' per site, regardless of the quality, type or size of the Web site.

The authors do not claim they found the ideal solution to the problem. What was proposed in this chapter is the result of the ongoing process of reviewing, adjusting and incorporating different ideas and thoughts into a template as little technical as possible in order for anyone to be able to use it even if that person is not an ICT expert. The authors opened a path which they believe could lead to a systematic way of assessing the quality of Web sites. The thing that encourages them to move further this type of research is the positive and insightful comments received in the conferences where this methodology is presented.

REFERENCES

Accessibility on the Web. (2004). International Center for Disability Resources on the Internet [online], Available: www.icdri.org [accessed: December 2004].

Anderson, B., & Fell, J. M. (US Government Export Portal) (2003). Taming the Red Dragon: Getting in and Succeeding in China, Available: www.export.gov/comm_svc/pdf/RedDragon.pdf [accessed: 2004].

Ardissono, L., Goy, A., Petrone, G., & Segnan, M. (2002, May). Personalization in B2C interaction. *Communications of the ACM*, *45*(5). doi:10.1145/506218.506245

Benjamin, R., Gladman, B., & Randel, B. (1998). Protecting IT systems from cyber crime. [accessed: December 2004]. *The Computer Journal*, *41*(7), 429–443. Available www.cs.ncl.ac.uk/research/pubs/trs/papers/631.ps. doi:10.1093/comjnl/41.7.429

Bohman, P. (2003). *Using Opera to check for accessibility*. Available: www.Webaim.org/techniques/articles/opera [accessed: 2004].

California State University. Sacramento (2002). *Web Design Principles*. Available: www.csus.edu/uccs/training/online/design/d_principles.htm.

Communicate in your customers' language (2000). WorldLingo Press Releases [online], Available: www.worldlingo.com/company_information/pr20000625_01.html [accessed: November 2004].

D-Lib Magazine. (2003). *Trends in the Evolution of the Public Web*. Available: www.dlib.org/dlib/april03/lavoie/04lavoie.html [accessed: 2004].

Evaluating Web Pages. Techniques to apply and questions to ask (2004). UC Berkeley – Teaching Library Internet Workshop [online], Available: www.lib.berkeley.edu/TeachingLib/Guides/Internet/Evaluate.html [accessed: December 2004].

Faster Growth Heightens Web Class Divide. (2002). TRN (Technology Research News) [online], Available: www.trnmag.com [accessed: December 2004].

Fishman, S. *Software Development: A legal guide*, Nolo Press, Berkeley, 1994, First Edition.

Hanrahan, M., Kwok, & Wei-Tai (2001). *A New Connectivity - How Technology is Changing the Marketing World*. Available: iaa2001.atalink.co.uk/html/p026.htm [accessed: 2004].

Helm, J. C. (2001). *Web based Application Quality Assurance*. Available: sce.cl.uh.edu/helm/Helm_html/papers/QW2001_PaperJune0101.pdf [accessed: 2004].

Hirsh, L. (2002, February 25). *Online Shopping and the Human Touch, E-Commerce Times: The E-Business and Technology Supersite* [online], Available: www.ecommercetimes.com/.../16483.html [accessed: December 2004].

Housley, S. (2004). *Web site Globalization*. Available: www.onevision.co.uk/xq/ASP/id.1072 [accessed: 2004].

Interface Design. (2002). *Web Style Guide 2nd Edition* [online], Available: www.Webstyleguide.com/interface/user-centered.hmtl [accessed: December 2004].

Internet Guide. (2002). Canada Government [online], Available: www.cio-dpi.gc.ca/ig-gi/m/wsm-msw/wsm-msw_e.asp [accessed: December 2004].

Iowa University. (1999). Designing accessible Web sites booklet, Available: www.uiowa.edu/infotech/WebAccess.htm [accessed: 2004].

Katerattanakul, P., & Siau, K. (1999). Measuring Information Quality of Web sites. In *Proceedings of the 20th International Conference of Information Systems* [online], Charlotte, North Carolina, United States, Available: portal.acm.org [accessed: December 2004].

Macromedia – Accessibility. (2004). Macromedia [online], Available: www.macromedia.com/macromedia/accessibility [accessed: December 2004].

Making your Web sites Accessible with Microsoft FrontPage (2002). Microsoft Corporation [online], Available: www.microsoft.com [accessed: December 2004].

McManis, B. L., Ryker, R., & Cox, K. C. (2001). An Examination of Web Usage in a global context. *Industrial Management & Data Systems*, *101*(9), 470–476. doi:10.1108/02635570110410654

Nielsen, J. (1999). User interface directions for the Web. *Communications of the ACM*, *42*(1), 65–72. doi:10.1145/291469.291470

Novell (2004). *Securing your Web site*. Novel [online], Available: www.novell.com/documentation/nw6p/index.html [accessed: December 2004].

Privacy and Data Sharing. A Performance and Innovation Report (2002). Cabinet Office (UK Government – Department of Constitutional Affairs) [online], Available: www.foi.gov.uk/sharing/pubs.htm [accessed: December 2004].

Pyatt, E. J., Pennsylvania State University (2004). Creating Accessible Web Sites, [accessed: Dec. 2004].

Roy, M. Christine, Dewit, O., Aubert, B. A. (2001). The impact of interface usability on trust in Web retailers. *Internet Research: Electronic Networking Applications and Policy*, *11*(5), 388–398. doi:10.1108/10662240110410165

Rutherford, E. (2002). *How to Avoid Global Website Disasters*. Available: www.cio.com/research/.../111400_disaster.html [accessed:2004].

Stickiness in an eTailing environment (2000). Northwestern University [online], Available: www.medill.northwestern.edu/imc/studentwork/projects/Sticky/index.htm [accessed: November 2004].

Sullivan, N. (2004). *Why design accessible Websites*. PostDiluvian Org. [online], Available: postdiluvian.org/~nicole/work/why.php [accessed: December 2004].

Sutcliffe, A. (2002). Heuristic Evaluation of Website Attractiveness and Usability. *Proceedings of the 35th Hawaii International Conference on System Sciences* [online], Available: csdl2.computer.org/comp/proceedings/hicss/2002/1435/05/14350137.pdf [accessed: December 2004].

Svet, B., & the State University of New York, Buffalo. (2003). *Personalization and Customization in eCommerce*. Available: www.cse.buffalo.edu/~sbraynov/seninar2003/presenttations/lecture.pdf [accessed: 2004].

Thelwall, M. (2000). Commercial Web sites: lost in cyberspace? *Internet Research*, *10*(2), 150–159. doi:10.1108/10662240010322939

Trends in the Evolution of the Public Web. (2003). *D-Lib Magazine* [online], Available: www.dlib.org/dlib/april03/lavoie/04lavoie.html [accessed: November 2004].

Universal Web Site Accessibility Policy for State Web Sites – Version 4.0, (2004). State of Connecticut Web site Accessibility Committee [online], Available: www.cmac.state.ct.us/access/policies/accesspolicy40.html [accessed: December 2004].

Verisign (2002). *Guide to securing your Web site for business.* Verisign [online], Available: www.verisign.com/resources/gd/secureBusiness/secureBusiness.html [accessed: December 2004].

Xanthidis, D., & Nicholas, D. (2004). Evaluating internet usage and ecommerce growth in Greece. *Aslib Proceedings*, *56*(6), 356–366. doi:10.1108/00012530410570390

Xanthidis, D., & Nicholas, D. (2007). A pilot qualitative study of eCommerce practices and attitudes of medium-large companies' executives in Greece. *WSEAS Transactions on Information Science and Applications*, *4*(2), 354–361.

Zhang, X., Keeling, B. K., & Pavur, J. R. (2000). Information Quality of commercial Web sites Home Pages: An explorative analysis. *ACM* [online], Available: portal.acm.org [accessed: November 2004].

APPENDIX

Companies: http://www.presspoint.gr/sectors.asp

(Sources: ASE (Athens Stock Exchange): companies in blue background, Presspoint.gr: companies in white background)
Table 5.

Sector 1:	**Food, beverages and tobacco industry**	**Web Site**
1.	Coca Cola Greece S.A.	www.cocacola.gr/
2.	Chipita International A.B.&E.E.	www.chipita.com
3.	Katseli & Sons A.B.E.E.	www.katselis.gr
4.	Eurofarma A.B.E.E.	www.evrofarma.gr
5.	CMA	www.cma-greece.gr
6.	Diageo	www.diageo.com
7.	Intercatering	www.intercatering.gr/
8.	Kraft Foods	www.kraftfoods.gr/
9.	Nestle Hellas	www.nestle.gr/online
10.	Athens Beers S.A	www.amstel.gr/
11.	Alatini S.A.	www.allatini.com.gr
12.	Vasiliou Wines	www.vassilioudomaine.gr/
13.	Delta Ice Creams S.A.	www.delta.gr/
14.	Dodoni Ice-Creams A.B.E.E	www.dodoni.com.gr/
15.	ELAIS S.A. Olive products businesses	www.elais.gr/
16.	Pedestrians Union	www.pezaunion.gr/
17.	Thraki S.A	www.thraki-sa.gr
18.	INO S.A.	www.inowines.gr/
19.	Kanakis ST. A.B.E.E.	www.stelioskanakis.gr/
20.	Kri-Kri Milk Company A.B.E.E.	www.krikri.gr/
21.	Lazaridis Wines S.A.	www.domaine-lazaridi.gr/
22.	Mevgal S.A.	www.mevgal.gr/
23.	Mega Farm	www.megafarm.gr/
24.	Melisa - Kikizas	www.melissa-kikizas.gr/
25.	Mparmpa Stathis Foods S.A.	www.geniki-trofimon.gr/
26.	Louli Mills S.A	www.loulisgroup.com
27.	Nostimo A.E.B.E	www.musses.gr/
28.	Xifias Fish S.A	www.xifias.gr/
29.	Chatzikraniotis & Sons	www.xatzikranioti.gr/
Sector 2:	Chemical industries	Web Site
1.	Veterin A.B.E.E.	www.veterin.com/
2.	PLIAS A.B.&E.E.	www.plias.gr/
3.	Druckfarben Hellas S.A.B.E.	www.druckfarben.gr/

Sector 1:	Food, beverages and tobacco industry	Web Site
4.	Ballis Chemicals A.E.B.E.	www.ballis.gr/profile
5.	Kerakoll Hellas	www.kerakoll.com/
6.	Neochimiki L.V. Lavrentiadis A.B.E.E.	www.neochimiki-lavrentiadis.gr/
7.	Famar S.A.	www.famar.gr/
Sector 3:	Transport equipment manufacturing	Web Site
1.	Neorion Shipyards S.A	www.neorion-shipyards.gr/
2.	Sfakianakis S.A.B.E.	www.suzuki.gr/
3.	Petropoulos, P., S.A.&B.E.	www.petropoulos.com/
4.	Hyundai Hellas	www.hyundai.gr/hyundai/
Sector 4:	Financial sector	Web Site
1.	Alpha Leasing S.A.	www.alpha.gr/introen.html
2.	Progress Funds S.A.	www.progressfund.gr/
3.	New Millennium Investments A.E.E.X.	www.newmillenniumaeex.gr/
4.	Dias fund A.E.E.X.	www.diasfund.gr
5.	Altius Investments S.A.E.X.	www.altius.gr/
6.	Credit Petropoulakis	www.credit-sec.gr/
7.	Eurocapital Financial Services	www.athenstock.com/
8.	EuroXX Finance	www.euroxx.gr/
9.	Investor ΕΠΕ	www.investor.gr/
10.	Aspis Bank	www.aspissec.gr/
11.	Eurodynamics S.A.E.X.	www.eudynamics.gr/
Sector 5:	Insurance and pension funding services	Web Site
1.	Ethniki Insurance .Α.Ε.Γ.Α.	www.ethniki-asfalistiki.gr/
2.	Agrotiki Insurances S.A.	www.agroins.com/
3.	Phoenix Metrolife S.A	www.phoenix-metrolife.com/
4.	Aspis Pronia S.A. General Insurances	www.aspis.gr/
5.	Europisti S.A.Γ.A.	www.europisti.gr/
6.	Alico AIG Life	www.alico.gr/
7.	ING Hellas	www.ing.gr/
8.	International Life Group	www.inlife.gr/
9.	Megaservice Ltd.	www.megaservice.gr/
10.	Eurobrokers S.A.	www.eurobrokers.gr/
11.	Interamerican Insurance S.A.	www.interamerican.gr/
12.	Syneteristiki ΑΕΕΓΑ	www.syneteristiki.gr/
Sector 6:	IT services	Web Site
1.	Logicdis S.A.	www.logicdis.gr
2.	Ipirotiki Software & Publications S.A.	www.ipirotiki.gr/
3.	Compucon Computer Applications A.B.E.E.	www.compucon.gr/GR/
4.	Logismos S.A.	www.logismos.gr/
5.	01 Pliroforiki A.E.	www.01p.gr/
6.	ABC Professional Services S.A.	www.abc.gr/

Sector 1:	Food, beverages and tobacco industry	Web Site
7.	ACE Advanced Applications S.A.	www.ace.gr/
8.	ACOM S.A.B.E.	www.acom.gr/
9.	Actis Information Systems S.A.	www.actis.gr/
10.	Active Computer Systems Ε.Π.Ε	www.active.gr
11.	Adacom S.A	www.adacom.com
12.	Ahead Rm A.E	www.aheadrm.com
13.	Algo Systems S.A.	www.algo.com.gr
14.	Alpha Grissin INFOTECH S.A.	www.alphagrissin.gr/
15.	Alpha IT S.A	www.alphait.gr/
16.	Alphyra Hellas	www.alphyra.gr/home/
17.	Altasoft	www.altasoft.gr/
18.	American Computers & Engineers Hellas S.A.	www.ace-hellas.gr/
19.	Anixter Greece Network Systems	www.anixter.gr/
20.	Apollo	www.apollo.gr/
21.	Areia S.A.	www.areianet.gr/
22.	Arion Software	www.arion.gr/
23.	Binary Logic Computers Ε.Π.Ε	www.mmpi.net/
24.	Cardisoft S.A.	www.cardisoft.gr/
25.	Datablue S.A.	www.datablue.gr
26.	Hipac S.A.B.E.	www.hipac.gr/
27.	Infomap S.A.	www.infomap.gr/
28.	Mantis IT S.A.E.	www.mantis.gr/
29.	Gnomon IT S.A.	www.gnomon.com.gr/
Sector: 7	Communications and Telecommunications Services	Web Site
1.	Vodafone – Panafon S.A.E.T.	www.vodafone.gr
2.	Forthent S.A.	www.forthnet.gr
3.	Lanet S.A.B.E.T.	www.lannet.gr/
4.	Intersat S.A.	www.intersat.gr/
5.	Algonet Telecommunications S.A	www.algonet.gr
6.	Chorus Call Hellas S.A.	www.choruscall.com/
7.	Com-Tonet S.A.	www.com-tonet.gr
8.	Cosmoline S.A.	www.cosmoline.com/
9.	Hellas Sat S.A.	www.hellas-sat.net/
10.	OTEGlobe	www.oteglobe.gr/
11.	Plural A.E.T.B.E.	www.plural.gr/
12.	Stet Hellas Telecommunications A.E.B.E	www.tim.gr/
13.	Teledome	www.teledome.gr/
14.	Telepassport Hellas	www.telepassport.gr/
15.	Tellas Telecommunications S.A	www.tellas.gr/
16.	Unitel Hellas S.A	www.unitel.gr/
17.	Uunet Hellas	www.gr.uu.net/

Sector 1:	**Food, beverages and tobacco industry**	**Web Site**
18.	Vivodi Telecommunications S.A	www.vivodi.gr/
19.	VoiceWeb S.A.	www.voiceWeb.gr/
20.	Winet	www.winet.gr/
21.	Newsphone Hellas A.E.E.	www.newsphone.gr/
22.	Mediatel Telephone Information S.A.	www.mediatel.gr/
23.	Cosmote Mobile Telecommunications S.A.	www.cosmote.gr/
24.	Q-Telecom	www.myq.gr/
Sector 8:	Health and social services	Web Site
1.	IASO S.A.	www.iaso.gr
2.	Euromedica S.A	www.euromedica.com.gr/
3.	Ygeia Diagnostic & Therapeutic Center, Athens S.A	www.hygeia.gr/
4.	Medicon Hellas A.E.	www.mediconsa.com/
5.	Biorehab Hellas	www.biorehab.gr/
6.	Gerolimatos Group of Companies	www.gerolymatos.gr/
7.	Thessaloniki Psychiatric Hospital	www.psychothes.gr/
Sector 9:	Media and printing (newspapers)	Web Site
1.	Lampraki Press Group S.A.	www.dol.gr/
2.	Kathimerini S.A.	www.kathimerini.gr
3.	Inform P. Lykos S.A	www.lykos.gr/
4.	Imako Media Net Group S.A.	www.imako.gr/
5.	Technical Press S.A	www.technicalpress.gr/el/
6.	Alpha Satellite TV S.A.	www.alphatv.gr/
7.	Filalthis	www.filathlos.gr/
8.	MAD TV Productions	www.mad.gr/
9.	Ellinika Grammata Publications	www.ellinikagrammata.gr/
10.	Asfalisi Net	www.asfalisinet.gr
11.	Amalthia Publications S.A	www.euro2day.gr/
12.	Alter TV S.A.	www.alter.gr/
13.	Greek Radio Television S.A.	www.ert.gr/
14.	Click FM ΕΠΕ	www.klikfm.gr/
15.	OPAP S.A.	www.opap.gr/
16.	Mega TV S.A	www.megatv.com/
17.	Alupress S.A	www.alupress.gr/
18.	Biokosmos News	www.bioshop.gr/
19.	Compupress S.A.	www.compupress.gr/
20.	Direction Publications S.A	www.direction.gr/
21.	Europress Publications	www.europress.gr/
22.	Knowledge Systems S.A.	www.business2005.gr/
23.	Metamedia	www.metohos.com
24.	Motor Press Hellas AEE	www.chip.gr/
25.	Newspaper Direct Hellas	www.newspaperdirect.gr

Sector 1:	**Food, beverages and tobacco industry**	**Web Site**
26.	Northmedia Publications	www.city231.gr/
27.	Option Press S.A.	www.optionpress.gr/
28.	Smartpress S.A.	www.smartpress.gr/
29.	Travel Times Publishing ΕΠΕ	www.traveltimes.gr/
Sector 10:	Metal/machinery manufacturing – Mineral and Cement	Web Site
1.	ELVAL S.A.	www.elval.gr/
2.	Metal Company Arkadias Cr. Rokas A.B.E.E.	www.rokasgroup.gr/
3.	Profil Pipe Company S.A.	www.tzirakian.com/
4.	FITCO S.A.	www.fitco.gr/
5.	S & B Industrial Minerals S.A	www.s.andb.gr/
6.	Naxos Marbles AEBE	www.naxos-marble.com/
7.	Pavlidis Marbles & Granites S.A	www.pavlidismg.com/
8.	F.H.L. H. Kyriakidis Marbles - Granites A.B.E.E.	www.fhl.gr/
9.	Betanet A.B.E.E.	www.betanet.gr/
10.	Heracles Cements Group	www.aget.gr/
11.	Iktinos Hellas S.A.	www.iktinos.gr/
12.	Mathios Refractories S.A	www.mathios.gr/
13.	Sidma S.A.	www.sidma.gr/
14.	Metka Metal Constructions Hellas S.A.	www.metka.gr/
15.	Grecian Magnesite AMBNEE	www.grecianmagnesite.com/
16.	Spider Metal Industry S.A.	www.spidersa.com/
17.	MEVACO A.B.E.E.	www.mevaco.gr/
Sector 11:	Education	Web Site
1.	Alba	www.alba.edu.gr/
2.	Alexander	www.alexanderinst.gr/
3.	Andim	www.andim.gr/
4.	Business Training Center	www.btc.com.gr/
5.	Centre of European Management Studies (CEMS)	www.cems.gr/
6.	Compact S.A.	www.compact.gr/
7.	Delta Singular Training S.A.	www.ds-training.gr/
8.	Didacta Training Group	www.didacta.gr/
9.	ECDL Hellas S.A.	www.ecdl.gr/
10.	Icon International Training	www.icon.gr/
11.	Infotest	www.certification.gr/
12.	Inte*learn	www.intelearn.gr/
13.	ITEC S.A.	www.itec.edu/
14.	New York College	www.nyc.gr/
15.	Proseed	www.proseed.gr/
16.	Akmi IEK	www.iek-akmi.gr/
17.	American College, Hellas	www.acg.edu/
18.	Xinis Educational Group	www.xinis.com/

Sector 1:	Food, beverages and tobacco industry	Web Site
19.	Crete Educational Institutes	www.sport-tourism.com/
20.	Aegean University	www.aegean.gr/
21.	Omiros Training Group	www.omiros.gr/
22.	Futurekids S.A S.A.	www.futurekids.edu.gr/
23.	Piraeuos University, Department of Industrial Management and Technology	www.tex.unipi.gr/
Sector 12:	Retail	Web Site
1.	Atlantic Supermarket A.E.E	www.atlantic.gr/
2.	Duty Free Shops, Greece S.A.	www.dutyfreeshops.gr/
3.	Promota Hellas S.A	www.promota.gr/
4.	AS Company S.A.	www.ascompany.gr/index.jsp
5.	Expert Suppliers S.A.	www.experthellas.gr/
6.	Glorybook-Economist Co.,Ε.Π.Ε.	www.glorybook.gr/
7.	Metro AEBE	www.metro.com.gr/
8.	Metropolis	www.metropolis.gr/
9.	Moda Bagno N. Varveris S.A.	www.modabagno.gr/
10.	Multirama A.B.E.E.	www.multirama.gr/
11.	Oriflame	www.oriflame.gr/
12.	Sara Lee Coffee and Tea Hellas S.A.	www.bravo.gr/
13.	Tupperware Hellas S.A.	www.tupperware.gr/
14.	Vadas A.E.B.E.E.	www.vardas.gr/
15.	Vassilias S.A.	www.vassilias.gr/
16.	Eikona – Ixos A.E.E	www.e-h.gr/
17.	Electornici Athens A.E.E.	www.electroniki.gr/
18.	Interflora S.A	www.interflora.gr/
19.	Kotsobolos A.E.B.E.	www.kotsovolos.gr
20.	Marinopoulos Group	www.marinopoulos.gr/
21.	Plaisio Computers A.E.B.E.	www.plaisio.gr/
22.	Hatzigeorgiou S.A	www.hatz.gr/
Sector 13:	Tourism [Hotels]	Web Site
1.	Louis Hotels	www.louishotels.com/
2.	Loutraki Club Hotel Casino	www.clubhotelloutraki.gr/
3.	Metropolitan Hotel	www.chandris.gr/
4.	Park Hotel Athens	www.athensparkhotel.gr/
5.	Rodos Park Suites Hotel	www.rodospark.gr/
6.	Astir Palace Vouliagmenis A.Ξ.Ε.	www.astir.gr/
7.	GEKE S.A.	www.president.gr/
8.	Divans Hotels Group	www.divanis.gr/
9.	Ioniki Hotels S.A.	www.ionianhe.gr/
10.	Capsis Tourist Group S.A.	www.capsis.gr/
11.	Porto Carras	www.portocarras.com/

Sector 1:	**Food, beverages and tobacco industry**	**Web Site**
Sector 14:	Government	Web Site
1.	Greek Parliament	www.parliament.gr
2.	National Administration Center	www.ekdd.gr
3.	Peiraias District	www.nomarhiapeiraia.gr
4.	Ministry of development	www.ypan.gr
5.	Ministry of foreign affairs	www.mfa.gr
Sector 15:	Business services	Web Site
1.	Promaxon S.A	www.procom.gr/
2.	Forever Print Recycling	www.foreverprint.gr/
3.	Euroconsultants S.A	www.euroconsultants.com.gr

This work was previously published in Emerging Markets and E-Commerce in Developing Economies, edited by K. Rouibah; O. Khalil; A. Ella Hassanien, pp. 293-317, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 1.9

A Review of Methodologies for Analyzing Websites

Danielle Booth
Pennsylvania State University, USA

Bernard J. Jansen
Pennsylvania State University, USA

ABSTRACT

This chapter is an overview of the process of Web analytics for Websites. It outlines how visitor information such as number of visitors and visit duration can be collected using log files and page tagging. This information is then combined to create meaningful key performance indicators that are tailored not only to the business goals of the company running the Website but also to the goals and content of the Website. Finally, this chapter presents several analytic tools and explains how to choose the right tool for the needs of the Website. The ultimate goal of this chapter is to provide methods for increasing revenue and customer satisfaction through careful analysis of visitor interaction with a Website.

INTRODUCTION

Web analytics is the measure of visitor behavior on a Website. However, what kind of information is available from Website visitors, and what can be learned from studying such information? By collecting various Web analytics metrics, such as number of visits, visitors, and visit duration, one can develop key performance indicators (KPIs) – a versatile analytic model that measures several metrics against each other to define visitor trends. KPIs use these dynamic numbers to get an in-depth picture of visitor behavior on a site. This information allows businesses to align their Websites' goals with their business goals for the purpose of identifying areas of improvement, promoting popular parts of the site, testing new site functionality, and ultimately increasing revenue. This chapter covers the most common metrics, different methods for gathering metrics, how to utilize key performance indicators, best key practices, and choosing the right Web analytics tool.

The first section addresses metrics, information that can be collected from visitors on a Website. It covers types of metrics based on what kind of data is collected as well as specific metrics and how they can be utilized. The following section discusses

DOI: 10.4018/978-1-59904-974-8.ch008

the two main methods for gathering visitor information -- log files and page tagging. For each method, this section covers the advantages and disadvantages, types of supported information, and examples for data format. Following this is a section on how to choose the key performance indicators (KPIs). This includes outlining several business strategies for integrating Web analytics with the rest of an organization as well as identifying the type of Website and listing several specific KPIs for each site type. The following section provides the overall process and advice for Web analytics integration, and the final section deals with what to look for when choosing analytics tools as well as a comparison of several specific tools. Finally, the conclusion discusses the future of Web analytics.

METRICS

In order to understand the benefits of Website analysis, one must first understand metrics – the different kinds of available user information. Although the metrics may seem basic, once collected, they can be used to analyze Web traffic and improve a Website to better meet its overall goals. According to Panalysis (http://www.panalysis.com/), an Australian Web analytics company, these metrics generally fall into one of four categories: site usage, referrers (or how visitors arrived at your site), site content analysis, and quality assurance. Table 1 shows examples of types of metrics that might be found in these categories.

Although the type and overall number of metrics varies with different analytics vendors, there is still a common set of basic metrics common to most. Table 2 outlines eight widespread types of

Table 1. Metrics categories (Jacka, n.d.)

Site Usage	Referrers	Site Content Analysis	Quality Assurance
• Numbers of visitors and sessions • How many people repeatedly visit the site • Geographic information • Search Engine Activity	• Which websites are sending visitors to your site • The search terms people used to find your site • How many people place bookmarks to the site	• Top entry pages • Most popular pages • Top pages for single page view sessions • Top exit pages • Top paths through the site • Effectiveness of key content	• Broken pages or server errors • Visitor response to errors

Table 2. Eight common metrics of Website analysis

Metric	Description	Category
Visitor Type	Who is accessing the Website (returning, unique, etc.)	Site Usage
Visit Length	The total amount of time a visitor spends on the Website	Site Usage
Demographics and System Statistics	The physical location and information of the system used to access the Website	Site Usage
Internal Search Information	Information on keywords and results pages viewed using a search engine embedded in the Website	Site Usage
Visitor Path	The route a visitor uses to navigate through the Website	Site Content Analysis
Top Pages	The pages that receive the most traffic	Site Content Analysis
Referrering URL and Keyword Analysis	Which sites have directed traffic to the Website and which keywords visitors are using to find the Website	Referrers
Errors	Any errors that occurred while attempting to retrieve the page	Quality Assurance

information that measure who is visiting a Website and what they do during their visits, relating each of these metrics to specific categories.

Each metric is discussed below.

Visitor Type

Since analyzing Website traffic first became popular in the 1990s with the Website counter, the measure of Website traffic has been one of the most closely watched metrics. This metric, however, has evolved from merely counting the number of hits a page receives into counting the number of individuals who visit the Website.

There are two types of visitors: those who have been to the site before, and those who have not. This difference is defined in terms of repeat and new visitors. In order to track visitors in such a way, a system must be able to determine individual users who access a Website; each individual visitor is called a unique visitor. Ideally, a unique visitor is just one visitor, but this is not always the case. It is possible that multiple users access the site from the same computer (perhaps on a shared household computer or a public library). In addition, most analytic software relies on cookies to track unique users. If a user disables cookies in their browser or if they clear their cache, the visitor will be counted as new each time he or she enters the site.

Because of this, some companies have instead begun to track unique visits, or sessions. A session begins once a user enters the site and ends when a user exits the site or after a set amount of time of inactivity (usually 30 minutes). The session data does not rely on cookies and can be measured easily. Since there is less uncertainty with visits, it is considered to be a more concrete and reliable metric than unique visitors. This approach is also more sales-oriented because it considers each visit an opportunity to convert a visitor into a customer instead of looking at overall customer behavior (Belkin, 2006).

Visit Length

Also referred to as Visit Duration or Average Time on Site (ATOS), visit length is the total amount of time a visitor spends on a site during one session. One possible area of confusion when using this metric is handling missing data. This can be caused either by an error in data collection or by a session containing only one page visit or interaction. Since the visit length is calculated by subtracting the time of the visitor's first activity on the site from the time of the visitor's final activity, what happens to the measurement when one of those pieces of data is missing? According to the Web Analytics Association, the visit length in such cases is zero (Burby & Brown, 2007).

When analyzing the visit length, the measurements are often broken down into chunks of time. StatCounter, for example, uses the following time categories:

- Less than 5 seconds
- 5 seconds to 30 seconds
- 30 seconds to 5 minutes
- 5 minutes to 20 minutes
- 20 minutes to 1 hour
- Greater than 1 hour (Jackson, 2007)

The goal of measuring the data in this way is to keep the percentage of visitors who stay on the Website for less than five seconds as low as possible. If visitors stay on a Website for such a short amount of time it usually means they either arrived at the site by accident or the site did not have relevant information. By combining this information with information from referrers and keyword analysis, one can tell which sites are referring well-targeted traffic and which sites are referring poor quality traffic.

Demographics and System Statistics

The demographic metric refers to the physical location of the system used to make a page request.

This information can be useful for a Website that provides region-specific services. For example, if an e-commerce site can only ship its goods to people in Spain, any traffic to the site from outside of Spain is irrelevant. In addition, region-specific Websites also want to make sure they tailor their content to the group they are targeting. Demographic information can also be combined with information on referrers to determine if a referral site is directing traffic to a site from outside a company's regions of service.

System statistics are information about the hardware and software with which visitors access a Website. This can include information such as browser type, screen resolution, and operating system. It is important that a Website be accessible to all of its customers, and by using this information, the Website can be tailored to meet visitors' technical needs.

Internal Search

If a Website includes a site-specific search utility, then it is also possible to measure internal search information. This can include not only keywords but also information about which results pages visitors found useful. The Patricia Seybold Group (http://www.psgroup.com/) identifies the following seven uses for internal search data:

- Identify products and services for which customers are looking, but that are not yet provided by the company.
- Identify products that are offered, but which customers have a hard time finding.
- Identify customer trends.
- Improve personalized messages by using the customers' own words.
- Identify emerging customer service issues
- Determine if customers are provided with enough information to reach their goals.
- Make personalized offers. (Aldrich, 2006)

By analyzing internal search data, one can use the information to improve and personalize the visitors' experience.

Visitor Path

A visitor path is the route a visitor uses to navigate through a Website. Excluding visitors who leave the site as soon as they enter, each visitor creates a path of page views and actions while perusing the site. By studying these paths, one can identify any difficulties a user has viewing a specific area of the site or completing a certain action (such as making a transaction or completing a form).

According to the Web Analytics Association, there are two schools of thought regarding visitor path analysis. The first is that visitor actions are goal-driven and performed in a logical, linear fashion. For example, if a visitor wants to purchase an item, the visitor will first find the item, add it to the cart, and proceed to the checkout to complete the process. Any break in that path (i.e. not completing the order) signifies user confusion and is viewed as a problem.

The second school of thought is that visitor actions are random and illogical and that the only path that can provide accurate data on a visitor's behavior is the path from one page to the page immediately following it. In other words, the only page that influences visitor behavior on a Website is the one they are currently viewing. For example, visitors on a news site may merely peruse the articles with no particular goal in mind. This method of analysis is becoming increasingly popular because companies find it easier to examine path data in context without having to reference the entire site in order to study the visitors' behavior (Web Analytics Association, n. d.).

Top Pages

Panalysis mentions three types of top pages: top entry pages, top exit pages, and most popular pages. Top entry pages are important because the first page a visitor views makes the greatest impression about a Website. By knowing the top entry page, one can make sure that page has relevant information and provides adequate navigation to important parts of the site. Similarly, identifying popular exit pages makes it easier to pinpoint areas of confusion or missing content.

The most popular pages are the areas of a website that receive the most traffic. This metric gives insight into how visitors are utilizing the Website, and which pages are providing the most useful information. This is important because it shows whether the Website's functionality matches up with its business goals; if most of the Website's traffic is being directed away from the main pages of the site, the Website cannot function to its full potential (Jacka, n. d.).

Referrers and Keyword Analysis

A referral page is the page a user visits immediately before entering to a Website, or rather, a site that has directed traffic to the Website. A search engine result page link, a blog entry mentioning the Website, and a personal bookmark are examples of referrers. This metric is important because it can be used to determine advertising effectiveness and search engine popularity. As always, it is important to look at this information in context. If a certain referrer is doing worse than expected, it could be caused by the referring link text or placement. Conversely, an unexpected spike in referrals from a certain page could be either good or bad depending on the content of the referring page.

In the same way, keyword analysis deals specifically with referring search engines and shows which keywords have brought in the most traffic. By analyzing the keywords visitors use to find a page, one is able to determine what visitors expect to gain from the Website and use that information to better tailor the Website to their needs. It is also important to consider the quality of keywords. Keyword quality is directly proportional to revenue and can be determined by comparing keywords with visitor path and visit length (Marshall, n. d.). Good keywords will bring quality traffic and more income to your site.

Errors

Errors are the final metric. Tracking errors has the obvious benefit of being able to identify and fix any errors in the Website, but it is also useful to observe how visitors react to these errors. The fewer visitors who are confused by errors on a Website, the less likely visitors are to exit the site because of an error.

GATHERING INFORMATION

How does one gather these metrics? There are two major methods for collecting data for Web analysis: log files and page tagging. Most current Web analytic companies use a combination of the two methods for collecting data. Therefore, it is important to understand the strengths and weaknesses of each.

Log Files

The first method of metric gathering uses log files. Every Web server keeps a log of page requests that can include (but is not limited to) visitor IP address, date and time of the request, request page, referrer, and information on the visitor's Web browser and operating system. The same basic collected information can be displayed in a variety of ways. Although the format of the log file is ultimately the decision of the company who runs the Web server, the following four formats are a few of the most popular:

- NCSA Common Log
- NCSA Combined Log
- NCSA Separate Log

- W3C Extended Log

The NCSA Common Log format (also known as Access Log format) contains only basic information on the page request. This includes the client IP address, client identifier, visitor username, date and time, HTTP request, status code for the request, and the number of bytes transferred during the request. The Combined Log format contains the same information as the common log with the following three additional fields: the referring URL, the visitor's Web browser and operating system information, and the cookie. The Separate Log format (or 3-Log format) contains the same information as the combined log, but it breaks it into three separate files – the access log, the referral log, and the agent log. The date and time fields in each of the three logs are the same. Table 3 shows examples of the common, combined, and separate log file formats (notice that default values are represented by a dash "-"):

Similarly, W3C provides an outline for standard formatting procedures. This format differs from the first three in that it aims to provide for better control and manipulation of data while still producing a log file readable by most Web analytics tools. The extended format contains user defined fields and identifiers followed by the actual entries, and default values are represented by a dash "-" (Hallam-Baker & Behlendorf, 1999). Table 4 shows an example of an extended log file.

There are several benefits of using system log files to gather data for analysis. The first is that it does not require any changes to the Website or any extra software installation to create the log files. Web servers automatically create these logs and store them on a company's own servers giving the company freedom to change their Web analytics tools and strategies at will. This method also does not require any extra bandwidth when loading a page, and since everything is recorded server-side, it is possible to log both page request successes and failures.

Using log files also has some disadvantages. One major disadvantage is that the collected data is limited to only transactions with the Web server. This means that they cannot log information independent from the servers such as the physical location of the visitor. Similarly, while it is possible to log cookies, the server must be specifically configured to assign cookies to visitors in order to do so. The final disadvantage is that while it is useful to have all the information stored on a company's own servers, the log file method is only available to those who own their Web servers.

Page Tagging

The second method for recording visitor activity is page tagging. Page tagging uses an invisible image to detect when a page has been successfully loaded and then uses JavaScript to send information about the page and the visitor back to a remote server. According to *Web Analytics Demystified* the variables used and amount of data

Table 3. NCSA Log comparison (IBM, 2004)

NCSA Common Log	**125.125.125.125 - dsmith [10/Oct/1999:21:15:05 +0500] "GET /index.html HTTP/1.0" 200 1043**
NCSA Combined Log	125.125.125.125 - dsmith [10/Oct/1999:21:15:05 +0500] "GET /index.html HTTP/1.0" 200 1043 "http://www.ibm.com/" "Mozilla/4.05 [en] (WinNT; I)" "USERID=CustomerA;IMPID=01234"
NCSA Separate Log	Common Log: 125.125.125.125 - dsmith [10/Oct/1999:21:15:05 +0500] "GET /index.html HTTP/1.0" 200 1043 Referral Log: [10/Oct/1999:21:15:05 +0500] "http://www.ibm.com/index.html" Agent Log: [10/Oct/1999:21:15:05 +0500] "Microsoft Internet Explorer - 5.0"

Table 4. W3C extended log file (Microsoft, 2005)

W3C Extended Log	#Software: Microsoft Internet Information Services 6.0 #Version: 1.0 #Date: 2002-05-24 20:18:01 #Fields: date time c-ip cs-username s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status sc-bytes cs-bytes time-taken cs(User-Agent) cs(Referrer) 2002-05-24 20:18:01 172.224.24.114 - 206.73.118.24 80 GET /Default.htm - 200 7930 248 31 Mozilla/4.0+(compatible;+MSIE+5.01;+Windows+2000+Server) http://64.224.24.114/

collected in page tagging are dependent on the Web analytics vendor. Some vendors stress short, easy to use page tags while others emphasize specific tags that require little post-processing. The best thing to look for with this method, however, is flexibility – being able to use all, part, or none of the tag depending on the needs of the page (Peterson, 2004).

There are several benefits to using this method of gathering visitor data. The first is speed of reporting. Unlike a log file, the data received via page tagging is parsed as it comes in. This allows for near real-time reporting. Another benefit is flexibility of data collection. More specifically, it is easier to record additional information about the visitor that does not involve a request to the Web server. Examples of such information include information about a visitor's screen size, the price of purchased goods, and interactions within Flash animations. This is also a useful method of gathering data for companies that do not run their own Web servers or do not have access to the raw log files for their site (such as blogs).

There are also some disadvantages of page tagging, most of which are centered on the extra code that must be added to the Website. This causes it to use more bandwidth each time a page loads, and it also makes it harder to change analytics tools because the code embedded in the Website would have to be changed or deleted entirely. The final disadvantage is that page tagging is only capable of recording page loads, not page failures. If a page fails to load, it means that the tagging code also did not load, and there is therefore no way to retrieve information in that instance.

Although log files and page tagging are two distinct ways to collect information about the visitors to a Website, it is possible to use both together, and many analytics companies provide ways to use both methods to gather data. Even so, it is important to understand the strengths and weaknesses of both. Table 5 shows the advantages and disadvantages of log file analysis and page tagging.

The Problems with Data

One of the most prevalent problems in Web analytics is the difficulty identifying unique users. In order to determine repeat visitors, most Web analytic tools employ cookies that store unique identification information on the visitor's personal computer. Because of problems with users deleting or disabling cookies, however, some companies have moved towards using Macromedia Flash Local Shared Objects (LSOs). LSOs act like a cookie, but standard browsers lack the tools required to delete them, anti-spyware software does not delete them because it does not see them as a threat, and most users do not know how to delete them manually. Awareness is growing, however, and Firefox and Macromedia are working against LSOs and providing users with tools to delete them (Permadi, 2005).

Sen, Dacin, and Pattichis (2006) cite various other problems with log data from Websites including large data size and messy data. Problems with large data size are caused by massive amounts of traffic to a Website and also the amount of information stored in each record. Records with missing

Table 5. Log files vs. page tagging

Log Files		Page Tagging	
Advantages	*Disadvantages*	*Advantages*	*Disadvantages*
Does not require changes to the Website or extra hardware installation	Can only record interactions with the Web server	Near real-time reporting	Requires extra code added to the Website
Does not require extra bandwidth	Server must be configured to assign cookies to visitors	Easier to record additional information	Uses extra bandwidth each time the page loads
Freedom to change tools with a relatively small amount of hassle	Only available to companies who run their own Web servers	Able to capture visitor interactions within Flash animations	Can only record successful page loads, not failures
Logs both page request successes and failures	Cannot log physical location		Hard to switch analytic tools

IP addresses and changes to Website content cause messy data. Even though the data may be hard to work with at first, once it is cleaned up, it provides an excellent tool for Web analytics.

CHOOSING KEY PERFORMANCE INDICATORS

In order to get the most out of Web analytics, one must know how to choose effectively which metrics to analyze and combine them in meaningful ways. This means knowing the Website's business goals and then determining which KPIs will provide the most insight.

Knowing Your Business Goals

Every company has specific business goals. Every part of the company works together to achieve them, and the company Website is no exception. In order for a Website to be beneficial, information gathered from its visitors must not merely show what has happened in the past, but it must also be able to improve the site for future visitors. The company must have clearly defined goals for the future and use this information to support strategies that will help it achieve those goals.

For a Website, the first step in achieving this is making sure the data collected from the site is actionable. According to the Web Analytics Association (McFadden, 2005), in order for a company to collect actionable data, it must meet these three criteria: "(1) the business goals must be clear, (2) technology, analytics, and the business must be aligned, and (3) the feedback loop must be complete" (Web Channel Performance Management section, para. 3).

There are many possible methods for meeting these criteria. One is Alignment-Centric Performance Management (Becher, 2005). This approach goes beyond merely reviewing past customer trends to carefully selecting a few key KPIs based on their future business objectives. Even though a wealth of metrics is available from a Website, this does not mean that all metrics are relevant to a company's needs. Reporting large quantities of data is overwhelming, so it is important to look at metrics in context and use them to create KPIs that focus on outcome and not activity. For example, a customer service Website might view the number of emails responded to on the same day they were sent as a measurement of customer satisfaction. A better way to measure customer satisfaction, however, might be to survey the customers on their experience. Although this measurement is subjective, it is a better representation of customer satisfaction because even if a customer receives a response the same day they send out an email, it does not mean that the experience was a good one (Becher, 2005).

Choosing the most beneficial KPIs using this method is achieved by following "The Four M's of Operational Management" as outlined by Becher (2005) which facilitate effective selection of KPIs:

- **Motivate:** Ensure that goals are relevant to everyone involved.
- **Manage:** Encourage collaboration and involvement for achieving these goals.
- **Monitor:** Once selected, track the KPIs and quickly deal with any problems that may arise.
- **Measure:** Identify the root causes of problems and test any assumptions associated with the strategy.

By carefully choosing a few, quality KPIs to monitor and making sure everyone is involved with the strategy, it becomes easier to align a Website's goals with the company's goals because the information is targeted and stakeholders are actively participating.

Another method for ensuring actionable data is Online Business Performance Management (OBPM) (Sapir, 2004). This approach integrates business tools with Web analytics to help companies make better decisions quickly in an ever-changing online environment where customer data is stored in a variety of different departments. The first step in this strategy is gathering all customer data in a central location and condensing it so that the result is all actionable data stored in the same place. Once this information is in place, the next step is choosing relevant KPIs that are aligned with the company's business strategy and then analyzing expected versus actual results (Sapir 2004).

In order to choose the best KPIs and measure the Website's performance against the goals of a business, there must be effective communication between senior executives and online managers. The two groups should work together to define the relevant performance metrics, the overall goals for the Website, and the performance measurements. This method is similar to Alignment-Centric Performance Management in that it aims to aid integration of the Website with the company's business objectives by involving major stakeholders. The ultimate goals of OBPM are increased confidence, organizational accountability, and efficiency (Sapir 2004).

Identifying KPIs Based on Website Type

Unlike metrics, which are numerical representations of data collected from a Website, KPIs are tied to a business strategy and are usually measured by a ratio of two metrics. By choosing KPIs based on the Website type, a business can save both time and money. Although Websites can have more than one function, each site belongs to at least one of the four main categories – commerce, lead generation, content/media, and support/self service (McFadden, 2005). Table 6 shows common KPIs for each Website type:

We discuss each Website type and related KPIs below.

Commerce

The goal of a commerce Website is to get visitors to purchase goods or services directly from the site, with success gauged by the amount of revenue the site brings in. According to Peterson, "commerce analysis tools should provide the 'who, what, when, where, and how' for your online purchasers (2004, p. 92)." In essence, the important information for a commerce Website is who made (or failed to make) a purchase, what was purchased, when purchases were made, where customers are coming from, and how customers are making their purchases. The most valuable KPIs used to answer these questions are conversion rates, average order value, average visit value, customer loyalty, and bounce rate (McFadden,

Table 6. The four types of Websites and examples of associated KPIs (McFadden, 2005)

Website Type	KPIs
Commerce	• Conversion rates • Average order value • Average visit value • Customer loyalty • Bounce rate
Lead Generation	• Conversion rates • Cost per lead • Bounce rate • Traffic concentration
Content/Media	• Visit depth • Returning visitor ratio • New visitor ratio • Page depth
Support/Self service	• Page depth • Bounce rate • Customer satisfaction • Top internal search phrases

2005). Other metrics to consider with a commerce site are which products, categories, and brands are sold on the site and internal site product search that could signal navigation confusion or a new product niche (Peterson, 2004).

A conversion rate is the number of users who perform a specified action divided by the total of a certain type of visitor (i.e. repeat visitors, unique visitors, etc.) over a given period. Types of conversion rates will vary by the needs of the businesses using them, but two common conversion rates for commerce Websites are the order conversion rate (the percent of total visitors who place an order on a Website) and the checkout conversion rate (the percent of total visitors who begin the checkout process). There are also many methods for choosing the group of visitors on which to base your conversion rate. For example, a business may want to filter visitors by excluding visits from robots and Web crawlers (Ansari, Kohavi, Mason, & Zheng, 2001), or they may want to exclude the traffic that "bounces" from the Website or (a slightly trickier measurement) the traffic that is determined not to have intent to purchase anything from the Website (Kaushik, 2006).

It is common for commerce Websites to have conversion rates around 0.5%, but generally good conversion rates will fall in the 2% range depending on how a business structures its conversion rate (FoundPages, 2007). Again, the ultimate goal is to increase total revenue. According to eVision, for each dollar a company spends on improving this KPI, there is \$10 to \$100 return (2007). The methods a business uses to improve their conversion rate (or rates), however, are different depending on which target action that business chooses to measure.

Average order value (AOV) is a ratio of total order revenue to number of orders over a given period. This number is important because it allows the analyzer to derive a cost for each transaction. There are several ways for a business to use this KPI to its advantage. One way is to break down the AOV by advertising campaigns (i.e. email, keyword, banner ad etc.). This way, a business can see which campaigns are bringing in the best customers and spend more effort refining their strategies in those areas (Peterson, 2005). Overall, however, if the cost of making a transaction is greater than the amount of money customers spend for each transaction, the site is not fulfilling its goal.

There are two main ways to correct this. The first is to increase the number of products customers order per transaction, and the second is to increase the overall cost of purchased products. A good technique for achieving this is through product promotions (McFadden, 2005), but many factors influence how and why customers purchase what they do on a Website. These factors are diverse and can range from displaying a certain security image on the site (MarketingSherpa, 2007) to updating the site's internal search (Young, 2007). Like many KPIs, improvement ultimately comes from ongoing research and a small amount of trial and error.

Another KPI, average visit value, measures the total number of visits to the total revenue. This is a measurement of quality traffic important to businesses. It is problematic for a commerce site when, even though it may have many visitors, each visit generates only a small amount of revenue. In that case, even if the total number of visits increased, it would not have a profound impact on overall profits. This KPI is also useful for evaluating the effectiveness of promotional campaigns. If the average visit value decreases after a specific campaign, it is likely that the advertisement is not attracting quality traffic to the site. Another less common factor in this situation could be broken links or a confusing layout in a site's "shopping cart" area. A business can improve the average visit value by using targeted advertising and employing a layout that reduces customer confusion.

Customer loyalty is the ratio of new to existing customers. Many Web analytics tools measure this using visit frequency and transactions, but there are several important factors in this measurement including the time between visits (Mason, 2007). Customer loyalty can even be measured simply with customer satisfaction surveys (SearchCRM, 2007). Loyal customers will not only increase revenue through purchases but also through referrals, potentially limiting advertising costs (QuestionPro).

Bounce rate is a measurement of how many people arrive at a homepage and leave immediately. There are two scenarios that generally qualify as a bounce. In the first scenario, a visitor views only one page on the Website. In the second scenario, a visitor navigates to a Website but only stays on the site for five seconds or less (Avinash, 2007). This could be due to several factors, but in general, visitors who bounce from a Website are not interested in the content. Like average order value, this KPI helps show how much quality traffic a Website receives. A high bounce rate may be a reflection of unintuitive site design or misdirected advertising.

Lead Generation

The goal for a lead generation Website is to obtain user contact information in order to inform them of a company's new products and developments and to gather data for market research; these sites primarily focus on products or services that cannot be purchased directly online. Examples of lead generation include requesting more information by mail or email, applying online, signing up for a newsletter, registering to download product information, and gathering referrals for a partner site (Burby, 2004). The most important KPIs for lead generation sites are conversion rates, cost per lead, bounce rate, and traffic concentration (McFadden, 2005).

Similar to commerce Website KPIs, a conversion rate is the ratio of total visitors to the amount of visitors who perform a specific action. In the case of lead generation Websites, the most common conversion rate is the ratio of total visitors to leads generated. The same visitor filtering techniques mentioned in the previous section can be applied to this measurement (i.e. filtering out robots and Web crawlers and excluding traffic that bounces from the site). This KPI is an essential tool in analyzing marketing strategies. Average lead generation sites have conversion rates ranging from 5-6% and 17-19% conversion rates for exceptionally good sites (Greenfield, 2006). If the conversion rate of a site increases after the implementation of a new marketing strategy, it indicates that the campaign was successful. If it decreases, it indicates that the campaign was not effective and might need to be reworked.

Cost per lead (CPL) refers to the ratio of total expenses to total number of leads, or how much it costs a company to generate a lead; a more targeted measurement of this KPI would be the ratio of total marketing expenses to total number of leads. Like the conversion rate, CPL helps a business gain insight into the effectiveness of its marketing campaigns. A good way to measure the success of this KPI is to make sure that the CPL for a specific marketing campaign is less than the overall CPL (WebSideStory, 2004). Ideally, the CPL should be low, and well-targeted advertising is usually the best way to achieve this.

Lead generation bounce rate is the same measurement as the bounce rate for commerce sites. This KPI is a measurement of visitor retention based off total number of bounces to total number of visitors; a bounce is a visit characterized by a visitor entering the site and immediately leaving. Lead generation sites differ from commerce sites in that they may not require the same level of user interaction. For example, a lead generation site could have a single page where users enter their contact information. Even though they only view one page, the visit is still successful if the Website is able to collect the user's information. In these situations, it is best to base the bounce rate solely off of time spent on the site. As with commerce sites, the best way to decrease a site's bounce rate is to increase advertising effectiveness and decrease visitor confusion.

The final KPI is traffic concentration, or the ratio of the number of visitors to a certain area in a Website to total visitors. This KPI shows which areas of a site have the most visitor interest. For this type of Website, it is ideal to have a high traffic concentration on the page or pages where users enter their contact information.

Content/Media

Content/media Websites focus mainly on advertising, and the main goal of these sites is to increase revenue by keeping visitors on the Website longer and also to keep visitors coming back to the site. In order for these types of sites to succeed, site content must be engaging and frequently updated. If content is only part of a company's Website, the content used in conjunction with other types of pages can be used to draw in visitors and provide a way to immerse them with the site. The main KPIs are visit depth, returning visitors, new visitor percentage, and page depth (McFadden, 2005).

Visit depth (also referred to as depth of visit or path length) is the measurement of the ratio between page views and unique visitors, or how many pages a visitor accesses each visit. As a general rule, visitors with a higher visit depth are interacting more with the Website. If visitors are only viewing a few pages per visit, it means that they are not engaged, and the effectiveness of the site is low. A way to increase a low average visit depth is by creating more targeted content that would be more interesting to the Website's target audience. Another strategy could be increasing the site's interactivity to encourage the users to become more involved with the site and keep them coming back.

Unlike the metric of simply counting the number of returning visitors on a site, the returning visitor KPI is the ratio of unique visitors to total visits. A factor in customer loyalty, this KPI measures the effectiveness of a Website to bring visitors back. A lower ratio for this KPI is best because a lower number means more repeat visitors and more visitors who are interested in and trust the content of the Website. If this KPI is too low, however, it might signal problems in other areas such as a high bounce rate or even click fraud. Click fraud occurs when a person or script is used to generate visits to a Website without having genuine interest in the site. According to a study by Blizzard Internet Marketing, the average for returning visitors to a Website is 23.7% (White, 2006). As with many of the other KPIs for content/media Websites, the best way to improve the returning visitor rate is by having quality content and encouraging interaction with the Website.

New visitor ratio is the measurement of new visitors to unique visitors and is used to determine if a site is attracting new people. When measuring this KPI, the age of the Website plays a role – newer sites will want to attract new people. Similarly, another factor to consider is if the Website is concerned more about customer retention or gaining new customers. As a rule, however, the new visitor ratio should decrease over time as the returning visitor ratio increases. New visitors can be brought to the Website in a variety of different ways, so a good way to increase this KPI is to try different marketing strategies and figure out which campaigns bring the most (and the best) traffic to the site.

The final KPI for content/media sites is page depth. This is the ratio of page views for a specific page and the number of unique visitors to that page. This KPI is similar to visit depth, but its measurements focus more on page popularity. Average page depth can be used to measure interest in specific areas of a Website over time and to make sure that the interests of the visitors match the goals of the Website. If one particular page on a Website has a high page depth, it is an indication that that page is of particular interest to visitors. An example of a page in a Website expected to have a higher page depth would be a news page. Information on a news page is constantly updated so that, while the page is still always in the same location, the content of that page is constantly changing. If a Website has high page depth in a relatively unimportant part of the site, it may signal visitor confusion with navigation in the site or an incorrectly targeted advertising campaign.

Support/Self Service

Websites offering support or self-service are interested in helping users find specialized answers for specific problems. The goals for this type of Website are increasing customer satisfaction and decreasing call center costs; it is more cost-effective for a company to have visitors find information through its Website than it is to operate a call center. The KPIs of interest are visit length, content depth, and bounce rate. In addition, other areas to examine are customer satisfaction metrics and top internal search phrases (McFadden, 2005).

Page depth for support/self service sites is the same measurement as page depth content/media sites – the ratio of page views to unique visitors. With support/self service sites, however, high page depth is not always a good sign. For example, a visitor viewing the same page multiple times may show that the visitor is having trouble finding helpful information on the Website or even that the information the visitor is looking for does not exist on the site. The goal of these types of sites is to help customers find what they need as quickly as possible and with the least amount of navigation through the site (CCMedia, 2007). The best way to keep page depth low is to keep visitor confusion low.

As with the bounce rate of other Website types, the bounce rate for support/self service sites reflects ease of use, advertising effectiveness, and visitor interest. A low bounce rate means that quality traffic is coming to the Website and deciding that the site's information is potentially useful. Poor advertisement campaigns and poor Website layout will increase a site's bounce rate.

Customer satisfaction deals with how the users rate their experience on a site and is usually collected directly from the visitors (not from log files), either through online surveys or through satisfaction ratings. Although it is not a KPI in the traditional sense, gathering data directly from visitors to a Website is a valuable tool for figuring out exactly what visitors want. Customer satisfaction measurements can deal with customer ratings, concern reports, corrective actions, response time, and product delivery. Using these numbers, one can compare the online experience of the Website's customers to the industry average and make improvements according to visitors' expressed needs.

Similarly, top internal search phrases applies only to sites with internal search, but it can be used to measure what information customers are most interested in which can lead to improvement in site navigation. This information can be used to direct support resources to the areas generating the most user interest, as well as identify which parts of the Website users may have trouble accessing. In addition, if many visitors are searching for a product not supported on the Website, it could be a sign of ineffective marketing.

Regardless of Website type, the KPIs listed above are not the only KPIs that can prove useful in analyzing a site's traffic, but they provide a good starting point. The main thing to remember is that no matter what KPIs a company chooses, they must be aligned with its business goals, and more KPIs do not necessarily mean better analysis – quality is more important than quantity.

KEY BEST PRACTICES

In this chapter, we have addressed which metrics can be gathered from a Website, how to gather them, and how to determine which information is important. But how can this help improve a business? To answer this, the Web Analytics Association provides nine key best practices to follow when analyzing a Website (McFadden, 2005). Figure 1 outlines this process.

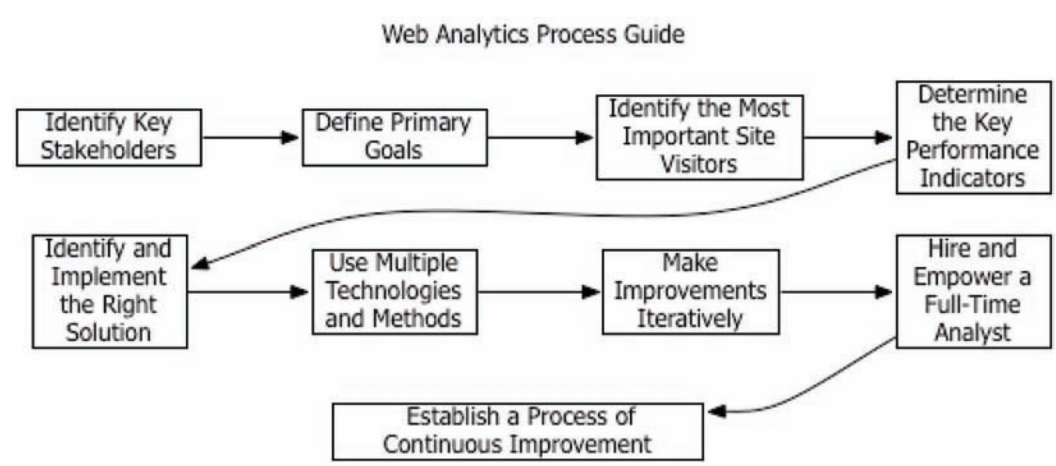
Identify Key Stakeholders

A stakeholder is anyone who holds an interest in a Website. This includes management, site developers, visitors, and anyone else who creates, maintains, uses, or is affected by the site. In order for the Website to be truly beneficial, it must integrate input from all major stakeholders. Involving people from different parts of the company also makes it more likely that they will embrace the Website as a valuable tool.

Define Primary Goals for Your Website

To know the primary goals of a Website, one must first understand the primary goals of its key

Figure 1. The best key practices of Web analytics



stakeholders. This could include such goals as increasing revenue, cutting expenses, and increasing customer loyalty (McFadden, 2005). Once those goals have been defined, discuss each goal and prioritize them in terms of how the Website can most benefit the company. As always, beware of political conflict between stakeholders and their individual goals as well as assumptions they may have made while determining their goals that may not necessarily be true. By going through this process, a company can make sure that goals do not conflict and that stakeholders are kept happy.

Identify the Most Important Site Visitors

According to Sterne, corporate executives categorize their visitors differently in terms of importance. Most companies classify their most important visitors as ones who either visit the site regularly, stay the longest on the site, view the most pages, purchase the most goods or services, purchase goods most frequently, or spend the most money (Sterne, n. d.). There are three types of customers – (1) customers a company wants to keep who have a high current value and high future potential, (2) customers a company wants to grow who can either have a high current value and low future potential or low current value and high future potential, and (3) customers a company wants to eliminate who have a low current value and low future potential. The most important visitor to a Website, however, is the one who ultimately brings in the most revenue. Defining the different levels of customers will allow one to consider the goals of these visitors. What improvements can be made to the Website in order to improve their browsing experiences?

Determine the Key Performance Indicators

The next step is picking the metrics that will be most beneficial in improving the site and eliminating the ones that will provide little or no insight into its goals. One can then use these metrics to determine which KPI you wish to monitor. As mentioned in the previous section, the Website type – commerce, lead generation, media/content, or support/self service – plays a key role in which KPIs are most effective for analyzing site traffic.

Identify and Implement the Right Solution

This step deals with finding the right Web analytics technology to meet the business's specific needs. After the KPIs have been defined, this step should be easy. The most important things to consider are the budget, software flexibility and ease of use, and how well the technology will work with the needed metrics. McFadden suggests that it is also a good idea to run a pilot test of the top two vendor choices (McFadden, 2005). We will expand on this topic further in the next section.

Use Multiple Technologies and Methods

Web analytics is not the only method available for improving a Website. To achieve a more holistic view of a site's visitors, one can also use tools such as focus groups, online surveys, usability studies, and customer services contact analysis (McFadden, 2005).

Make Improvements Iteratively

When analyzing a Website's data, it is helpful to add gradual improvements to the Website instead of updating too many facets of the Website at once. By doing this one can monitor if a singular change is an improvement or if it is actually hurting the site.

Hire and Empower a Full-Time Analyst

It is important to put a person in charge of the data once it is collected. According to the Web Analytics Association, a good analyst understands business needs (which means communicating well with the stakeholders), has knowledge of technology and marketing, has respect, credibility, and authority, and is already a company employee. Although it may seem like hiring a full-time analyst is expensive, many experts agree that the return on revenue should be more than enough compensation to recoup the cost (McFadden, 2005).

Establish a Process of Continuous Improvement

Once the Web analysis process is decided upon, continuous evaluation is paramount. This means reviewing the goals and metrics and monitoring new changes and features which are added to the Website. It is important that the improvements are adding value to the site and meeting expectations.

SPECIFIC TOOLS

Choosing a Tool

Once the company decides what it wants out of the Web analysis, it is time to find the right tool. Kaushik outlines ten important questions to ask Web analytics vendors (2007):

1. *What is the difference between your tool and free Web analytics tools?* Since the company who owns the Website will be paying money for a service, it is important to know why that service is better than free services (for example, Google Analytics). Look for an answer that outlines the features and functionality of the vendor. Do not look for answers about increased costs because of privacy threats or poor support offered by free analytics tools.
2. *Do you offer a software version of your tool?* Generally, a business will want to look for a tool that is software based and can run on their own servers. If a tool does not have a software version but plans to make one in the future, it shows insight into how prepared they are to offer future products if there is interest.
3. *What methods do you use to capture data?* If you remember from the first section, there are two main ways to capture visitor data from a Website – log files and page tagging. Ideally, one should look for a vendor that offers both, but what they have used in the past is also important. Because technology is constantly changing, look for a company that has kept up with these changes in the past by providing creative solutions.
4. *Can you help me calculate the total cost of ownership for your tool?* The total cost of ownership for a Web analytics tool depends on the specific company, the systems they have in place, and the pricing of the prospective Web analytics tool. In order to make this calculation, one must consider the following:
    a. Cost per page view.
    b. Incremental costs (i.e. charges for overuse or advanced features).
    c. Annual support costs after the first year.
    d. Cost of professional services (i.e. installation, troubleshooting, or customization).
    e. Cost of additional hardware you may need.
    f. Administration costs (which includes the cost of an analyst and any additional employees you may need to hire).
5. *What kind of support do you offer?* Many vendors advertise free support, but it is

important to be aware of any limits that could incur additional costs. It is also important to note how extensive their support is and how willing they are to help.

6. *What features do you provide that will allow me to segment my data?* Segmentation allows companies to manipulate their data. Look for the vendor's ability to segment your data after it is recorded. Many vendors use JavaScript tags on each page to segment the data as it is captured, meaning that the company has to know exactly what it wants from the data before having the data itself; this approach is less flexible.
7. *What options do I have to export data into our system?* It is important to know who ultimately owns and stores the data and whether it is possible to obtain both raw and processed data. Most vendors will not provide companies with the data exactly as they need it, but it is a good idea to realize what kind of data is available before a final decision is made.
8. *Which features do you provide for integrating data from other sources into your tool?* This question deals with the previous section's Key Best Practice #6: Use Multiple Technologies and Methods. If a company has other data it wants to bring to the tool (such as survey data or data from your ad agency), bring them up to the potential analytics vendor and see if it is possible to integrate this information into their tool.
9. *What new features are you developing that would keep you ahead of your competition?* Not only will the answer to this question tell how much the vendor has thought about future functionality, it will also show how much they know about their competitors.
10. *Why did you lose your last two clients? Who are they using now?* The benefits of this question are obvious -- by knowing how they lost prior business, the business can be confident that it has made the right choice.

Some examples of free and commercially available analytics tools are discussed below.

Free Tools

One of the most popular free analytics tools on the Web now is Google Analytics (previously Urchin). Google Analytics (http://www.google.com/analytics/) uses page tagging to collect information from visitors to a site. In addition to expanding on the already highly regarded Urchin analytics tool, it also provides support for integrating other analytic information (for example, WordPress and AdWords). Google Analytics reports many of the KPIs discussed in the previous sections including depth of visit, returning visitors, and page depth.

There is, however, concern about privacy issues regarding Google Analytics because Google uses their default privacy policy for their analytics tools, but the company assures its Google Analytics users that only account owners and people to whom the owners give permission will have access to the data (Dodoo, 2006). Microsoft also provides a free Web analytic software called Gatineau (Thomas, 2007).

Paid Tools

InfoWorld provides an in-depth analysis comparing the top four Web analytic companies – Coremetrics, WebTrends, Omniture, and WebSideStory HBX (Heck, 2005). They created a scoring chart and measured each vendor on reporting, administration, performance, ease-of-use, support, and value. Coremetrics received a score of 8.3 with its highest ratings in administration and support. It is a hosted service that offers special configurations for financial, retail, and travel services. WebTrends also earned a score of 8.3 with its highest rating in reporting. This tool is expensive, but it offers a wide range of performance statistics and both client and server hosting. Omniture is next in line with a score of 8.4 with its highest ratings in reporting and support. It is an ASP reporting application that excels in providing relevant reports. WebSideStory had the highest score of 8.7 with its highest ratings in reporting, administration, ease-of-use, and support. This tool is easy to use and is appropriate for many different types of businesses.

CONCLUSION

The first step in analyzing your Website and Website visitors is understanding and analyzing your business goals and then using that information to carefully choose your metrics. In order to take full advantage of the information gathered from your site's visitors, you must consider alternative methods such as focus groups and online surveys, make site improvements gradually, hire a full-time analyst, and realize that your site's improvement is a process and not a one-time activity. Using these key best practices and choosing the right analytics vendor to fit your business will save your company money and ultimately increase revenue.

As Web analytics continues to mature, the methods vendors use to collect information are becoming more refined. One article speculates that companies will find concrete answers to the problems with cookies and unique visitors (Eisenberg, 2005). The Web analytics industry as a whole is also expanding. According to Eisenberg (2005), a recent Jupiter report predicts an increase in the Web analytics industry – 20 percent annually. More and more businesses are realizing the benefits of critically analyzing their Website traffic and are taking measures to improve their profits based off these numbers. Regardless of business size and objective, an effective Web analytics strategy is becoming increasingly essential for online success.

REFERENCES

Aldrich, S. E. (2006, May 2). *The Other Search: Making the Most of Site Search to Optimize the Total Customer Experience*. Patricia Seybold Group. Retrieved March 7, 2007, from WebSideStory database.

Ansari, S., Kohavi, R., Mason, L., & Zheng, Z. (2001). Integrating E-Commerce and Data Mining: Architecture and Challenges. *IEEE International Conference on Data Mining*.

Avinash, A. (2007, June 26). *Bounce Rate: Sexiest Web Metric Ever?* Retrieved December 2, 2007, from http://www.mpdailyfix.com/2007/06/bounce_rate_sexiest_web_metric.html.

Becher, J. D. (2005, March). Why Metrics-Centric Performance Management Solutions Fall Short. *DM Review Magazine*. Retrieved March 7, 2007, from http://www.dmreview.com/article_sub.cfm?articleId=1021509.

Belkin, M. (2006, April 8). *15 Reasons why all Unique Visitors are not created equal*. Retrieved March 7, 2007, from http://www.omniture.com/blog/node/16.

Burby, J. (2004, July 20). *Build a Solid Foundation With Key Performance Indicators, Part 1: Lead-Generation Sites*. Retrieved March 7, 2007, from http://www.clickz.com/showPage.html?page=3382981.

Burby, J., & Brown, A. (2007). *Web Analytics Definitions*. Retrieved October 30, 2007, from http://www.webanalyticsassociation.org/attachments/committees/5/WAA-Standards-Analytics-Definitions-Volume-I-20070816.pdf.

CCMedia. (2007, August 30). *How to Obtain a Cost-effective Operational Model for Support/Self-service Websites?* Retrieved December 5, 2007, from www.webnibbler.com/en/WhitePaper/Online%20Support%20Website.pdf.

Dodoo, M. (2006, March 3). *Privacy & Google Analytics*. Retrieved March 7, 2007, from http://www.marteydodoo.com/2006/05/03/privacy-google-analytics/.

Eisenberg, B. (2005, April 1). *Web Analytics: Exciting Times Ahead*. Retrieved March 7, 2007, from http://www.clickz.com/showPage.html?page=3493976

eVision. (2007, September 27). *Websites that convert visitors into customers: Improving the ability of your Website to convert visitors into inquiries, leads, and new business*. Retrieved March 7, 2007, from http://www.evisionsem.com/marketing/webanalytics.htm

FoundPages. (2007, October 25). *Increasing Conversion Rates*. Retrieved October 31, 2007, from http://www.foundpages.com/calgary-internet-marketing/search-conversion.html

Greenfield, M. (2006, January 1). *Use Web Analytics to Improve Profits for New Year: Focus on four key statistics*. Retrieved March 7, 2007, from http://www.practicalecommerce.com/articles/132/Use-Web-Analytics-to-Improve-Profits-for-New-Year/

Hallam-Baker, P. M., & Behlendorf, B. (1999, February 4). *Extended Log File Format*. Retrieved March 7, 2007, from http://www.w3.org/TR/WD-logfile.html

Heck, M. (2005, February 18). *Chart Your Website's Success*. Retrieved March 7, 2007, from http://www.infoworld.com/Omniture_SiteCatalyst_11/product_56297.html?view=1&curNodeId=0&index=0

IBM. (2004, May 19). *Log File Formats*. Retrieved October 29, 2007, from http://publib.boulder.ibm.com/tividd/td/ITWSA/ITWSA_info45/en_US/HTML/guide/c-logs.html

Jacka, R. *Getting Results From Your Website*. Retrieved October 30, 2007, from http://www.panalysis.com/downloads/gettingresults.pdf

Jackson, M. (2007, January 22). *Analytics: Deciphering the Data*. Retrieved March 7, 2007, from http://www.ecommerce-guide.com/resources/article.php/3655251

Kaushik, A. (2006, November 13). *Excellent Analytics Tip #8: Measure the Real Conversion Rate & 'Opportunity Pie.'* Retrieved November 3, 2007, from http://www.kaushik.net/avinash/2006/11/excellent-analytics-tip-8-measure-the-real-conversion-rate-opportunity-pie.html

Kaushik, A. (2007, January 23). *Web Analytics Tool Selection: 10 Questions to ask Vendors*. Retrieved March 7, 2007, from http://www.kaushik.net/avinash/2007/01/web-analytics-tool-selection-10-questions-to-ask-vendors.html

MarketingSherpa. (2007, October 20). *Security Logo in Email Lifts Average Order Value 28.3%*. Retrieved December 4, 2007, from https://www.marketingsherpa.com/barrier.html?ident=30183

Marshall, J. *Seven Deadly Web Analytics Sins*. Retrieved March 7, 2007, from http://www.clicktracks.com/insidetrack/articles/7_deadly_webanalytics_sins01.php

Mason, N. (2007, February 6). *Customer Loyalty Improves Retention*. Retrieved March 7, 2007, from http://www.clickz.com/showPage.html?page=3624868

McFadden, C. (2005, July 6). *Optimizing the Online Business Channel with Web Analytics*. Retrieved March 7, 2007, from http://www.Webanalyticsassociation.org/en/art/?9

Microsoft. (2005, August 22). *W3C Extended Log File Examples*. Retrieved March 7, 2007, from http://technet2.microsoft.com/WindowsServer/en/library/b5b8a519-8f9b-456b-9040-018358f2c0c01033.mspx?mfr=true

Permadi, F. (2005, June 19). *Introduction to Flash Local Shared-Object*. Retrieved March 7, 2007, from http://www.permadi.com/tutorial/flashSharedObject/index.html

Peterson, E. T. (2004). *Web Analytics Demystified*. Celilo Group Media.

Peterson, E. T. (2005, July 31). *Average Order Value*. Retrieved November 3, 2007, from Web Analytics Demystified Blog Website: http://blog.webanalyticsdemystified.com/weblog/2005/07/average-order-value.html

QuestionPro. *Measuring Customer Loyalty and Customer Satisfaction*. Retrieved November 21, 2007, from http://www.questionpro.com/akira/showArticle.do?articleID=customerloyalty.

Sapir, D. (2004, August). Online Analytics and Business Performance Management. *BI Report*. Retrieved March 7, 2007, from http://www.dmreview.com/editorial/dmreview/print_action.cfm?articleId=1008820

Search, C. R. M. (2007, May 9). *Measuring Customer Loyalty*. Retrieved November 4, 2007, from http://searchcrm.techtarget.com/general/0,295582,sid11_gci1253794,00.html

Sen, A., Dacin, P. A., & Pattichis, C. (2006, November). Current trends in Web data analysis. *Communications of the ACM*, *49*(11), 85–91. doi:10.1145/1167838.1167842

Sterne, J. *10 Steps to Measuring Website Success*. Retrieved March 7, 2007, from http://www.marketingprofs.com/login/join.asp?adref=rdblk&source=/4/sterne13.asp

Thomas, I. (2007, January 9). *The rumors are true: Microsoft 'Gatineau' exists*. Retrieved March 7, 2007, from http://www.liesdamnedlies.com/2007/01/the_rumors_are_.html

Web Analytics Association. *Onsite Behavior - Path Analysis*. Retrieved March 7, 2007, from http://www.Webanalyticsassociation.org/attachments/contentmanagers/336/1%20Path%20AnAnalys.doc

WebSideStory. (2004). *Use of Key Performance Indicators in Web Analytics*. Retrieved December 2, 2007, from www.4everywhere.com/documents/KPI.pdf

White, K. (2006, May 10). *Unique vs. Returning Visitors Analyzed*. Retrieved March 7, 2007, from http://newsletter.blizzardinternet.com/unique-vs-returning-visitors-analyzed/2006/05/10/#more-532

Young, D. (2007, August 15). Site Search: Increases Conversion Rates, Average Order Value And Loyalty. *Practical Ecommerce*, Retrieved November 15, 2007, from http://www.practicalecommerce.com/articles/541/Site-Search-Increases-Conversion-Rates-Average-Order-Value-And-Loyalty/

KEY TERMS

Abandonment Rate: KPI that measures the percentage of visitors who got to that point on the site but decided not to perform the target action.

Alignment-Centric Performance Management: Method of defining a site's business goals by choosing only a few key performance indicators.

Average Order Value: KPI that measures the total revenue to the total number of orders.

Average Time on Site (ATOS): See *visit length*.

Checkout Conversion Rate: KPI that measures the percent of total visitors who begin the checkout process.

Commerce Website: A type of Website where the goal is to get visitors to purchase goods or services directly from the site.

Committed Visitor Index: KPI that measures the percentage of visitors that view more than one page or spend more than 1 minute on a site (these measurements should be adjusted according to site type).

Content/Media Website: A type of Website focused on advertising.

Conversion Rate: KPI that measures the percentage of total visitors to a Website that perform a specific action.

Cost Per Lead (CPL): KPI that measures the ratio of marketing expenses to total leads and shows how much it costs a company to generate a lead.

Customer Satisfaction Metrics: KPI that measures how the users rate their experience on a site.

Customer Loyalty: KPI that measures the ratio of new to existing customers.

Demographics and System Statistics: A metric that measures the physical location and information of the system used to access the Website.

Depth of Visit: KPI that measures the ratio between page views and visitors.

Internal Search: A metric that measures information on keywords and results pages viewed using a search engine embedded in the Website.

Key Performance Indicator (KPI): A combination of metrics tied to a business strategy.

Lead Generation Website: A type of Website that is used to obtain user contact information in order to inform them of a company's new products and developments, and to gather data for market research.

Log File: Log kept by a Web server of information about requests made to the Website including (but not limited to) visitor IP address, date and time of the request, request page, referrer, and information on the visitor's Web browser and operating system.

Log File Analysis: Method of gathering metrics that uses information gathered from a log file to gather Website statistics.

Metrics: Statistical data collected from a Website such as number of unique visitors, most popular pages, etc.

New Visitor: A user who is accessing a Website for the first time.

New Visitor Percentage: KPI that measures the ratio of new visitors to unique visitors.

Online Business Performance Management (OBPM): Method of defining a site's business goals that emphasizes the integration of business tools and Web analytics to make better decisions quickly in an ever-changing online environment.

Order Conversion Rate: KPI that measures the percent of total visitors who place an order on a Website.

Page Depth: KPI that measures the ratio of page views for a specific page and the number of unique visitors to that page.

Page Tagging: Method of gathering metrics that uses an invisible image to detect when a page has been successfully loaded and then uses JavaScript to send information about the page and the visitor back to a remote server.

Prospect Rate: KPI that measures the percentage of visitors who get to the point in a site where they can perform the target action (even if they do not actually complete it).

Referrers and Keyword Analysis: A metric that measures which sites have directed traffic to the Website and which keywords visitors are using to find the Website.

Repeat Visitor: A user who has been to a Website before and is now returning.

Returning Visitor: KPI that measures the ratio of unique visitors to total visits.

Search Engine Referrals: KPI that measures the ratio of referrals to a site from specific search engines compared to the industry average.

Single Access Ratio: KPI that measures the ratio of total single access pages (or pages where the visitor enters the site and exits immediately from the same page) to total entry pages.

Stickiness: KPI that measures how many people arrive at a homepage and proceed to traverse the rest of the site.

Support/Self Service Website: A type of Website that focuses on helping users find specialized answers for their particular problems.

Top Pages: A metric that measures the pages in a Website that receive the most traffic.

Total Bounce Rate: KPI that measures the percentage of visitors who scan the site and then leave.

Traffic Concentration: KPI that measures the ratio of number of visitors to a certain area in a Website to total visitors.

Unique Visit: One visit to a Website (regardless of if the user has previously visited the site); an alternative to unique visitors.

Unique Visitor: A specific user who accesses a Website.

Visit Length: A metric that measures total amount of time a visitor spends on the Website.

Visit Value: KPI that measures the total number of visits to total revenue.

Visitor Path: A metric that measures the route a visitor uses to navigate through the Website.

Visitor Type: A metric that measures users who access a Website. Each user who visits the Website is a unique user. If it is a user's first time to the Website, that visitor is a new visitor, and if it is not the user's first time, that visitor is a repeat visitor.

Web Analytics: The measurement of visitor behavior on a Website.

This work was previously published in Handbook of Research on Web Log Analysis, edited by B. Jansen; A. Spink; I. Taksa, pp. 143-164, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 1.10
Basics to Develop Web Services for Human Resources

Roman Povalej
University of Karlsruhe (TH), Germany

Peter Weiß
University of Karlsruhe (TH), Germany

INTRODUCTION

In this article certain pillars as basics are presented being necessary to develop Web services (W3C, 2007) supporting human resource (HR) processes like assessing, hiring, modeling information systems, staffing, and so forth; by the help of these Web services. Current HR information systems in general do not adequately support tasks related to cross-organizational or global skills and competence management. In the following, the topic is presented which relates to knowledge management especially to "communities of practice," as well as related topics such as e-skills and ICT (information and communication technologies) professionalism; the latter currently being broadly discussed by experts in Europe.

DOI: 10.4018/978-1-59904-883-3.ch013

HR managers of a company or an organization are challenged through the need to formalize skills requirements and to continuously monitor the skills demand inside the company. Obtaining ICT skills are not a one-time event. Technological change advances at a high speed and requires that skills need continually to be kept up-to-date and relevant (The European e-Skills Forum [ESF], 2005).

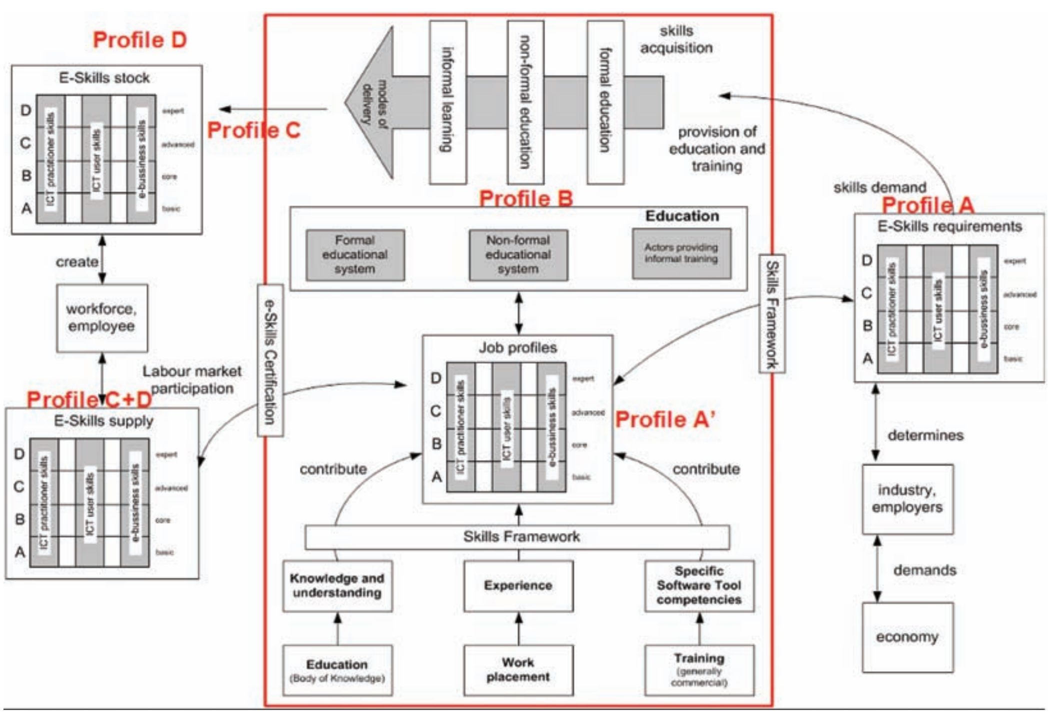
During the last years, new concepts have emerged which intend to empower learners and individuals to steer learning processes to a large extent on their own. Learning objectives tend to be increasingly individual in character (ESF, 2005).

In this context, providing an appropriate infrastructure which supports the continuing professional development (CPD) of employees is today a key issue. CPD processes require a respective infrastructure encompasses besides qualifications, skills/competence frameworks and body of knowledge, as well required standards for competence, skills, and appropriate career and development services. Standards encompass educational and industry-oriented performance standards which in turn are expressed preferably through a common language as competence and skills standards. The governance and administration of the CPD process require the availability of flexible and personalized certification services which offer the formal validation of individuals' learning achievements independent of where and how they were acquired.

Recruitment and internal mobility are typical use cases to identify the right people wanted, by matching their assessment outcomes with job positions requested. In case a company needs new ICT competence, the HR manager in charge has to draw up a job advertisement, which describes the required knowledge, skills, competences (KSC), and qualifications for the specific job role. The information is generally structured in a job profile, which mirrors ideally the companies core competences; but how can this be efficiently done or supported by a concept method and/or tool, how can be potential employees be addressed independently of their national origin? Further scenarios can be derived from Figure 1.

Noticeably, information about requirements is exchanged amongst various players along the value chain. Hence, looking at ICT skills requirements necessitates looking at the belonging processes as a whole. This relates to a common language which allows exchanging information amongst the various players involved and exchange of data stored in respective HR-systems. The latter, in particular, motivates data services which facilitate sharing of data required for the execution of tasks along the HR processes (such as recruitment of ICT professionals). Typically, systems do not have the required interoperability because the applied data models differ in structure and semantics. In the following, we look at how to achieve better interoperability of systems and

Figure 1. Supply and demand of e-skills



an enhanced sharing of information about ICT skills requirements typically stored in job or ICT worker profiles.

Competence-based professional profiles are a referred concept to define and exchange information about respective professional standards. The profile is the description of competence required to operate on a process, on a service, or for a definite role, inside a direction or a function, a working team, or a project. It shows as much explicit as possible which type of know-how needs an organization has in terms of competence. In a most diffused configuration, the competence based profile is defined with reference to an organizational position or role. The competence based profile of a role has to be built up without any reference to single persons that occupied this role or are candidates to cover the role in the future.

In the following, the article looks at a common language and ICT worker profiles as a constituent part and central element of an adequate management of HR processes in general and CPD processes specifically. The article presents a conceptual model which allows expanding existing barriers of HR systems. Web services are seen as an appropriate concept to implement required functionality and processes.

BACKGROUND

The competitiveness of European industry depends on both the effective use of ICT for industrial and business processes and the KSC and qualifications of existing and new employees (ESF, 2005).

Today, nearly every area of economic activity is affected by ICT and the pace of technological change and short technology life cycles makes these the most dynamic of occupations (Cedefop, 2005); scilicet companies as well as ICT employees (and also others) have an ongoing need to keep their KSC and qualifications up-to-date to prevail against competitors. In respect thereof, ICT certifications are useful to validate ICT employees' market value and often they rank a key role in the hiring process.

In the first place, ICT companies, but actually all companies, employ people with ICT skills as well as people with a range of other relevant skills. Often the number of ICT practitioners in such ICT user organizations (e.g., banks, manufacturing companies, and airlines) is greater than the number employed in ICT companies.

New competency-based systems of certification of individual employees and/or others were, are, and further on, will be established within the information technology field and offered on the market; this is mainly driven by ICT vendors, the expected profit is highly valued.

Competence encompasses "[…] the demonstration of relevant, up-to-date skills and capabilities appropriate to a particular task or role with practical experience to complement theoretical knowledge" (Hughes & Thompson, 2007). This includes as well, necessary soft skills such as interpersonal and communication skills as well as an understanding of the business domain. We argue that performance components are an adequate means to develop necessary skills and to avoid that qualification measures coincide. Figure 1 shows the supply and demand of e-skills.

E-skills present a mainly political driven joint initiative of stakeholders in Europe such as the European Commission, ICT professionals associations, and international ICT vendor industry associations. Subsequently, we argue why e-skills require joint activities and concerted actions.

Each stakeholder determines their own needed and relevant e-skills requirements (profile A). Education institutions determine their curricula (profile B) which consist of e-skills which should be taught to the students. Industry and employers determine their e-skills requirements (profile A) to manage their tasks and challenges in the economy. Sometimes the requirements of these two groups do not exactly match, for example, parts of the curricula are in general concepts, methods, and tools about databases; however, the industry demands product-related knowledge about a specific database, often they do not notice that the parts of the curricula will be actual on the long run, whereas product-related knowledge about one specific database could be shortly unusable, because the belonging software product is no longer offered and used. After education, every person becomes part of the e-skills stock (profile C) with their own portfolio (profile D), and with their participation at the labor market they build the e-skills supply (profile C+D). The relevant job profiles are managed within companies' skills framework (profile A') and are influenced by the e-skills requirements (profile A), e-skills stocks (profile C and profile D), and e-skills supply (profile C+D). It is a big challenge to compare and/or to match these (job) profiles (profile A, B, C, D, C+D, A') among themselves because of lack of a common language. The degree of difficulty is much higher on the European level.

Web service adoption in the industry has allowed organizations to share information with their partners, providers, and customers in a standardized manner. Web services provide a successful way to communicate distributed applications in a platform-independent and loosely coupled manner. The integration of applications is one of the main challenges when building ICT solutions. Integration is often achieved using costly customized solutions for every pair of applications. Web service technologies are a set of standards that allow software interfaces to be defined using XML (extensible markup language) as the message format and the Internet infrastructure for message transport. While lowering costs, Web service technologies by themselves do not ensure that two businesses use the same data structures or business protocols; neither do they provide the means to resolve potential conflicts (Lausen, Brujin, Keller, & Lara, 2006). In order to deal with existing interoperability issues, ontology-based models for describing Web services are proposed. Hence, Web services are in the wider context of analysis, we will look at issues of common semantics to enable communication of and automated data exchange between various applications in HR systems.

We see that electronic data interchange and integration has potential to make internal HR management processes more flexible and faster, as well as more efficient and effective.

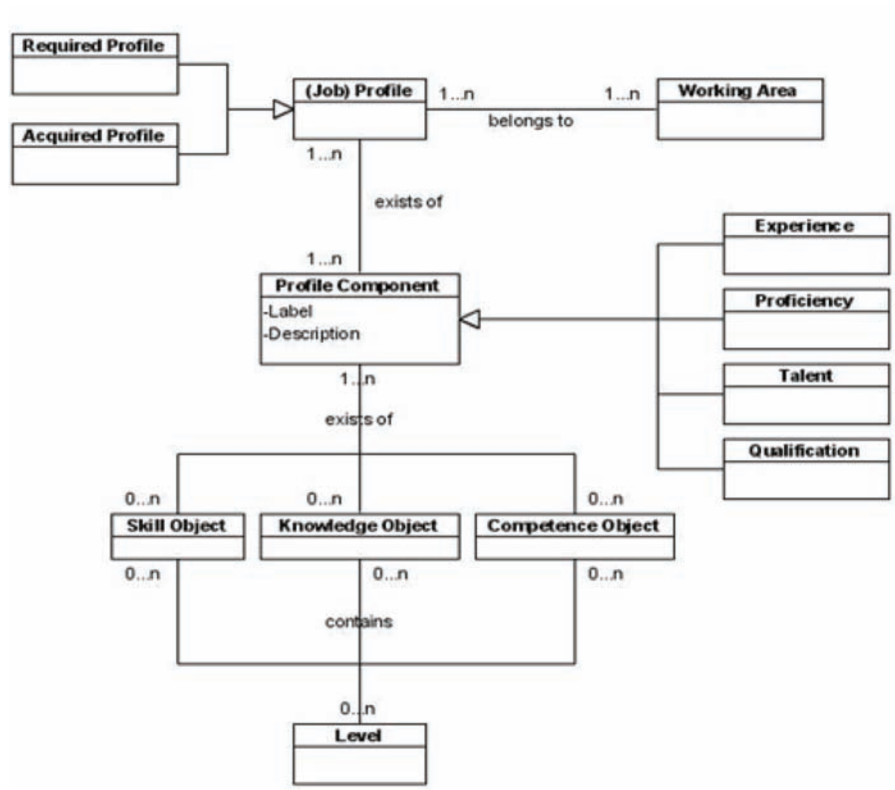
CONCEPTUAL DATA MODEL

During schooling and education, for (the first) letter of application, for a successful professional advancement, and more: the whole working life belongs to the crucial points KSC and qualifications. There are standardized data models describing competences in a re-usable way: IEEE Reusable Competency Definition (RCD, 2007) and HR-XML (2004) (further information in Povalej and Weiß, 2007). When certain competences are once defined they can be re-used again and again, for example, in different job advertisements or describing several job profiles. Further on, employees can use the already defined competences to describe their own ones.

Here are some clues based on the results of the analyses of these two standardized data models which could be extracted for our conceptual data model (Povalej & Weiß, 2007): the re-usability and the flexibility are strengthened through an object oriented approach; one competency consists of the components skills, context, and proficiency level; context and proficiency level has to be clearly definable which is a precondition for profile matching and domain-wide re-usability; free description area out of RCD is needed and it represents the smallest skills unit; additional, the component grade is introduced. Figure 2 displays the connections and dependencies between relevant objects or classes of a job profile model.

A job profile describes among others, KSC and qualifications which employees have or should have to cope with work tasks successfully within their working areas. Hence, there are two kinds

Figure 2. Example of a job profile model



of profiles: (1) the required profile contains all information about KSC and qualification which employees should have to manage their tasks in a company and (2) the acquired profile contains all information about KSC and qualification which employees gained in their life. Every (job) profile belongs to at least one working area. Further, on the (job) profile exists of one or more profile components—each profile component represents experience, proficiency, talent, or qualification. Every profile component exists of none, one, or many KSC objects, whereby every component contains none or one level.

The code in Figure 3 is an extract from a HR-XML schema describing competence and level belonging to a job profile.

Figure 4 displays exemplarily the component profile of our conceptual model (Povalej & Weiß, 2007).

CONCEPTUAL META-FRAMEWORK

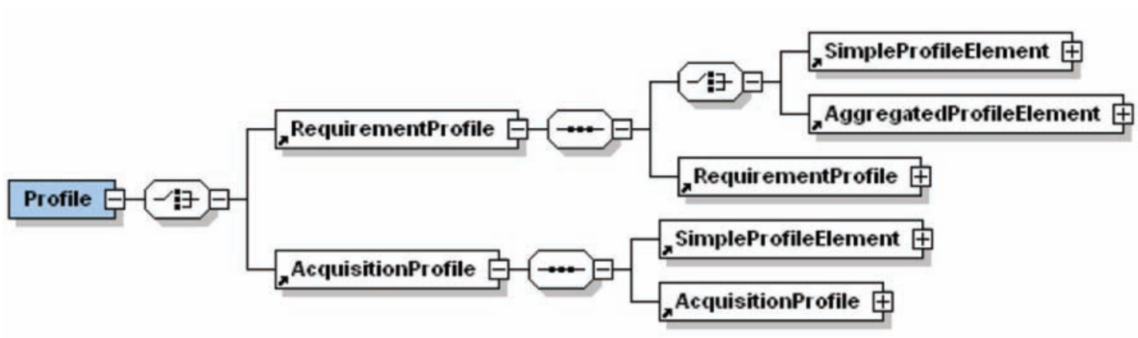
However, in order to achieve transparency and comparability of ICT worker profiles, data structures need to be analyzed and their semantics or meta data needs to be described. Subsequently, we look at the modeling of profile structure and semantics. Meta data is preferably described and stored in the form of ontologies. Preparatory work to create ontologies requires analyzing the underlying data structure in a formal way, which then determines our structure of data sets stored in profiles. We present the results of our analysis of meta data applied in major ICT skills/competence frameworks.

The smallest building block or unit of qualifications has to be defined (CEN/ISSS, 2005; Weiß, Povalej, & Stucky, 2005; Weiß, Stucky, Dolan, & Bumann, 2005).

Figure 3. Exemplary extract from a HR-XML schema

```
[...]
<xsd:simpleType name="CompetenceWeightType">
        <xsd:restriction base="xsd:string">
                <xsd:enumeration value="compLevelTarget"/>
                <xsd:enumeration value="compLevelActual"/>
[...]
<xsd:complexType name="CompetenceType">
        <xsd:sequence>
                <xsd:element name="CompId">
                        <xsd:complexType>
        <xsd:attribute name="id" type="xsd:string" use="required"/>
[...]
```

Figure 4. Exemplary component profile (Source: Povalej & Weiß, 2007)



In the following section, the developed matrix of profiles P is described (further information: Weiß, Povalej, & Stucky, 2005; Povalej & Weiß, 2007). In our approach, a profile constitutes on three sub-systems to be defined and specified by a respective modeler/developer: 1) the general description of qualification component (QC) (e.g., constructs described as learning outcome (LOC) by KSC), 2) the specific description of the working area (WA), job role, function, business process, and so forth, and 3) the performance level required. These three elements combined, define the qualification unit (QU) as the smallest describable and measurable unit within a qualification or profile P. These results are summarized in the definition of ICT worker sector profile P. An ICT worker profile is defined as the selection of the tupel (WA, QC, L) and can be described as matrix P with the job role q as follows (notation: Figure 5) (our approach corresponds to the central concepts or terms suggested by Cedefop, 2003):

$$P_q = L_q = \left(l_{qij}\right)_{j=1...n}^{i=1...m} = \left(wa_i, qc_j\right)_{j=1...n}^{i=1...m} \qquad (1)$$

FUTURE TRENDS

CEN/ISSS (2007) is working on a European e-competence framework. A team of experts has been appointed and will produce guidelines for the developments of a European e-competence framework for consensus building. The team involves experts from stakeholders of leading frameworks such as SFIA in UK, AITTS in

Figure 5. Overview of applied notation (Source: Weiß, Povalej, & Stucky, 2005)

Element/ concept	Notation
P	set of profiles; $P = \{q_1, \ldots q_m\}$
L	set of levels (vertical)
QC	qualification component as general description of the smallest unit of a qualification to be described; QC may be expressed as set of LOC on basis of KSC in relation to the intended application context (QC = LOC) (e.g. general European Qualification Framework descriptions)
WA	set of working areas (horizontal level e.g. 'sales and marketing', 'management and administration'; this is applied synonymously with the terms 'field of work', 'function', 'business process' etc.); $WA = \{wa_1, \ldots wa_m\}$; if higher granularity is required, WA may be subdivided in further sub-categories
q	job role e.g. software developer, information system analyst, etc.
i, j	indices to define KSC category based on typology (WA, QC)
m, n	indices for existing numbers of categories for WA, QC

Germany, CIGREF in France, and ECCO in Italy. The European e-competence framework is addressed to HR managers, ICT managers, and ICT practitioners/ professionals as well as training institutions and educational bodies. The framework intends to contribute to the daily work of HR managers as an international planning and development tool.

The results will be a basic structure for the development of vocabularies and skills/competence dictionaries. Building vocabularies and dictionaries requires in depth domain knowledge of ICT competences in industry. Hence, the European e-competence framework delivers the content which populates profiles and instantiates described meta data structure.

KSC described in coherence with the framework structure allows exchanging and reusing this information in various HR and CPD processes. The framework builds a grid which is spanned mainly by the dimensions competence areas and respective descriptors to specify attitudes, skills, and knowledge. Furthermore, levels and the orientation towards ICT processes are considered. The next steps require to populate the grid and to build dictionaries and catalogues of competence descriptions. The reuse of data and the exchange of information between various systems can be seen as particular challenges.

Portfolios as well as e-portfolios are closely linked to (job) profiles. Hence, consistently, there should be a more detailed review about the connection between these areas. Butler (2006) made the following recommendation for the success of e-portfolios: "Institutions need to be aware of the impact that an electronic portfolio development will have. Electronic portfolios need to be an integral part of a programme of study, not an 'added-on' assessment, which may necessitate the review and restructuring of courses. The type of portfolio required, its purpose and its audience need to be clearly articulated."

CONCLUSION

A detailed consideration of the supply and demand of e-skills (or further knowledge, skills, competences, and qualifications) on the labor market is crucial for the development of a common European framework. It has been pointed out that a common language and a shared common framework are needed.

Further, this article accentuated the formal description of the smallest unit as a building block of qualifications and for analyzing ICT worker profiles. Qualification units are the smallest element to be scrutinized in respective competence assessments and in the certification process. By the help of our developed and presented conceptual data model (derived from standardized data models) as well as our conceptual framework (derived from existing frameworks using the meta-level), ICT competence profiles, or in general, profiles can be modeled at the meta-, reference, and application level (Weiß, Povalej, & Stucky, 2005).

Hence, there are now the necessary basics to develop further Web services for human resources. Central aims are to create the "right" conditions to attract talented people by respective career guidance, information services, and by highlighting flexible entry points to the ICT profession—these all offered as Web services.

The next steps are to develop further the meta-framework for KSC and qualifications. Further on, our up to now work is also the basic for further development of KSC and qualification catalogues and repositories. At the same time the conceptual models will be implemented to real world systems, considering these catalogues and repositories, and also in the future, Web services for a higher transparency, comparability, transferability, and interchangeability.

REFERENCES

Anderson, J. E., Barret, K. S., & Schwager, P. H. (2005). Informing the HR hiring decision of IT personnel: The HR professional's view of IT certification, education, & experience. *Informing Science Journal*, *8*, 281–302.

Butler, P. (2006, October). A review of the literature on portfolios and electronic portfolios. *Creative Commons Attribution-NonCommercial-ShareAlike 2.5 License*.

Cedefop (The European Centre for the Development of Vocational Training). (2003). Credit Transfer In Vet NOVEMBER 2002-OCTOBER 2003. Author: Isabelle le Mouillour, University of Kassel, Germany. Retrieved October 15, 2007, from http://communities.trainingvillage.gr/credittransfer-eqf

Cedefop (The European Centre for the Development of Vocational Training). (2005). European reference levels for education and training promoting credit transfer and mutual trust. Study commissioned to the Qualifications and Curriculum Authority, England. Authors: Mike Coles, Tim Oates. *Cedefop Panorama series; 109, Luxembourg: Office for Official Publications of the European Communities, 2005.*

CEN/ISSS ICT-Skills Workshop Phase 2. (2005). *CEN Workshop Agreement (CWA)*. Retrieved October 15, 2007, from http://www.uninfo.polito.it/WS_ICT-Skills

CEN/ISSS ICT-Skills Workshop Phase 3. (2007). Retrieved October 15, 2007, from http://www.cen.eu/CENORM/BusinessDomains/businessdomains/isss/activity/wsict-skills.asp

ESF (The European E-skills Forum). (2005). *E-skills for Europe: Towards 2010 and Beyond*. Luxembourg: Office for Official Publications of the European Communities, Synthesis Report. 2005.

HARMONISE. (2007). Survey of certification schemes for ICT professionals across Europe towards harmonisation. Project of CEPIS Council of European Professional Informatics Societies, technical report, confidential. Retrieved October 15, 2007, from http://www.cepis-harmonise.org

HR-XML Consortium. (2004). *Competencies*. Retrieved October 15, 2007, from http://www.hr-xml.org

Hughes, C., &, Thompson, C. (2007, June). The international IT professional practice programme: ICT certifications for informatics professionals. *UPGRADE, VII*(3).

Lausen, H., de Bruijn, J., Keller, U., & Lara, R. (2006, October). Semantic web services with WSMO. *UPGRADE, VII*(5), 34–37.

Povalej, R., & Weiß, P. (2007, September). Conceptual data model and meta framework: Modeling ICT competence profiles. *I-KNOW 2007.* Retrieved from http://www.i-know.at/

RCD. IEEE Learning Technology Standards Committee RCD. (2007). *Draft standard for reusable competency definitions (RCD).* Retrieved May 12, 2007, from http://ieeeltsc.org/wg20Comp/Public/IEEE_1484.20.1.D3.pdf

Weiß, P., Povalej, R., & Stucky, W. (2005, December). *Learning across boundaries: Developments towards a European framework for ICT skills*. Paper presented at the International Conference on Internet & Information Technology in Modern Organizations.

Weiß, P., Stucky, W., Dolan, D., & Bumann, P. (2005): E-skills certification in Europe towards harmonisation. In P. Cunningham & M. (Eds), *Innovation and the knowledge economy, issues, applications, case studies.* Amsterdam: IOS Press.

World Wide Web Consortium. (W3C). (2007). *Web services activity statement*. Retrieved October 15, 2007, from http://www.w3.org/2002/ws/Activity

KEY TERMS

Certification: One process that formally validates the knowledge, skills, and competences (and/or know-how, know-what, know-why) acquired by an individual, following a standard assessment procedure.

Data Modeling: The process of analyzing and identifying data objects and their relationships amongst each other within certain contexts like business or others where they are used. It is a first step into object-oriented programming,

E-Portfolio: A Web-based information management system that uses electronic media and services. In the first line, it is used for self-expression, for example, trying to establish contacts for private or professional matters.

Framework: Set of assumptions, concepts, conditions, methods, preconditions, values, practices, and so forth, that constitutes a way of viewing reality respectively, that supports through making available blueprints within certain contexts, like ICT development or building projects.

Human Resources: A modern conceptuality in the economy, it describes persons as manpower and/or resources. Through this, the importance of employees should be emphasized.

Information and Communication Technologies (ICT): Technologies which provides for the electronic input, storage, retrieval, processing, transmission, and dissemination of information. Often, ICT are linked with a particular context, for example, ICT in education, health care, or libraries.

Job Profiles: Written definitions of primary job duties, key responsibilities, necessary KSC, and qualification, education, experiences, proficiencies, talent, personal characteristics sought in a candidate, and so forth. They belong to certain working areas and are subdivided into components which exist of objects like KSC, including the belonging level. There are required and acquired ones.

Meta Modeling: The analysis, building up, and development of constraints, concepts, frames, methods models, rules, theories, and so forth, practicable and useful for modeling within certain domains to cope with predefined challenges. In general, a model is an abstraction of objects in the real world; a meta-model is an abstraction of model itself.

Qualification: Official record (certificate, diploma) of achievement which recognizes successful completion of education or training, or satisfactory performance in a test or examination.

Web Services: Enterprise applications that exchange data, share tasks, and automate processes over the Internet.

This work was previously published in Encyclopedia of Human Resources Information Systems: Challenges in e-HRM, edited by T. Torres-Coronas; M. Arias-Oliva, pp. 84-91, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 1.11
Web 2.0 and E-Discovery

Bryan Kimes
Altria Client Services, Inc., USA

ABSTRACT

Companies today face an overwhelming amount of digital information, and many of them are involved at some point in civil litigation. When a company is in the discovery (pretrial) phase of civil litigation, it usually exchanges information, including documents, with the opposing party in the litigation. The Federal Rules of Civil Procedure, which govern civil litigation in federal courts, were amended in 2006 to provide additional guidance to parties with regard to electronically stored information. The management teams of many U.S. corporations are working with their IT departments and lawyers in order to understand the sources of electronically stored information that may be potentially relevant to their litigation. Over the last 20 years, technology has grown increasingly more complex, from the early mainframe and personal computers to sophisticated e-mail and instant messaging applications that enable users to send and receive millions of messages every day. This chapter addresses the issues companies may face related to the discovery of electronically stored information as a result of new communication technologies, including Web 2.0 applications.

INTRODUCTION

Companies today are dealing with new technologies and a growing amount of digital information. Each year, the global information growth is larger than any previous year, and that trend is not showing any signs of slowing down. In 2002, an estimated 5 exabytes of digital information were created worldwide. For reference, 5 exabytes is the equivalent of 37,000 libraries the size of the Library of Congress (University of California at Berkley, 2003). The annual growth is roughly doubling each year, and U.S. corporations are contributing to this trend. Electronic communication technologies, including e-mail as well as next-generation Web 2.0 social technologies (e.g., instant messaging [IM], blogs, and wikis), are also contributing to this rapid growth. For example, Cellular-News reported that Gartner predicts that an estimated 1.9 trillion instant messages were sent worldwide in 2007 ("SMSs to Surpass 2 Trillion

DOI: 10.4018/978-1-60566-122-3.ch006

Messages in Major Markets in 2008," 2007). These new technologies are transforming modern organizations and their employees. Corporate technology managers are assessing the potential benefits of these new technologies relative to the burden and risk they present to the organization, including that from litigation.

This chapter is not intended to be a comprehensive review of the phases of civil litigation or the Federal Rules of Civil Procedure (FRCP). In addition, the author of this chapter is not a lawyer and the information in this chapter is generalized and not intended for specific situations, and should not be substituted for the advice of a lawyer. The following is an overview of the business issues related to managing electronically stored information that may be required in litigation. By way of background, many companies will face civil litigation, which is when one individual or organization sues another seeking relief, oftentimes for money damages.

E-DISCOVERY BACKGROUND

The *United States v. Microsoft* case, which was tried in 1998, is a famous civil case in which the United States Department of Justice (DOJ) filed an antitrust claim against the software giant alleging it used its monopoly power of the Windows operating system to bundle its Internet Explorer Web browser. The government claimed that other Web browsers, such as Netscape, were not able to compete in the Web browser market due to Microsoft's positioning of Internet Explorer within the Windows operating system. The DOJ won in the trial court, but an appeals court overturned the decision and the case later settled. The Microsoft case was full of headlines and was an example of a case where e-mail played a critical role. The DOJ trial lawyers used internal Microsoft e-mails, many authored by senior executives, including Bill Gates, to demonstrate the company's alleged strategy to "cut off Netscape's air supply" (United States Department of Justice, 1999).

Civil litigation follows specific procedures. Cases in United States federal courts follow the Federal Rules of Civil Procedure. The Federal Rules define the rules into different categories, including Commencement of Action, Pleadings and Motions, Parties, Deposition and Discovery, Trial, and Judgment. The remainder of this chapter will focus on the discovery phase of civil litigation. During the discovery phase of a case, thousands or even millions of documents may be collected, reviewed, and produced from the files of either party in the lawsuit.

Let us use a hypothetical example to illustrate the process. The Acme Corporation, a fictional product manufacturer, is involved in civil litigation. The case involves a former employee, John Doe. John filed a lawsuit against Acme alleging he was terminated as a result of age discrimination. John Doe, who is the plaintiff, hired a lawyer to represent him in his suit against Acme, the defendant. Acme also hired a law firm to represent it in the case. When the case reached the discovery phase, John's lawyer asked Acme for documents that related to John's termination. Acme's management, working with its lawyers, determined which employees had documents that related to his termination. It instructed specific employees to not discard the documents that related to the issues in the case. Acme's IT director worked with his management to determine the locations where electronic documents that related to the case were stored.

Now, let us walk through three variations of the Acme example at different points in time. In the first variation, Acme terminated John in 1990. In 1990, most companies did not have company-wide e-mail accounts or network servers. Wireless cellular telephones were just getting started and smart phones were still a science fiction concept. The computer systems at Acme, including those used by John and his management, were probably running MS-DOS or Windows 2.0 (the classic 286 PC). If Acme was on the cutting edge of computer technology in 1990, it may have had Windows 3.0, which ran on the 386 powerhouse. In order to respond to the John Doe lawsuit, Acme's management contacted the employees that may have had records that related to John's termination, including the personnel files from John's manager, the human resources (HR) files that related to John and Acme's employment policies. In order to find the electronic documents, Acme's management went to the IT director who managed all of Acme's computer systems. The IT director managed a relatively small system that included a mainframe computer system. The IT director worked with management to find material regarding John's case.

Now let us assume Acme terminated John in 2000. In 2000, Acme had been using e-mail for several years now and employees regularly use e-mail to correspond with each other to get their work done, including personnel matters. Acme had Windows 98 running on employees' desktops, which allowed employees to save documents, for example, in Microsoft Word, to their computers. Employees saved documents to the computers' desktops, local drives, and some network file servers. Acme's management followed a similar process for collecting the documents as it did in 1990. However, because the electronically stored information at Acme has multiplied since 1990, Acme's job to identify the relevant documents has become more complicated. Acme management asked various employees that worked with John, including his manager, to identify the documents that related to the case. It also worked with the IT director to understand how the computer systems were managed and how employees used those systems. Acme's cost associated with the discovery phase of the case has likely grown due to the increased volume of documents that Acme's lawyers needed to review and potentially produce to John's lawyers.

Finally, let us assume that Acme terminated John in 2007. Since 2000, Acme implemented several new technologies to help its employees be more productive and communicate faster. In 2007, Acme used Microsoft Vista with SharePoint. It had Microsoft Exchange 2007 with unified messaging (voice mail integrated into users' e-mail account). Many senior employees have wireless handheld devices that integrate into their e-mail accounts as well as instant text messaging. Acme employees use these new technologies and create more documents on their desktops and rely on electronic communication tools, including e-mail and instant messaging. In 2007, the IT director has to address a more complicated network of computer systems to help management identify those systems that contain documents that relate to John's lawsuit. As a result of the increase in the number of documents, Acme's costs associated with the discovery phase of the case is likely to be even more expensive and time consuming in 2007 than in 2000 due to the higher volume of electronically stored information.

Acme is certainly not alone. Companies all over the world are facing a new generation of technology as new employees, who have been teething on instant messaging, MySpace, and Facebook, join the workforce. For example, many companies convene meetings using Web-based electronic rooms, for example, WebEx, with shared desktops. These next-generation meeting rooms allow participants to attend meetings without ever leaving their desks, and many participants may never meet each other face to face.

Technology managers, like the IT director at Acme, may have a difficult task when they are asked to identify systems and electronically stored information that may contain documents relevant to a lawsuit. Additionally, the technology generation gap is often short relative to the time it takes for courts and litigators to deal with these new issues in actual cases. It can take years for a dispute to make its way into a courtroom. Technology managers are faced with making decisions about how to implement new technology tools without knowing what, how, or when they will have to produce documents in the future.

FEDERAL RULES OF CIVIL PROCEDURE

Companies that find themselves involved in civil litigation follow certain rules. If the case is in a federal court, they use the FRCP; if the case is in a state court, the parties follow the state's rules of civil procedure, which often follow, if not verbatim, the FRCP. Again, this chapter will not be an exhaustive review of the FRCP, but it will touch on a few rules that relate to electronically stored information.

On December 1, 2006, amendments to the Federal Rules of Civil Procedure were submitted by Congress and approved by the Supreme Court. The amended rules specifically address electronically stored information. Let us start with a rule that was not amended, FRCP 26(b)(1), which defines the scope and limits of Discovery.

- **Rule 26(b)(1): Discovery Scope and Limits***In General*. Parties may obtain discovery regarding any matter, not privileged, that is relevant to the claim or defense of any party, including the existence, description, nature, custody, condition, and location of any books, documents, or other tangible things and the identity and location of persons having knowledge of any discoverable matter. For good cause, the court may order discovery of any matter relevant to the subject matter involved in the action. Relevant information need not be admissible at the trial if the discovery appears reasonably calculated to lead to the discovery of admissible evidence. All discovery is subject to the limitations imposed by Rule 26(b)(2)(i), (ii), and (iii). (U.S. House of Representatives Committee on the Judiciary, 2006)

Now let us return to the Acme example. Why is Acme's IT director's job more complicated in 2007 than in 1990? I suggest to you that it is the result of the explosion of electronic documents, specifically e-mail and other documents stored on Acme's computer systems.

Below is a summary of the electronic-discovery amendments to the Federal Rules of Civil Procedure that this chapter will discuss.

- **Rule 26(a)(1)(B): General Provisions Governing Discovery; Duty of Disclosure; Required Disclosures; Methods to Discover Additional Matter**

A copy of, or a description by category and location of, all documents, electronically stored information, and tangible things that are in the possession, custody, or control of the party and that the disclosing party may use to support its claims or defenses, unless solely for impeachment. ("Law Library/Court Rules," n.d.)

What Might this Mean for Acme? When Acme's IT director reviewed this amended rule with management and outside lawyers, they discussed, among other things, the computer systems at Acme and thought about a summary of the electronically stored information, such as in databases, servers, and electronic archives, that may be relevant to John's case against Acme.

- **Rule 26(b)(2)(B): General Provisions Governing Discovery; Duty of Disclosure; Discovery Scope and Limits; Limitations** A party need not provide discovery of electronically stored information from sources that the party identifies as not reasonably accessible because of undue burden or cost. On motion to compel discovery or for a protective order, the party from whom discovery is sought must show that the information is not reasonably accessible because of undue burden or cost. If that showing is made, the court may nonetheless order discovery from such sources if the requesting party shows good cause, considering the limitations of Rule 26(b)

(2)(C). The court may specify conditions for the discovery. (“Law Library/Court Rules,” n.d.)

What Might this Mean for Acme? When Acme’s IT director reviewed this rule with management and lawyers, they discussed, among other things, what sources of potentially relevant electronic information that they thought were not reasonably accessible and that they did not intend to search. If John’s lawyer wanted Acme to provide documents from one of the sources that Acme disclosed as not reasonably accessible, then John’s lawyer could serve a motion to compel the production of information from the sources identified as not reasonably accessible. Acme will then have to explain why the information that John’s lawyer is seeking would require undue burden or cost to obtain.

Now let us look at another hypothetical exchange. John’s lawyer wants Acme to provide information from an old HR computer system that is no longer active, and Acme considers the old system not reasonably accessible because it would require the IT director to hire a specialized vendor to restore the old system. John’s lawyer files a motion to compel in order to obtain the information contained within the computer system. Acme’s IT director explained to the court that the request is unduly burdensome; specifically, it will cost Acme US$500,000 to restore the legacy computer system. Acme feels that the burden is excessive relative to the amount John’s lawyer is seeking in the case, which is only US$100,000 in money damages. The court will have to weigh these factors when determining whether to permit discovery.

- **Rule 34(a) & (b): Production of Documents, Electronically Stored Information, and Things and Entry Upon Land for Inspection and other Purposes; Procedure** (a) SCOPE: Any party may serve on any other party a request (1) to produce and permit the party making the request, or someone acting on the requestor’s behalf, to inspect, copy, test, or sample any designated documents or electronically stored information—including writings, drawings, graphs, charts, photographs, sound recordings, images, and other data or data compilations stored in any medium from which information can be obtained—translated, if necessary, by the respondent into reasonably usable form, or to inspect, copy, test, or sample any designated tangible things which constitute or contain matters within the scope of Rule 26(b) and which are in the possession, custody or control of the party upon whom the request is served;

  (b) PROCEDURE: The request shall set forth, either by individual item or by category, the items to be inspected, and describe each with reasonable particularity. The request shall specify a reasonable time, place, and manner of making the inspection and performing the related acts. The request may specify the form or forms in which electronically stored information is to be produced. Without leave of court or written stipulation, a request may not be served before the time specified in Rule 26(d). The party upon whom the request is served shall serve a written response within 30 days after the service of the request. A shorter or longer time may be directed by the court or, in the absence of such an order, agreed to in writing by the parties, subject to Rule 29. The response shall state, with respect to each item or category, that inspection and related activities will be permitted as requested, unless the request is objected to, including an objection to the requested form or forms for producing electronically stored information, stating the reasons for the objection. If objection is made to part of an item or category, the part shall be specified and inspection permitted of the remaining parts. If objection is made to the requested form or forms for producing electronically stored information—or if no form was specified in the request—the responding party must state the form or forms it intends to use. The party submitting the request may move for an order under Rule 37(a) with respect to any objection to or other failure to respond to the request or any part thereof, or any failure to permit inspection as requested. Unless the parties otherwise agree, or the court otherwise orders:

(i) a party who produces documents for inspection shall produce them as they are kept in the usual course of business or shall organize and label them to correspond with the categories in the request;
(ii) if a request does not specify the form or forms for producing electronically stored information, a responding party must produce the information in a form or forms in which it is ordinarily maintained or in a form or forms that are reasonably usable; and
(iii) a party need not produce the same electronically stored information in more than one form. ("Law Library/ Court Rules," n.d.)

What Might this Mean for Acme? When the Acme IT director reviewed this rule with management and lawyers, they likely discussed, among other things, the options regarding how they would produce documents for John's lawyers. They discussed an upcoming meeting between their lawyers and John's lawyers, during which they would discuss the form of the document production. The Acme IT director has already met with some vendors that provide specialized tools to assist with document productions. The IT director understands that John's lawyer may specify a form of production and if he does not, then Acme will have to produce the documents in the form they are ordinarily maintained at Acme or in a reasonably usable format. Let us assume John's lawyer agreed to accept a production of the relevant documents in Adobe's Portable Document Format (PDF). In addition, Acme provided an accompanying index of metadata including the date, title, author, recipient, document type, and extracted full text.

The amended federal rules have provided organizations and their IT departments with more instruction regarding electronically stored information. However, complying with the obligations associated with electronic discovery can be very difficult for organizations, even those with a modest technology portfolio. Organizations with newer generation technologies may find themselves in a much more complicated situation when faced with discovery in civil litigation.

In a recent *Richmond Journal of Law & Technology* article, Doug Rogers, citing the Sedona Principals,[1] stated "it is unreasonable to expect parties to take every conceivable step to preserve all potentially relevant electronically stored information":

For instance, a party should not have to preserve all possible sources of e-mail simply because it is possible that an existing or former employee, who claims she was harassed, may have forwarded an e-mail from her supervisor to other employees, or because one of the e-mails may have been exported into a .pst file outside of the e-mail folders in the cache of a computer.

Comment 5a to Sedona Principle 5 explains: The obligation to preserve relevant evidence is generally understood to require that the producing party make reasonable and good faith efforts to identify and manage the information that it has identified as reasonably likely to be relevant. Satisfying this obligation must be balanced against the right of a party to continue to manage its electronic information in the best interest of the enterprise, even though some electronic information is necessarily overwritten on a routine basis by various computer systems. (Rogers, 2008, p. 25-26)

Companies like Acme are not alone when working through these issues. Over the past few years, many specialized litigation support vendors have emerged that offer a variety of technology and process solutions to assist organizations and their IT departments in managing the discovery process. In July 2007, one such vendor, Autonomy, acquired another company, Zantaz, which is an e-mail archiving firm, for US$375 million. Soon after the Zantaz acquisition, Autonomy reported a US$70 million deal with an unspecified customer. The customer is believed to be a large global bank that is preparing for a wave of civil litigation in the wake of the subprime mortgage lending crisis. The value of this transaction, which is Autonomy's largest, is reported to be almost double Autonomy's average transaction with a single customer. If this type of transaction is any indication of future business, Autonomy could see a return on the Zantaz acquisition much faster than many anticipated (Mellor, 2008).

Companies and organizations around the world are evolving their technology footprints. Many commentators believe the volume of documents that employees create on new applications are also increasing at a rapid rate, some say exponentially. The legal environment is also evolving and new cases are emerging that deal with new issues regarding electronically stored information. One case that has gained some notoriety is the Columbia Pictures case (*Columbia Pictures Indus. v. Bunnell et al.*). In this case, Columbia Pictures was ordered to capture and preserve data stored in temporary random access memory (RAM), which is very volatile and can be overwritten by the computer system when shutting down, starting up, or simply opening and closing applications. Some commentators have stated that the judge's holding is limited to the specific facts at issue in this case and generally considered data in computers' temporary RAM to be excluded from discovery because they are so transitory (Jacoby, 2007). Some commentators have expressed that Rule 34, which describes "other data or data compilations stored in any medium from which information can be obtained," would not apply because the RAM data were not easily obtainable. In this case, however, the judge stated that the data stored in this computer system's RAM was "extremely relevant and may be key," and that it was not available anywhere else (Jacoby).

A notable distinction of the temporary, or ephemeral, data at issue in the Columbia Pictures case is the way the specified computer system was configured. Many computer systems use RAM as a memory buffer that will randomly (hence the name *random access*) write some data to help support the application. For example, take an ordinary desktop computer. When a user turns the computer on, the computer will allocate some RAM to support the operating system. If the user then opens Microsoft Word, the computer will allocate more RAM for the application and the associated documents that are open. If the user then opens Excel and closes Word, the RAM assigned to Word is sometimes overwritten in order for the computer to run Excel. The computer, during normal operation, will recycle RAM memory to whatever application(s) needs it. The information stored in RAM, if viewed by a human, would most likely appear as gibberish as it is bits and bytes of computer code that the computer uses to function properly.

Nevertheless, the computer systems at Columbia Pictures, based on its own testimony, had a logging feature that was turned off. The court ordered Columbia to prospectively activate the logging feature and that the prior deletion of RAM was not a violation of the preservation order in the case (Jacoby, 2007).

Instant Messaging is now a ubiquitous technology across grade schools, college campuses, shopping malls, and movie theaters (much to my chagrin). Instant messaging took off in the mid-1990s with the popular ICQ ("I seek you"). This was followed by AOL's Instant Messenger in the late 1990s. Today, there is a variety of open-source instant messaging, or text messaging, applications that run on desktop computers, mobile phones, and other wireless devices. Many instant messaging applications, especially those on mobile phones, are similar to RAM in that many of them do not keep a permanent log of past messages. Many mobile messaging devices keep messages for a short period of time, sometimes only minutes or hours, before they are overwritten or deleted.

IM is slowly invading the workplace. Instant messaging allows employees to communicate faster than e-mail and slightly slower than a real-time phone call. Users can carry on several conversations at the same time. They can create buddy lists, which allow for quicker communication as well as the ability to see if someone is online or not. In a *Wall Street Journal* online article, Mamberto (2007) cites a 2006 study of the American Management Association and the ePolicy Institute that found roughly one third of employees are using IM, many of them without their employers' knowledge. Many corporations and organization are reluctant to deploy IM systems due to security concerns, loss of productivity, and inappropriate use. Despite these reservations, Gartner, an IT research firm, predicts IM will dominate 95% of the workforce of large companies within the next 5 years (Mamberto).

Desktop-computer-based IM applications are slightly different than mobile-phone-based systems, notably with regard to storage. Many desktop-based IM applications provide options for longer term storage, similar to conventional e-mail systems. When Acme's IT director implemented desktop IM, he or she may have worked with the records, compliance, and law department to determine whether Acme needed to include a storage component in the event it would need to save some messages for a longer term purpose, either for business or litigation purposes.

Web 2.0 technologies, including IM, offer employees at large organizations instant access to colleagues across the organization. These technologies flatten out the hierarchical structure of many organizations. Andrew McAfee, associate professor at Harvard Business School, states employees "increasingly react to situations and problems on the fly, not solely by hierarchy" (Mamberto, 2007, p. 2).

The American Bar Association (ABA; American Bar Association Legal Technology Resource Center, 2008) cited a January 2008 *Detroit Free Press* report (Wendland, 2008) that IM text messages obtained from the offices of Detroit Mayor Kwame Kilpatrick suggested he had a 2-year affair with his chief of staff Christine Beatty. Over 14,000 text messages between Kilpatrick and Beatty, who were both married at the time, were exchanged that allegedly illustrated a sexual affair including during official business trips. The lawsuit, which was filed by a former police officer, alleged that he and another officer were fired in retaliation because of their role in an internal investigation that would potentially expose the mayor's affair. Mayor Kilpatrick testified that he did not have an affair with his chief of staff and that he did not retaliate by firing two police officers. The City of Detroit tried to avoid producing the damaging IM text messages, but eventually they were disclosed. The case eventually settled for US$8 million.

The ABA article (2008) further reported that an estimated 20 billion IM text messages are sent each month in the United States alone. The article cited a Gartner report, which estimated that 1.9 trillion text messages were sent worldwide in 2007. The report predicted a 19% increase to 2.3 trillion text messages in 2008. When Acme's IT director implemented desktop IM, he or she worked with the records, compliance, and law department to consider how to train employees on the proper use of IM, similar to the corporate e-mail system.

Blogs (Web logs), wikis (collaborative Web sites), and RSS (really simple syndication) are also on the rise within companies and organizations. These new technologies can offer significant advantages to organizations. Most of the advantages are in response to the limitation of e-mail. For example, a user can subscribe to relevant work sites using RSS feeds as opposed to relying on e-mails from colleagues. E-mail can be a very inefficient way to keep employees updated on a given topic because many employees will overuse the carbon copy (CC) function as a way to inform people on the distribution list. There are two problems with this technique. First, the person sending the message may not know all the people that really need to be informed of the topic. Second, many people on the distribution list may not even care about the message.

One of the advantages of RSS as a communication tool is that users can decide what they are interested in and will subscribe to those topic areas. The RSS technology will notify users of updates and direct them to the source of the information. Many of the RSS sites will keep track of the pages viewed so users do not need to worry about reading articles twice because it is tracked by the system ("The Logic of Blogs and Wikis in the Enterprise," 2005).

Now, let us return to the Acme example, specifically, a project to develop a new product. The Acme project team created an electronic workroom where project team members could save and update important documents related to the product. The workroom had several blogs and wikis for employees to discuss current issues related to the project, including materials, sales strategies, and budgets. Managers who were not working on the project day to day could use the RSS tool to subscribe to the workroom. The manager would be notified when new content had been added to the workroom on a specific topic.

Blogging is still a relatively new technology and companies are still learning and assessing the security risks related to employee blogging. John Browning (2007) of the *Houston Business Journal*, citing a 2006 survey of the Employment Law Alliance, found that about 5% of Americans have a personal blog, while only 15% of their employers have a blogging policy. Additionally, Browning stated that an estimated 27% of adults read blogs according to the Pew Internet and American Life Project. Employees may make seemingly innocent statements on a public blog that can put a company, and the employee, at great risk. For example, if one of the Acme employees working on the new product was at home spending time on a personal blog, he or she may post an entry about how great Acme is and that it has a great future—referencing the new product without actually mentioning it by name. The employee is a hardworking and dedicated individual who really believes in Acme and is proud to work there. The employee may not even have to mention the new product by name but could indirectly disclose trade secrets or forward-looking financial information that could be damaging to Acme. Before the lightening fast World Wide Web, an innocent comment to a neighbor in the driveway presented some risk to the company, but in the days of online blogs, where users from all over the world can easily search and find just about anything on anything, a seemingly innocent comment is now within infinitely greater reach. Again, the employee probably had no intention of making a public statement about Acme's financial future. Many companies have existing policies that govern employee communication, which would generally cover employee blogging. However, the Society for Human Resource Management reports that a small number (3%) of firms have implemented blogging policies specifically (Browning).

CONCLUSION

Employees are faced with new technologies every year, and the current wave of Web 2.0 technologies will soon be replaced with a new generation of tools that will enable employees to share information faster and easier. Information technology managers must make decisions regarding these changing technologies and must work with their management to assess the pros and cons of implementing them. Many IT managers realize that employees will create and store documents on these new systems and those documents may be needed for future litigation. Furthermore, as the overall volume of documents increases as a result of newer technologies, the overall cost to organizations will also likely increase due to the time and expense for lawyers to collect, review, and produce these documents in litigation. As the volume of discoverable electronically stored information grows, the real winners may be the specialized vendors, like Autonomy, that offer specialized document management and litigation support services to companies dealing with huge volumes of data. If the current data-growth trends continue, coupled with increased user adoption of Web 2.0 technologies, then companies, including their management, IT managers, and lawyers, will have a lot of work ahead of them.

REFERENCES

American Bar Association Legal Technology Resource Center. (2008). *Text messaging and the importance of a records retention policy.* Retrieved February 25, 2008, from http://meetings.abanet.org/ltrc/index.cfm?&data=20080204#E6AADE50-02C7

Browning, J. (2007). Adopting a blogging policy could limit your exposure to legal risk. *Houston Business Journal.* Retrieved February 25, 2008, from http://houston.bizjournals.com/houston/stories/2007/05/28/smallb2.html

Jacoby, C. J. (2007). E-discovery update: Discovery of ephemeral digital information. *Law and Technology Resources for Legal Professionals.* Retrieved February 25, 2008, from http://www.llrx.com/columns/fios19.htm

Law library/court rules. (n.d.). *LexisNexis Applied Discovery.* Retrieved February 25, 2008 from http://www.lexisnexis.com/applieddiscovery/lawLibrary/courtRules.asp

Mamberto, C. (2007). Instant messaging invades the office: Companies say it spurs broader collaboration—and scares some bosses. *Wall Street Journal Online.* Retrieved February 25, 2008, from http://online.wsj.com/article/SB118523443717075546.html

Mellor, C. (2008). Sub-prime crisis god news for legal discovery software. *Techworld.* Retrieved February 25, 2008, from http://www.techworld.com/storage/news/index.cfm?newsid=11042

Rogers, D. (2008). A search for balance in the discovery of ESI since December 1 2006. *Richmond Journal of Law & Technology.* Retrieved May 5, 2008, from http://law.richmond.edu/jolt/v14i3/article8.pdf

SMSs to surpass 2 trillion messages in major markets in 2008. (2007). *Cellular-News.* Retrieved February 25, 2008, from http://www.cellular-news.com/story/28126.php

The logic of blogs and wikis in the enterprise. (2005). *The Gilbane Report*, *12*(10). Retrieved February 25, 2008, from http://gilbane.com/gilbane_report.pl/104/Blogs__Wikis_Technologies_for_Enterprise_Applications.html

United States Department of Justice. (1999). *U.S. Department of Justice proposed findings of fact.* Retrieved February 25, 2008, from http://www.justice.gov/atr/cases/f2600/2613-1.htm

University of California at Berkeley's School of Information Management and Systems. (2003). *How much information?* Retrieved January 2, 2008, from http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm

U.S. House of Representatives Committee on the Judiciary. (2006). *Federal rules of civil procedure*. Retrieved February 25, 2008, from http://judiciary.house.gov/media/pdfs/printers/109th/31308.pdf

Wendland, M. (2008). Most text messages just vanish. *Detroit Free Press Online*. Retrieved February 25, 2008, from http://www.freep.com/apps/pbcs.dll/article?AID=/20080125/COL11/801250422/1081/COL

ENDNOTE

1 The Sedona Principals: Best Practices, Recommendation & Principles for Addressing Electronic Document Production is a project of The Sedona Conference Working Group on Best Practices for Electronic Document Retention and Production. The Sedona Conference is a non-profit organization of lawyers, consultants, academics and jurist.

This work was previously published in Social Software and Web 2.0 Technology Trends, edited by P. Deans, pp. 68-78, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 1.12
The Power and Promise of Web 2.0 Tools

G. Andrew Page
The University of Alaska Anchorage, USA

Radwan Ali
Kennesaw State University, USA

ABSTRACT

The key idea that sets constructivism apart from other theories of cognition was launched about 60 years ago by Jean Piaget. It was the idea that what is called knowledge does not and cannot have the purpose of producing representations of an independent reality, but instead has an adaptive function (Von Glasersfeld, 1996, p.3). In this chapter, a variety of Web 2.0 applications and their affordances are presented and discussed in relation to constructivism in higher education. The aim is to explain how these applications can be used in higher education to promote interactive and engaging learning environments. Recommendations for harnessing the potential of these tools along with practical examples will assist facilitators of higher education with creative means to design their courses and thus promote Learning 2.0.

DOI: 10.4018/978-1-60566-654-9.ch010

WHAT IS WEB 2.0?

The Internet has presented great opportunity for global human participation by transcending geographical, cultural, religious, social classification, and political barriers. As the proliferation of learning with technology increases, there is also amplification in the array of technological possibilities for a variety of asynchronous and synchronous interactions. Therefore, it becomes necessary to provide insight into the effective use of these technologies and the facilitation of e-learning. According to Schrum and Hong (2002) the goals of teaching with technology should include facilitating higher-level, thinking skills, such as analysis, synthesis, and evaluation. Alexander describes this phenomenon:

Web 2.0 is defined as a way of creating webpages focusing on microcontent and social connections between people. It also exemplifies that digital content can be copied, moved, altered, remixed, and linked, based on the needs, interests, and abilities of users—quite against the grain of both traditional and recently expanded copyright (Alexander, 2008, p.151).

There are many creative Web 2.0 applications, tools, and services available online (See Appendix A for a list of digital resources and links). These tools or web-based learning objects have the potential to engage and involve the learner with technology as opposed to having a student learn from a computer module or digital lecture. There are many Web 2.0 tools that can be found at different sites on the web. For the purpose of this chapter, we will discuss blogs, wikis, widgets, nings, plugins, social networking (MySpace and Facebook), and virtual environments (Second Life). What issues will Web 2.0 solve? What are the potential learning outcomes from using Web 2.0? It is the promises of positive educational outcomes that web-based tools can potentially produce through quality interaction at a meta-level that are intriguing many educators and trainers.

This chapter has implications for instructors, students, instructional designers, and administrators involved with e-learning in higher education. This chapter also provides a synthesis of e-learning issues and an overview of Web 2.0 tools for promoting a constructivist environment.

The 5th edition of *The Principles of Instructional Design* (Gagne, Wager, Golas, Keller, 2005), states a very important instructional design question which is often overlooked. The question is "For what problem is instruction the solution?"(p.23). This question is especially relevant considering the numerous challenges and limitless potential of Web 2.0 tools. By using Web 2.0 tools the theory of constructivism can be applied to reach new levels of digital creation and e-learning.

Constructivism is the process of linking new understanding to old, modifying and enriching existing knowledge, and enhancing depth of comprehension about a topic. McFedries refers to Web 2.0 as functioning as a platform (p.68) because of the ability to delete, edit, and add content and work collaboratively with others in a synchronous approach.

"Constructivism affords a knowledge building process that engages active learners with the physical and social world" (Twomey-Fosnot, 1996, p.30). These interactive online tools include portals, blogs and video blogs, widgets, plugins, wikis, conferencing, games, survey instruments and games. Using these Web 2.0 tools in an educational setting builds on Vygotsky's (1978) view of interactive learning within the zone of proximal development (ZPD) and how the social process is crucial to the development of thought and behavior patterns. The application and integration of quality educational content is an important part of the constructivist framework. This is the adaptive function of constructivism alluded to by Von Glasersfeld (1996).

When interactive technology is applied in accordance with constructivist principles, it has been called "learning 2.0" (Brown & Adler, 2008). Learning 2.0 is about actively using technology. The multiliteracies of education (e.g., critical thinking, creating content, collaboration) are fostered due to the constructivist nature and exchange of knowledge that is part of the Web 2.0 phenomenon (Alexander, 2008). Learners taking the initiative to create content and learning by being engaged and involved is at the heart of constructivist practice. Brown and Adler (2008) acknowledge the impact of Web 2.0 and how it can add a new dimension to communication and participation:

The latest evolution of the Internet, the so-called Web 2.0, has blurred the line between producers and consumers of content and has shifted attention from access to information toward access to other people. New kinds of online resources—such as social networking sites, blogs, wikis, and virtual communities— have allowed people with common interests to meet, share ideas, and collaborate in innovative ways. Indeed, the Web 2.0 is creating a new kind of participatory medium that is ideal for supporting multiple modes of learning. (p.18)

THE USE OF WEB 2.0 APPLICATIONS TO PROMOTE CONSTRUCTIVIST LEARNING

This section will consider the benefits of constructivism and match them with the characteristics of a few of Web 2.0 applications. Constructivism, as a learning theory, provides an appropriate fit for the use of Web 2.0 applications. Constructivism as a learning framework can be served immensely by using the tools and affordances (Jonassen, 1999) presented by the Web 2.0 applications. Constructivism suggests that learners can enjoy their experience by being more involved in their learning than by being listeners (Franklin, 2000; Ge & Tok, 2003). The idleness or boredom aspects of passive listening prevent learners from realizing the potential of the subject matter being discussed. However, by engaging in relevant learning activities, the learner will have the opportunity to apply the concept learned. Web 2.0 applications can be very engaging for different reasons: the novelty of the technologies, the aesthetic appeal, and their interactive modes.

Web 2.0 applications present many affordances for constructivist learning. Some of these applications include social networking, blogs, podcasting, media sharing, and virtual worlds. The common trait of these applications is shareability. These applications and similar others allow for global interaction where participants contribute and use others' contributions. The contribution can be as simple as a few paragraphs posted to a blog or as complex as creating avatars on a Second Life (SL) island. The main premise is that these applications represent great potential for learning. In this section, a sample of five Web 2.0 applications is used to illustrate the use of Web 2.0 applications as contexts for constructivist learning. This sample includes social networking and bookmarking, blogs and wikis, podcasts, and media sharing, and virtual worlds

Blogs have been around for several years and we are seeing the impact of their use for positive learning outcomes. Basically, blogs are web-based journals that allow participants to contribute to or create discussion threads. The threads can discuss different subjects separately or concurrently. There are possibly millions of blogs on the web whose subjects include religion, politics, medicine, and car repairs in addition to others. Blogs are characterized by their asynchronous nature and ease of use. However, they also provide opportunities for audio and video files. Blogs can be used for constructivist learning in many ways. First, a blog can serve as an individual learning journal. A learner may start a blog to keep inventory of his/her learning. This inventory will help learners build from one phase of their learning to another. By doing so, learners use previous knowledge to generate new insight. Another example of a blog being used for educational purposes is a class group project. Learners can be engaged with discussion topics and debates in a flexible learning space that allows for the incorporation of videos, graphics, and audio and the enhanced creativity. Blogs are user-generated web-based journals that offer opinions and information and that may include text, images, and links to other blogs and web pages. Some blogs are confined to personal expressions, but others make provision for reactions and comments from readers. In higher education, blogs have been used as a means by which students can collaborate asynchronously. A blog can also serve as a group knowledge builder. A group of learners can join forces in order to complete tasks or projects. Each user is usually assigned a section of the task to complete and contribute to the blog. Through this collaboration, useful content is usually produced as a solution for the task or project. The produced content is presented as the fruition of the learning experience. By collaborating with others, a learner can be benefit by adding to his/her knowledge through the different members' contributions.

A wiki is a collaborative website that allows visitors to add, remove, edit, and change content, typically without the need for registration. It also allows for linking among any number of pages. This ease of interaction and operation makes a wiki an effective tool for mass collaborative authoring. (e.g., Wikipedia.org). Similar to blogs in many aspects, Wikis can serve numerous constructivist learning opportunities such as content production. Wikis are made of many web pages that allow contributors to add and modify content, theirs and others'. The modification opportunities make it possible for all wiki members to share their knowledge in addition to assessing that of other members. By assuming the editor (modifier) role, a wiki contributor is improving his/her writing and critical thinking. Through this practice, new knowledge may be found and added to the learner's previous knowledge. The aggregation of faculty publications in a wiki is described by Conner (2007) as a means to socially construct an online library resource and promote research productivity.

An example of an educational wiki in higher education would be Ask Dr. Wiki (http://www.askdrwiki.com/). This is a medical wiki devoted to creating a free source of medical information. Individuals can publish clinical notes, pearls, X-ray images, angiograms and more on the site. Using this wiki, anyone with a medical background can contribute or edit medical articles.

Media Sharing and Construction

Podcasts and media sharing include video, audio, photos, slides, and others to present other constructivist learning opportunities through producing and sharing digital content. Media sharing is appealing because of its interactive and personal nature. The content can be audio or video files. The main premise here is that learners can generate new knowledge by creating these podcasts and learn from the creation process. Through podcasts, learners can share information and broadcast many useful materials. Learners, then, can use podcasts as tools to learn from and teach others and provide follow-up on previous discourse. Dearstyne (2007) notes the trends in collaboration such as the increase in blogs, mashups, and wikis. Mashups are websites, or other applications, that integrate and aggregate content from more than one source into an integrated application (e.g., combining data on a topic of interest with geographical data such as Google Maps).

Virtual worlds applications such Second Life are interactive environments where participants assume personalities (avatars) to interact with other individuals and objects such as rooms, documents, presentations, and videos. By interacting with the different objects, a learner will gain a bit of knowledge from each object and then connect these different bits from the different objects to form the new experience. Similarly, the virtual world setting provides group collaboration contexts, and consequently, opportunities for constructing content. The application of collaborative and social constructivist-oriented activities, such as virtual spaces and a "user-centric approach" that promotes higher level learning are described by Sommerville and Nino (2007). The social synchronicity achieved through these media allows distant learners to connect in real-time and promote connectedness and community.

The group experience and the community membership provide positive confines that embrace collaboration and individual creativity. While the participants are making products, they will have opportunities to acquire and gain feedback from other community members (Jonassen, 1999). That will provide the different participants with new skills and knowledge as benefits of the feedback. These applications present opportunities where a learner can physically actively engage through the human computer interface (HCI) technology, virtually, and mentally active. An example of how a learner can be physically engaged in a virtual environment (and arguably a web 3.0

phenomenon) is in the work of Mitch Kapor and Philippe Bossut (2008) who have developed a hands-free navigation device for SecondLife. A video of this prototype is available at: http://www.youtube.com/watch?v=2t52gkAwJq8

Consequently, by being mentally active, learners will expand their cognitive abilities to move beyond memorization. Within that practice, long-term critical thinking becomes the norm for these learners. Furthermore, as the learners' critical thinking evolves, a few positive habits such as planning and organization are gained. These habits become second nature for these learners in their potential careers. Web 2.0 activities are similar in those regard. For example, learners who use blogs to track their learning (writing) and to collaborate with others (building camaraderie) will develop better writing and interpersonal skills to help in their respective careers. The main point is that many of these learning experiences are transferrable, and can be potentially helpful in many different contexts.

Perhaps the most important aspect of constructivism is that learners are charged with their learning. That charge will breed a sense of ownership and accordingly, learners become more interested in the process. Similarly, Web 2.0 applications encourage all participants to create their own work that can be seen by millions around the world and have the potential for cross-hemispheric collaboration. Because they own that work and the fact that it is in the public spotlight, those participants become more motivated to do work of high quality. There is an interesting dynamic and digital responsibility that is inherent to content creation on the web. The participants will seek and explore new venues and resources from which to learn to make better products. These results contribute to personal involvement. Furthermore, the experience will generate a long-term memory of the new knowledge and skills that are gained (Doolittle, 1999).

Another significant aspect which relates Web 2.0 applications to constructivist settings is that designing and implementing authentic activities to which the learners can relate can be a great motivation for the learner to participate (Zaulkernan, 2006). That authenticity can pique the learners' interest and curiosity. For example, the use of podcasting and vodcasting (video casting) by music students enhances the learning experience and enables the transmission of content to mobile devices (m-Learning) such as iPods and cell phones.

How will a facilitator use the different Web 2.0 applications as contexts for constructivist learning experiences? This question should be considered in relation to learning objectives, activities, and outcomes.

Implementing Web 2.0 technologies, while establishing learning objectives, is the beginning in creating a constructivist approach to learning. To accommodate for the potentially different levels of motor skills among the learner, the facilitator must include learning the technology as one of the learning objectives of the learning experience. By doing so, the learner will understand the importance of these technologies for completing their tasks. For example, a facilitator may create a list of learning objective as shown below in Table 1 to promote Learning 2.0.

Accordingly, the activities in the same learning experience should be planned to help the participants use the respective Web 2.0 technologies while completing their tasks toward achieving the learning outcomes. In order to apply an activity to achieve Learning Objective 5, a facilitator may require the learners to find a blog and monitor the interaction or to set up an RSS feed to manage the flow of created content. Furthermore, the learners are asked to explore how to participate in the respective blogs they found. Since participation is multi-layered the teacher could also encourage the participants to use other Web 2.0 technologies on the blog such as widgets or avatar creation (meez.com). Assessments of the participating learners should include measure to assess whether they learned and used the respective Web 2.0 technolo-

Table 1. An example of implementing Web 2.0

The learners will be able to
1. Identify what a learning strategy is.
2. Explain what a lesson plan is.
3. Create a lesson plan.
4. Explain what a blog is.
5. Use a blog to post his/her lesson plan.
6. Provide and receive feedback.

gies. Using survey instruments such as Zoomerang or SurveyMonkey can provide feedback to faculty about the level of learning.

WHY ADOPT WEB 2.0 TOOLS?

There are numerous benefits of using Web 2.0 tools: access to social networking tools, knowledge sharing, digital publishing, and cost effectiveness, are just a few. *Computerworld* magazine (2008) reports that Web 2.0 tools such as wikis, blogs, and social networks are being used by many corporations (e.g., Intel Corporation) to solve complex problems, to promote training, and to build an online community. Higher education is also using these emerging tools in creative ways to promote a higher quality of interactivity and collaboration among learners. Palloff and Pratt (2005) discuss the idea that teaching and learning involve more than taking old teaching models of lecture and transferring them to a different medium. They describe an engaged learner who collects, create, and recreates knowledge. The definition of dialogue is enhanced to included Web 2.0 tools such as blogs, synchronous audio and video conferencing. Social constructivist theory emphasizes the negotiation of meaning and construction of shared understandings through dialogue (Jonassen, Davidson, Collins, Campbell, & Haag, 1995; Bonk & Kim, 1998).

There are numerous issues with the quality of online or blended learning courses, including how course objectives are accomplished, and the levels of interaction with content, peers and the instructor (Legon, 2007). *Can Web 2.0 tools increase the Zone of Proximal Development (ZPD)?* Research centered on making the transition to the Web 2.0 world and the use of podcasting audio technologies to promote collaborative learning uncovered some interesting findings. Lee, Mcloughlin and Chan (2008) found that students were empowered by the creation of rich content, as was evident in their own reflections and analysis of constructivist thought. However, Sandars and Schroter (2007) surveyed 3000 medical students and found that while there was a high familiarity and interest with a wide variety of Web 2.0 technologies, there was hesitancy in how to use these technologies. The authors conclude that there needs to be additional training in the instructional approaches for achieving desired educational outcomes through the power of technology.

The increase of online interactivity through these tools offers solutions to communicating ideas collaboratively not only effectively and efficiently, but also with a high degree of creativity. Interactivity is complex and important to the learning process. There is interactivity between the learner and the teacher, the learner and peers, and the learner and the technology. Woo and Reeves (2007) deconstruct the notion of interaction in online learning through the social constructivist lens and show how it can promote meaningful learning. They conclude:

The bottom line is that to increase the learning effects of online interaction, we should, first of all, understand clearly the nature of interaction within the framework of social constructivist learning theory. Once we gain such an in-depth understanding, we should be able to engage in productive research and development to identify the necessary design principles for implementing more effective interaction activities within web-based learning environments. (p.23)

Likewise, according to Jonassen (1995), constructivists use technologies for purposes that are (1) applied to real-world situations, (2) problems and constructs, and (3) authentic and appropriate social context. There are numerous benefits to integrating the Web 2.0 tools as learners are empowered through the creation of content and the many forms of communication. What are some of these Web 2.0 tools and how are they being used in higher education?

WEB 2.0 AND HIGHER EDUCATION

Thompson (2007) reasoned that "Today's institutions of higher education (IHEs) need to consider how they can move from being Education 1.0 institutions because their competitors are revising how they provide services and coursework using Web 2.0 applications" (p.40). However, Eijkman (2008) believes that ". . . the effective educational use of Web 2.0 will in due time lead to a radical reframing of our educational thinking and practices and a redesign of digital learning spaces around interdependent acculturation . . ." (p. 102). From the standpoint of the learner, Karpinski (2008) found that a majority of learners surveyed were unfamiliar with Web 2.0 resources. While students may be familiar with the Web 2.0 tools there was hesitancy on their part to adopt them because of (1) difficulty of use (2) skepticism about the quality of content, and (3) the lack of perceived need. Clearly there needs to be more of an acculturation of Web 2.0 tools so that current and future educators can be more confident in effectively using them in their practice. In some cases there will be a need for an epistemic instructional transformation for faculty who are in various stages of adopting new technologies.

Campus Technology magazine published an article in April 2008 about creative ways that college instructors have integrated these tools into their practice. Examples include the use of a wiki that allows students to submit exam questions for consideration. A portal is a system that has many functions including serving as an online community and a means to catalog or organize different digital creations and provide them to members in an effective manner.

Alexander (2006) describes Web 2.0 tools and how this "heterogeneous mix of relatively familiar and emergent technologies" (p.33) is used in a social networking context. Another similar example is the University of Alaska Anchorage's Web 2.0 wiki which not only defines what a Web 2.0 tool is but how this online learning instrument can be used in an educational setting. Graduate students enrolled in an instructional design course collaborated to create the content on the wiki. Each learner was given editor status so that they could post their content defining and describing the technology but also sharing their creativity by sharing how they used the tool for a positive learning outcome.

In addition to the most popular (blogs and wikis), other Web 2.0 tools include social bookmarking (de.li.ci.ous), collaborative webpage design (jot.com), and (Gnosh.org). Wix.com is a free online tool that helps the web designer with the development of a Flash interface website. This intuitive tool offers "what you see is what you get" (WYSIWYG) design that is dynamic and fluid.

Conferencing software is a tool that allows for the synchronous sharing of ideas. Commercial products include HorizonLive, Elluminate, and Adobe Breeze. WizIQ is a free online conferencing software which provides the interactivity of a whiteboard, webcam, direct messaging, and more, (http://wiziq.com). Understandably, Alexander concludes: “Meanwhile, academic implementations are bubbling” (2006, p.44) with the advent of Web 2.0.

While there have been studies on interactivity and collaborative learning through community building (e.g. Davies, Ramsay, Linfield, & Couperthwaite, 2005; Palloff & Pratt, 2005; Sherry, 2000), Hung (2001) contends there still remains much research on *how* the social interactions that are an inherent Web 2.0 component support the Vygotskian claim that cognition begins at the social level. What does this mean in terms of utilizing Web 2.0 technologies for facilitation of Learning 2.0? What is the educational utility of the Web 2.0 tools?

THE OPPORTUNITIES OF WEB 2.0

The purpose of this section is to convey how Web 2.0 applications represent and provide opportunities for constructivist learning. The capacity to apply and construct new knowledge, to encourage critical thinking, promote collaboration, and to use guided scaffolding to support higher level thinking and learning is at the very center of the constructivist paradigm. It has been argued that higher education needs to move “beyond the individual mind to include learning that is built up by mediated conversations among members of peer groups, local learning communities, and broader cultural systems” (Sherry, 2000, p. 21). Conrad and Donaldson (2004) also take this point of view: “Engaged learning stimulates learners to actively participate in the learning situation, and thus gain the most knowledge from being a member of an online learning community” (p.7). Lave and Wenger (1991) state: “A learning curriculum is not something that can be considered in isolation, manipulated in arbitrary didactic terms, or analyzed apart from the social relations that shape legitimate peripheral participation” (p. 97). But what is meant by “actively participate” and which technologies and approaches support positive learning experiences and promote quality interaction or legitimate peripheral participation?

Social Networking

Social networking Sites (SNS) applications allow participants to belong to “friends” communities where individuals are invited to be friends. Two well-known applications are MySpace and Facebook. On the other hand, social bookmarking applications are used to help users organize information by saving different web sites in one location because the user deems these web sites important for learning and research. Tufekci (2008) did a study on social networking sites (SNS) by college students and draws an important distinction between the “expressive Internet” with an emphasis on the social interactions and networking and the “instrumental” Internet that focuses on e-commerce and news. Users of SNS are also much more frequent users of the expressive Internet.

Global access and democracy can easily intertwine to create new possibilities. For example, what if classrooms in Tasmania were paired with their counterparts in Alaska to study polar science from an Antarctic and Arctic perspective? The learners could collaborate to write and illustrate a report that could involve topics such as global warming, environmental, or ecological science. The learners could communicate through web-based synchronous and asynchronous mediums.

The Internet has made digital space and the ZPD much larger and the world much smaller. Two video game players can enjoy the same game via separate individual web access points without regard to their location or time zone; that is the larger space. These games can be embedded within a blog, webpage, or wikis (See, for example, http://onlinegames.net for the extractable and embeddable code). With the possibility that one of these players may be in the USA and other may be in China, the world reach becomes smaller. The world of Web 2.0 applications such as MySpace (http://www.myspace.com/) and FaceBook is also small; through them, a person may have hundreds, if not thousands, of friends around the world. These and similar social networking applications make global interaction and cross-hemispheric collaboration easier and provide settings for global communities.

Another example of a Web 2.0 tool that promotes interactivity and community is the Ning. A Ning aggregates Web 2.0 tools such as a webpage discussion forum, file sharing, within its multi-functional framework. While MySpace and Facebook have been used for several years, a Ning allows the user to create their own network and not just be a participant in a network that has already been created. It is a means by which private or public groups with a shared interest can construct digital space and interact through a variety of mediums online. The Ning at http://Classroom20.ning.com is an example of constructed digital space for educators who can share how they are integrating technology into their practice. It is a dynamic community that is changing because of the content constructed by participants. Nings have been created for a myriad of groups including those focused on action research, Scholarship of Teaching and Learning (SoTL), computer enthusiasts, history buffs, etc. It takes only a few minutes to construct a Ning and create a viable network that aggregates the virtual social presence of learners, their thoughts and discussions, and digital artifacts. The Ning has enormous potential for providing structured community and while it is in its early stages of development it has been widely adopted by higher education departments, workgroups, ESL classes, significant interest groups, and so on.

Web 2.0 tools support many forms of social networking and collaboration, including those which a more revolutionary sense of emancipatory Freirean education, such as Flash mobs, groups that get together physically to demonstrate or protest using digital technologies (e.g. LiveJournal.com). There are countless possibilities.

Widgets and Plugins

Other tools include widgets or interactive add-ons to supplement the learning experience. These widgets can be embedded into course blogs to achieve a variety of results. There are numerous widgets to help with the learning of language and testing of skills on a variety of topics, from tracking the financial markets, to creating artwork like Matisse. A good example of a widget used in education is to have it deliver course content that the students can embed in their individual web sites. Various widgets include embedded videos, photos, podcasts, maps, all aggregated on one page (pageflakes.com) (*The Chronicle of Higher Education*, 2008).

Web 2.0 tools may fill a specific need or address a certain problem, such as the management of the enormous amount of digital information. For example, a pragmatic plug-in tool for browsers is Real Simple Syndication (RSS). RSS feeds allow the user to aggregate various resources (e.g. blogs, wikis libraries, periodicals, websites) when new content has been added. Another example is Zotero, a free plug-in for anyone who needs assistance when managing reference data from online database searches. It helps organize the citations of research articles and makes the writing experience flow easier because the knowledge can be arranged, classified, and shared. It also facilitates the importing and exporting of information. For example, you can capture citations from webpages, store documents, images, and other indexed citations from large databases. The user has the ability to co-create digital libraries and be notified via Really Simple Syndication (RSS) when a relevant research publication is tagged for retrieval from similar articles. Filters can also be set up for advanced searches. Users can also write annotations that will be saved with each reference.

FUTURE TRENDS

The question "What issues will Web 2.0 address?" was asked earlier in this chapter. The new approaches to learning with digital media have the potential to enhance the higher education learning experience. We are at the nexus of a historic transformation in how learning is shared, transmitted, collected, and constructed. The social web has seen a profound shift in higher education and the building of community through virtual social networking sites. As the facilitation of online learning evolves (artificial intelligence and non-human facilitation, web 3.0?) it becomes necessary to adopt Web 2.0 tools in the design of effective and efficient e-learning.

CONCLUSION

This chapter discussed the promise of using Web 2.0 technologies to extend the classroom. Some of the advantages associated with implementation of Web 2.0 were listed and discussed. The potential for a revolution of learning that involves constructivism will inherently involve these tools. "Constructing an understanding requires that the students have opportunities to articulate their ideas, to test those ideas through experimentation and conversation, and to consider connections between the phenomena that they are examining and other aspects of their lives" (Julyan & Duckworth, 1996, p.58). We can use these Web 2.0 tools to transform our practice and enhance the reality of learning. The dialectic between the individual and society through the use of Web 2.0 tools is expanding dramatically as these new possibilities suggest that learning is a constructive building process of meaning-making. Faculty should adopt these constructivist tools or run the risk of "engage or enrage" (Prensky, 2005). Faculty cannot continue to teach as they were taught but rather they must employ the interactive Web 2.0 tools to engage learners. This is the power and promise of Web 2.0 tools.

REFERENCES

Addison, C. (2006). A new chapter in web development? *Development in Practice*, *16*(6), 623–627. doi:10.1080/09614520600958348

Alexander, B. (2006). A new wave of innovation for teaching and learning. *Educause* (March/April) 33-44. Retrieved April 5, 2008 from http://www.educause.edu/ir/library/pdf/ERM0621.pdf

Alexander, B. (2008). Web 2.0 and emergent multiliteracies. *Theory into Practice*, *47*(2), 150–160. doi:10.1080/00405840801992371

Bonk, C. J., & Kim, K. A. (1998). Extending sociocultural theory to adult learning. In C. M. Smith & T. Pourchot (Eds.), *Adult learning and development: Perspectives from educational psychology* (pp. 67-88). Mahwah, NJ: Lawrence Erlbaum.

Brown, J. S., & Adler, R. P. (2008). Minds on fire: Open education, the long tail, and learning 2.0. *EDUCAUSE Review*, (January/February): 17–32.

Computerworld. (2008). *IT will "take head out of sand" on Web 2.0 in 2008, 42*(6), 5.

Connor, E. (2007). Using Wiki technology to build a faculty publications database. *Journal of Electronic Resources in Medical Libraries*, *4*(4), 11–25. doi:10.1300/J383v04n04_02

Conrad, R. M., & Donaldson, J. A. (2004). *Engaging the online learner: Activities and resources for creative instruction*. San Francisco: Jossey-Bass.

Davies, A., Ramsay, J., & Linfield, H. (2005). Building learning communities: Foundations for good practice. *British Journal of Educational Technology*, *36*(4), 615–628. doi:10.1111/j.1467-8535.2005.00539.x

Dearstyne, B. W. (2007). Blogs, mashups, & wikis. *The Information Management Journal*, *41*(4), 24–33.

Dede, C. (1996). Distance learning: Distributed learning: Making the transformation. *Learning and Leading with Technology*, *23*, 25–30.

Doll, W. (1993). *A post-modern perspective on curriculum.* New York: Teachers College Press.

Doolittle, P. (1999). *Constructivism and online education.* Virginia Polytechnic Institute and State University. Retrieved July 2008 from: edpsychserver.ed.vt.edu/workshops/tohe1999/text/doo2.pdf

Eijkman, H. (2008). Web 2.0 as a non-foundational network-centric learning space. *Campus-Wide Information Systems*, *25*(20), 93–104. doi:10.1108/10650740810866567

Franklin, G. (2000). Augmenting print-based distance learning with other technologies can enhance its effectiveness for adult learners. *Journal of Educational Technology Systems*, *28*(4), 327–334. doi:10.2190/AJ1U-VD37-9J59-PH24

Gagne, R. M., Wager, W. W., Golas, K. C., & Keller, J. M. (2005). *Principles of instructional design*. (5th ed) Belmont, CA: Wadsworth Publishing.

Ge, S., & Tok, M. (2003). Enhancing online education using collaboration solutions. *Journal of Educational Technology Systems*, *31*(4), 361–380. doi:10.2190/BGQM-PFF7-84AD-XCA8

Glogoff, S. (2005). Instructional blogging: Promoting e-learning, student-centered learning, and peer input. *Innovate*, *1* (5). http://www.innovateonline.info/index.php?view=article&id=126

Hung, D. (2001). Design principles for web-based learning: Implications for Vygotskian thought. *Educational Technology*, *41*(3), 33–40.

It's a Web 2.0 world: Are you ready for it? (2003, April). *Campus Technology* (pp. 43-50).

Jonassen, D. (1999). Designing constructivist learning environments. In Reigeluth, C. M. (1999). *Instructional Design Theories and Models: A New Paradigm of Instructional Theory, 2*, 215-241. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Jonassen, D., Davidson, M., Collins, M., Campbell, J., & Haag, B. B. (1995). Constructivism and computer-mediated communication in distance education. *American Journal of Distance Education*, *9*(2), 7–26.

Julyan, C., & Duckworth, E. (1996). A constructivist perspective on teaching and learning science. In C. Twomey-Fosnot (Ed.), *Constructivism: Theory, perspectives, and practice*. (pp. 55-72). Columbia University. Teachers College Press.

Karpinski, J. L. (2008). Disconnected in a connected world: Knowledge and understanding of Web 2.0 tools at the University of Pennsylvania Medical Center. *Medical Reference Services Quarterly*, *27*(1), 53–72.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, MA: Cambridge University Press.

Lee, M. J., Mcloughlin, C., & Chan, A. (2008). Talk the talk: Learner-generated podcasts as catalysts for knowledge creation. *British Journal of Educational Technology*, *39*(3), 501–521. doi:10.1111/j.1467-8535.2007.00746.x

Legon, R. (2007). Online courses lead the way in quality assurance. *Sloan-C View: Perspectives in Quality Online Education 6*(1). Retrieved July 20, 2007 from http://www.sloan-c.org/publications/view/v6n1/viewv6n1.htm

McFedries, P. (2006). The web, take two. *IEEE Spectrum*, *43*(6), 68–68. doi:10.1109/MSPEC.2006.1638049

Palloff, R., & Pratt, K. (2005). *Collaborating online: Learning together in community*. San Francisco: Jossey-Bass.

Prensky, M. (2006). Engage me or enrage me: What today's learner's demand. *Educause* October, 60-64. Retrieved September 17, 2008 from http://net.educause.edu/ir/library/pdf/erm0553.pdf

Sandars, J., & Schroter, S. (2007). Web 2.0 technologies for undergraduate and postgraduate medical education: an online survey. *Postgraduate Medical Journal*, *83*(986), 759–762. doi:10.1136/pgmj.2007.063123

Schrum, L., & Hong, S. (2002). Dimensions and strategies for online success: voices from experienced educators. *JALN*, *6(1)*. Retrieved August 11, 2002 from http://www.aln.org/alnWeb/journal/jaln-vol6issue1.htm#schrum

Sherry, L. (2000). The nature and purpose of online conversations: A brief synthesis of current research. *International Journal of Educational Telecommunications*, *6*(1), 19–52.

Somerville, M. M., & Nino, M. (2007). Collaborative co-design: A user-centric approach for advancement of organizational *learning*. *Performance Measurement & Metrics*, *8*(3), 180–188. doi:10.1108/14678040710841054

The Chronicle of Higher Education. (2008, September 11). *Professor uses web 'Widgets' to share course content*. Retrieved October 10, 2008, from http://chronicle.com/wiredcampus/article/3307/professor-uses-web-widgets-to-share-course-content.

Thompson, (2007). Web 2.0 *takes on colleges and universities: The dawn of education 2.0*. Retrieved April 29, 2008 from http://www.masternewmedia.org/news/2007/04/20/Web_20_takes_on_colleges.htm

Tufekci, Z. (2008). Grooming, gossip, Facebook and Myspace. *Information Communication and Society*, *11*(4), 544–564. doi:10.1080/13691180801999050

Twomey-Fosnot, C. (1996). Constructivism: A psychological theory of learning. (pp. 8-33). In C. Twomey-Fosnot (Ed.), *Constructivism: Theory, perspectives, and practice.* New York: Teachers College Press.

Von Glasersfeld, E. (1996). Introduction: Aspects of Constructivism. In C. Twomey-Fosnot (Ed.). *Constructivism: Theory, perspectives and practice* (pp. 3-7). Teachers College, Columbia University.

Woo, Y., & Reeves, T. C. (2007). Meaningful interaction in web-based learning: A social *constructivist* interpretation. *The Internet and Higher Education*, *10*(1), 15–25. doi:10.1016/j.iheduc.2006.10.005

Zualkernan, I. A. (2006). A framework and a methodology for developing environments. *Educational Technology & Society*, *9*(2), 198–212.

APPENDIX

Tool	Example	Link
Open Source Tools	Open Office Linux The Open Source Initiative Source Forge	http://Openoffice.org http://www.linux.org http://www.opensource.org/ http://sourceforge.net/
Open CourseWare	Massachusetts Institute of Tech (MIT) Carnegie-Melon University Utah State University	http://ocw.mit.edu http://www.cmu.edu/oli/ http://ocw.usu.edu/
Portals	Merlot (Multimedia Educational Resources for Learning and Online Teaching) UPortal	http://merlot.org http://www.uportal.org/
Collaborative design & Project Management	Jot Gnosh Project Zoho	http://jot.com http://gnosh.com http://projects.zoho.com/jsp/home.jsp
Survey Instruments	Zoomerang Survey Monkey	http://zoomerang.com http://surveymonkey.com
Video Blogs	Video Blogs	http://www.vblog.com/
Directory of Web 2.0 Tools	http://www.go2Web20.net/ Centre for Learning & Performance Technologies	http://www.allthingsWeb2.com/ http://c4lpt.co.uk/handbook/elearning20.html
Social Networks	Ning.com	http://education.ning.com http://www.classroom20.com/
Blogs	Blogger.com	http://asdtechnology.blogspot.com
Wikis	PBwiki.com wikispaces.com	http://pbwiki.com http://wikispaces.com
Video Blogs	Vobbo	http://www.vobbo.com/
Webpage design	Weebly FreeWebs Google websites	http://Weebly.com http://webs.com http://sites.google.com/
Conferencing Software	HorizonLive Elluminate Live! Breeze, WizIQ;	http://Elluminate.com www.wimba.com http://www.adobe.com/resources/breeze/ http://www.wiziq.com
Social bookmarking	De.li.ci.ous	http://delicious.com/
Widgets	Widget Box	http://www.widgetbox.com/tag/education http://pageflakes.com
Online Answers	One Big U	http://www.onebigu.com/user/homepage.php
Games	Online Games	http://onlinegames.net

This work was previously published in Information Technology and Constructivism in Higher Education: Progressive Learning Frameworks, edited by C. R. Payne, pp. 146-158, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 1.13
Integrating Web 2.0 Technologies within the Enterprise

R. Todd Stephens
AT&T Corporation Collaboration & Online Services, USA

ABSTRACT

This chapter examines the elements of the new Web 2.0 technology base and reviews the lessons learned when implementing these technologies. Collaborative applications have made enormous inroads into the enterprise and bring unprecedented speed and transparency to communications. Researchers and practitioners alike are focusing on how collaborative applications can replace the one-way communications inherent to Intranet sites. This chapter is intended for individuals who are looking toward the possibility of integrating these new technologies into the core communication medium. Unfortunately, there are still large barriers such as politics, turf battles, integration, and poor usability with the current product set. A company's ability to manage information effectively over its life cycle, including sensing, collecting, organizing, processing, and maintaining information, is crucial to the long term success in a global economy. The success or failure of this integration may very well create or lose a competitive advantage for the enterprise. What is missing is a framework or roadmap in which organizations can plan out their execution of We 2.0 deployment.

DOI: 10.4018/978-1-60566-106-3.ch037

INTRODUCTION

Enterprises are being transformed from an old business model built around the command and control aspects information management to a new one where collaboration is the essential component. We are witnessing this transformation outside the enterprise with the success of Web 2.0 technologies like Wikipedia, YouTube, and Flickr. Yet, within the walls of the organization, progress is slow. This chapter will examine why the adoption rates for Collaborative and Social Software are low and what can be done to improve them. This research emerges from three Fortune 500 companies which the resulting methodology has worked to increase the adoption rates ten fold.

Traditional implementation methodologies focused on the hardware, software, and the associated functionality. Our research indicates that while these are important they do not lead to mass adoption of the application by the enterprise. Many functions that information workers perform are dictated by the business and current transactional systems like CRM, ERP, or accounting systems. Collaboration and Social Software need to be integrated as situational applications and although they are optional, they are required to reach higher levels of performance. Members of complex teams are less likely to utilize collaborative tools in the absence of other influences such as executive encouragement, modeling collaborative behavior, creating a “gift” culture, training, supporting a strong sense of culture, assigning team leaders, building, and understanding roles (Erickson & Gratton, 2007). Our research indicates barriers to adoption including social issues, cultural issues, awareness issues, educational issues, and political issues. This research will focus on the awareness and educational issues since most organizations implementing Web 2.0 will face these first issues first.

BACKGROUND

Web 1.0 Intranets

The term Web 1.0 emerged from the research around Web 2.0. Basically, Web 1.0 focused on a read only Web interface while Web 2.0 focuses on a read-write interface where value emerges from the contribution of a large volume of users. The Internet as well as the Intranet initially focused on the command and control of the information itself. Information was controlled by a relative small number of resources but distributed to a large number which spawned the massive growth of the Web itself. Like television, the Web allowed for the broadcasting of information to a large number of users.

Inside the organization, the Intranet has changed the way organizations structure and operate their business. Specifically, the Intranet has centralized communications and corporate information as well as built a sense of community across organizational boundaries (McNay, 2000). Typical organizations will have office-based employees in various locations, telecommuting, and off-shoring staff. The traditional day by day communication landscape has changed from personal to electronic. The migration to electronic communications emerged as standards, technology and infrastructure matured. This allowed more information sharing and community building to occur without a requirement of physical location. Over the past several years Intranets have emerged as the key delivery mechanism for application and business information. Intranets may be thought of as providing the infrastructure for intra-organizational electronic commerce (Chellappa & Gupta, 2002). This allows organizations to utilize the technology to achieve its organizational goals and objectives. Web 1.0 allowed the organization to govern the information flow and focus on achieving the business goals.

Unfortunately, most technologies fail to deliver competitive advantages over an extended period of time. Investments in information technology, while profoundly important, are less and less likely to deliver a competitive edge to an individual company (Carr, 2003). This is especially true in the world of the Web 1.0 since much of the knowledge and information is disseminated all over the world as quickly as it gets published. Organizations are beginning to see that the command and control model is no longer effective at developing a high performance work force which opens the door for the next evolution in technologies as described by the Web 2.0 framework.

Web 2.0 Defined

While Web 2.0 has been debated by researchers as to who and when the concepts emerged, little argument exists that the technology and demand has arrived. Unlike Web 1.0, this new technology encourages user participation and derives its greatest value when large communities contribute content. User generated metadata, information, and designs enable a much richer environment where the value is generated by the volume of employees. Sometimes referred to as sharing, collaboration, aggregate knowledge, or community driven content, social software creates the foundation of collective intelligence (Weiss, 2005). Much of the Web 2.0 technology is difficult to nail down with an exact definition; the basic truth is that Web 2.0 emphasizes employee interaction, community, and openness (Millard & Ross, 2006). Along with these characteristics, Smith and Valdes (2005) added simple and lightweight technologies and decentralized processing to the mix. O'Reilly (2005) defined Web 2.0 as a platform, spanning all connected devices; Web 2.0 applications are those that make the most of the intrinsic advantages of that platform: delivering software as a continually-updated service that gets better the more people use it, consuming and remixing data from multiple sources, including individual users, while providing their own data and services in a form that allows remixing by others, creating network effects through an "architecture of participation," and going beyond the page metaphor of Web 1.0 to deliver rich user experiences. While Web 2.0 has many and often confusing definitions most include the concepts of Weblogs, Wikis, Really Simple Syndication (RSS) Functionality, social tagging, mashups, and user defined content.

Weblogs or Blogs

Weblogs or blogs have become so ubiquitous that many people use the term synonymous for a "personal Web site" (Blood, 2004). Unlike traditional Hypertext Markup Language (HTML) Web pages, blogs offer the ability for the non-programmer to communicate on a regular basis. Traditional HTML style pages required knowledge of style, coding, and design in order to publish content that was basically read only from the consumer's point of view. Weblogs remove much of the constraints by providing a standard user interface that does not require customization. Weblogs originally emerged as a repository for linking but soon evolved to the ability to publish content and allow readers to become content providers. The essence of a blog can be defined by the format which includes small chunks of content referred to as posts, date stamped, reverse chronological order, and content expanded to include links, text, and images (Baoill, 2004). The biggest advancement made with Weblogs is the permanence of the content which has a unique Universal Resource Locator (URL). This allows the content to be posted and along with the comments to define a permanent record of information. This is critical in that having a collaborative record that can be indexed by search engines will increase the utility and spread the information to a larger audience. With the advent of software like Wordpress and Typepad, along with blog service companies like blogger.com, the weblog is fast becoming the communication medium of the new Web.

Sample Weblog URLs

- Andrew McAfee Web 2.0 Blog (http://blog.hbs.edu/faculty/amcafee/)
- Randy Basler's Boeing Blog (http://boeingblogs.com/randy/)
- Jonathan Schwartz's Sun Blog (http://blogs.sun.com/jonathan/)
- Rough Type by Nicholas Carr (http://www.roughtype.com)

Wikis

A Wiki is a Web site that promotes the collaborative creation of content. Wiki pages can be edited by anyone at anytime. Informational content can be created and easily organized within the wiki environment and then reorganized as required (O'Neill, 2005). Wikis currently are in high demand in a large variety of fields, due to their simplicity and flexibility nature. Documentation, reporting, project management, online glossaries, and dictionaries, discussion groups, or general information applications are just a few a examples of where the end user can provide value (Reinhold, 2006). The major difference between a wiki and blog is that the wiki user can alter the original content while the blog user can only add information in the form of comments. While stating that anyone can alter content, some large scale wiki environments have extensive role definitions which define who can perform functions of update, restore, delete, and creation. Wikipedia, like many wiki type projects, have readers, editors, administrators, patrollers, policy makers, subject matter experts, content maintainers, software developers, and system operators (Riehle, 2006), all of which create an environment open to sharing information and knowledge to a large group of users.

Sample Wiki URLs

- Disney's Parent Wiki (http://family.go.com/parentpedia)
- Wikipedia (http://www.wikipedia.org/)
- Reuters Financial Glossary (http://glossary.reuters.com/)
- Internet 2 (https://wiki.internet2.edu/confluence/dashboard.action)

RSS Technologies

Originally developed by Netscape, RSS was intended to publish news type information based upon a subscription framework (Lerner, 2004). Many Internet users have experienced the frustration of searching Internet sites for hours at a time to find relevant information. RSS is an XML based content-syndication protocol that allows Web sites to share information as well as aggregate information based upon the users needs (Cold, 2006). In the simplest form, RSS shares the metadata about the content without actually delivering the entire information source. An author might publish the title, description, publish date, and copyrights to anyone that subscribes to the feed. The end user is required to have an application called an aggregator in order to receive the information. By having the RSS aggregator application, end users are not required to visit each site in order to obtain information. From an end user perspective, the RSS technology changes the communication method from a search and discover to a notification model. Users can locate content that is pertinent to their job and subscribe to the communication.

Sample RSS URLs

- Newsgator (http://www.newsgator.com/)
- FeedBurner (http://www.feedburner.com/)
- Pluck (http://www.pluck.com/)
- Blog Lines (http://www.bloglines.com/)

Social Tagging

Social tagging describes the collaborative activity of marking shared online content with keywords or tags as a way to organize content for future navigation, filtering, or search (Gibson, Teasley, & Yew, 2006). Traditional information architecture utilized a central taxonomy or classification scheme in order to place information into specific pre-defined bucket or category. The assumption was that trained librarians understood more about information content and context than the average user. While this might have been true for the local library with the utilization of the Dewey Decimal system, the enormous amount of content on the Internet makes this type of system un-manageable. Tagging offers a number of benefits to the end user community. Perhaps the most important feature to the individual is able to bookmark the information in a way that is easier for them to recall at a later date. The benefit of this ability on a personal basis is obvious but what about the impact to the community at large. The idea of social tagging is allowing multiple users to tag content in a way that makes sense to them; by combining these tags, users create an environment where the opinions of the majority define the appropriateness of the tags themselves. The act of creating a collection of popular tags is referred to as a folksonomy which is defined as a folk taxonomy of important and emerging content within the user community (Ahn, Davis, Fake, Fox, Furnas, Golder, Marlow, Naaman, & Schachter, 2006). The vocabulary problem is defined by the fact that different users define content in different ways. The disagreement can lead to missed information or inefficient user interactions (Boyd, Davis, Marlow, & Naaman, 2006). One of the best examples of social tagging is Flickr which allows user to upload images and "tag" them with appropriate metadata keywords. Other users, who view your images, can also tag them with their concept of appropriate keywords. After a critical mass has been reached, the resulting tag collection will identify images correctly and without bias.

Sample Social Tagging URLs

- Flickr (http://www.flickr.com/)
- YouTube (http://www.youtube.com/)
- Del.icio.us (http://del.icio.us/)
- Technorati (http://technorati.com/)

Mashups: Integrating Information

The final Web 2.0 technology describes the efforts around information integration or sometimes referred to as "mashups." These applications can be combined to deliver additional value that the individual parts could not deliver on their own. One example is HousingMaps.com that combines the Google mapping application with a real estate listing service on Craiglists.com (Jhingran, 2006). Other examples include Chicagocrime.org who overlays local crime statistics onto Google Maps so end users can see what crimes were committed recently in the neighborhood. Another site synchronizes Yahoo! Inc.'s real-time traffic data with Google Maps. Much of the work with Web services will enable greater extensions of mashups and combine many different businesses and business models. Organizations, like Amazon and Microsoft are embracing the mash-up movement by offering developers easier access to their data and services. Moreover, they're programming their services so that more computing tasks, such as displaying maps onscreen, get done on the users' Personal Computers rather than on their far-flung servers (Hof, 2005)

Sample Mashup URLs

- Housing Maps: (http://www.housingmaps.com/)
- Chicago Crime (http://www.chicagocrime.org)
- Healthcare Product (http://www.vimo.com/)
- Global Disease Map (http://healthmap.org/)

User Contributed Content

One of the basic themes of Web 2.0 is user contributed information. The value derived from the contributed content comes not from a subject matter expert, but rather from individuals whose small contributions add up. One example of user contributed content is the product review systems like Amazon.com and reputation systems used with ebay.com. A common practice of online merchants is to enable their customers to review or to express opinions on the products they have purchased (Hu & Liu, 2004). Online reviews are a major source of information for consumers and demonstrated enormous implications for a wide range of management activities, such as brand building, customer acquisition and retention, product development, and quality assurance (Hu, Pavlou, & Zhang, 2006). A person's reputation is a valuable piece of information that can be used when deciding whether or not to interact or do business with. A reputation system is a bi-directional medium where buyers post feedback on sellers and vice versa. For example, eBay buyers voluntarily comment on the quality of service, their satisfaction with the item traded, and promptness of shipping. Sellers comment about the prompt payment from buyers, or respond to comments left by the buyer (Christodorescu, Ganapathy, Giffin, Kruger, Rubin, & Wang, 2005). Reputation systems may be categorized in three basic types: ranking, rating, and collaborative. Ranking systems use quantifiable measures of users' behavior to generate and rating. Rating systems use explicit evaluations given by users in order to define a measure of interest or trust. Finally, collaborative filtering systems determine the level of relationship between the two individuals before placing a weight on the information. For example, if a user has reviewed similar items in the past then the relevancy of a new rating will be higher (Davis, Farnham, & Jensen, 2002).

SAMPLE USER CONTRIBUTED CONTENT URLS

- Amazon.com (http://www.amazon.com)
- Ebay (http://www.ebay.com)
- Trip Advisor (http://www.tripadvisor.com/)
- Review Centre (http://www.reviewcentre.com/)

Web 1.0 Compared to Web 2.0

While the differences between Web 1.0 and 2.0 are grey at best, we can attempt to draw some segmentation by reviewing the high level characteristics. Table 1 provides a side by side comparison of these technologies.

In the Web 1.0 environment, information was largely static and controlled by a few resources. Specifically, the individual or organization that produced this information pushed information to the end user by either controlling the access or limiting the feedback options. Web 2.0 turns that model around and create a far greater dynamic environment where each consumer has the ability to contribute to the overall value of the information itself. Instead of searching and browsing topics, Web 2.0 users are allowed to publish and subscribe to the content which results is a more bottom up implementation. The following section will review how these new technologies can be integrated into the current knowledge environments that have traditionally followed the command and control model of information.

Enterprise 2.0

Enterprise 2.0 is a term used to describe the integration of the Web 2.0 technology portfolio inside of the organization. Both the producers and

Table 1. Characteristics of Web 1.0 and Web 2.0

Web 1.0 Characteristics	Web 2.0 Characteristics
Static Content	Dynamic Content
Producer Based Information	Participatory Based Information
Messages Pushed to Consumer	Messages Pulled by Consumer
Institutional Control	Individual Enabled
Top Down Implementation	Bottom Up Implementation
Users Search and Browse	Users Publish and Subscribe
Transactional Based Interactions	Relationship Based Interactions
Goal of Mass Adoption	Goal of Niche Adoption
Taxonomy	Folksonomy

consumers of the information will reside inside the organization. If either of the customer classifications involve outside entities then the Web 2.0 tag should be used. McAfee (2006) indicates a new wave of business communication tools which allow for more spontaneous, knowledge-based collaboration. These new tools, the author contends, may well supplant other communication and knowledge management systems with their superior ability to capture tacit knowledge, best practices and relevant experiences from throughout a company and make them readily available to more users. For all its appeal to the young and the wired, Web 2.0 may end up making its greatest impact in business. And that could usher in more changes in corporations, already in the throes of such tech-driven transformations as globalization and outsourcing. Indeed, what some are calling Enterprise 2.0 could flatten a raft of organizational boundaries; between managers and employees and between the company and its partners and customers (Hof, 2005).

Barriers to Adoption

This chapter does not address issues around infrastructure or software selection. The research wanted to look at the barriers to adoption assuming all other variables are constant and normally taken into account on most implementations. The following barriers of adoption are not related to the specific technology. Rather, they focus on the end user and the major issues impacted them. This makes sense in Web 2.0 the end user contributes as much to the success of the implementation as any other component.

Awareness Issues

The awareness issue describes an environment where of the majority of users have never heard of Web 2.0, Enterprise 2.0, Collaboration, and Social Software. More importantly, end users have not heard of the internal product offering, if one exists. Communication is one of the most critical aspects of letting people know that a collaborative or social application is available. Traditional information technology solutions were focused on a single business process and the aspects of marketing and branding were unnecessary. However, for enterprise services this awareness can be one of the most critical functions performed early in the product's life cycle. Like e-mail and desktop Office applications, you want a high degree of awareness across the entire enterprise. A high degree of awareness would be some where between 90-100 percent of the information workers within the organization.

Educational Issues

End users may have heard of Web 2.0 through the media but they still not understand how the technology can be used in a business setting. Once an end user becomes aware of an application, the next phase is to ensure that they understand how the application should be used. The educational area is critical since most employees above the age of 35 have not used these new types of technologies.

Cultural and Social Issues

When organizations have overcome the awareness and educational gaps, then we can make the statement that the majority of the organization knows the technology is available and what can be done with it. They may still choose to use their older technology that has been used in the past. Not with standing political pressure, we are looking at cultural or social issues. These issues can emerge when end users fear change, afraid of new methods, or prefer to work in a command and control model.

Political Issues

The final area focuses on the political pressure organizations place on users. Political pressure may focus around strategic direction, vendor associations, or organizational structures. In smaller organizations, these issues may not be as big an impact as in a large distributed environment.

INTEGRATION OF WEB 2.0 TECHNOLOGIES

The studied organization is a Fortune 500 telecommunications company that has gone through several acquisitions over the past few years. With the integration of three companies, the presented framework went through several different field trials over the course of three years. This allowed the research to apply the framework into three different companies in order to test the validity in a real business environment. The framework was developed after five years of trials in traditional knowledge management systems. The application of the framework to collaborative tools started in 2004 and progressed through 2007. The initial deployment focused on Microsoft's Sharepoint which is a collaborative tool that has most of the Web 2.0 elements described in the prior section. (See Figure 1.)

Common Situation

While all three implementations varied by size of the firm, number of employees, and basic infrastructure, the implementations had one common characteristic. Flat line growth occurred within six months in each of the deployments. Flat line growth occurs when new orders show no growth over a three to six month period. Figure 2 provides the different site metrics collected prior to the implementation of the proposed framework. The lines have been cut off to indicate the point in time the framework was applied to the organization.

In all three cases, the program had a solid beginning but reached a level of saturation between 90 and 120 collaborative sites. For clarity, the number of collaborative sites continued to grow but could not outpace the same number of deletions. A deletion occurs when a program, project, or resource no longer needs the collaborative or social software environment. On average, the leveling off of site demand occurred between five and six months.

Figure 1. Adoption rates prior to implementing the framework

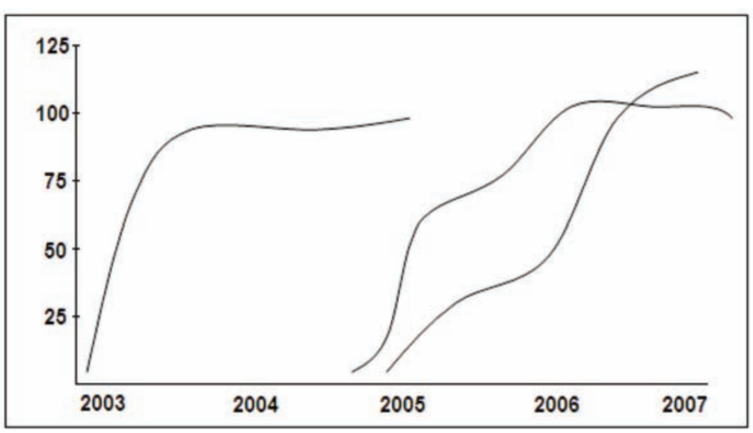
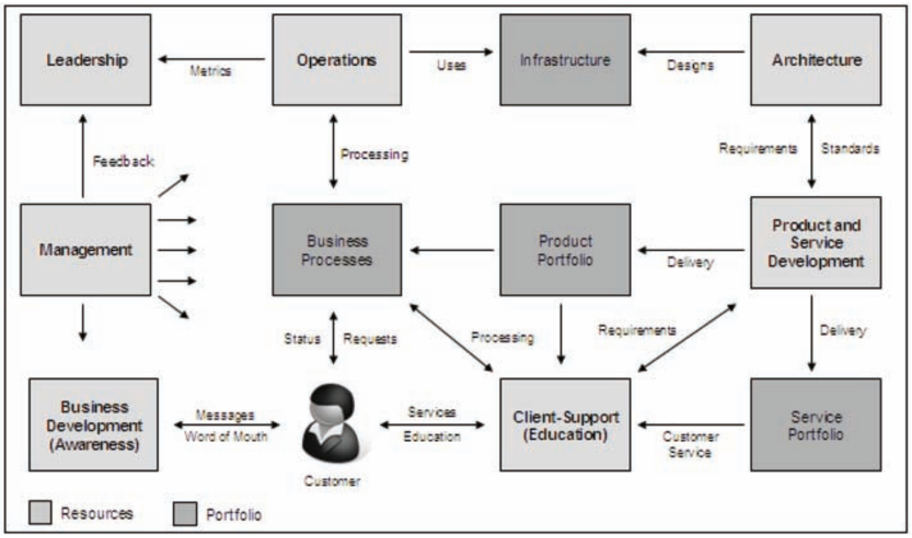


Figure 2. Implementation business model



Business Model Framework

At the highest level, a business model is how an organization creates value. Timmers (1998) define a business in respect to the architecture for the product, service, and information flows, the benefits for the various actors, and the sources of revenue. In reality, an organization can have a variety of business models, each is simply an artificial representation of reality which detracts focus from certain aspects while concentrating on others (Kittl, Petrovic, & Teksten, 2001). This research uses the concepts of a business model, not in terms of commerce, but focused on the various interacting parts required to deliver business value. Figure 2 provides an overview of the business model used to describe functional designation of work. Some researchers refer to this as the resource model.

The model describes seven functional areas and four portfolio or product areas. Leadership and management are two functional areas that will not be covered in this chapter to any depth. The basic idea is that all information technology activities would include coordination, communication, and cooperation activities which are usually performed by these two groups.

Starting at the top of Figure 2, traditional implementations of information technology would include the operational and architecture functions. Operations would include activities such as hardware monitoring, software installation, backup, recovery, security, and maintenance. Computer Operations is a critical function to ensure delivery of a reliable, scalable, and functional infrastructure. This area must be governed with a high degree of control in order to maintain the stability of the environment. Architecture focuses on the design, planning, and software selection within the enterprise. Generally speaking, architecture includes the activities of defining and modeling the environment which may include the following architectures: business, application, data, information, technology, and product architecture (Pereira & Sousa, 2004). Traditionally, these components focused on ensuring that the environment did not fail from an infrastructure point of view. The vast majority of Web 2.0 implementations will focus on these core elements to ensure that the program operates effectively. The idea of an implementation failing would indicate the hardware or software failed to deliver the business value consistently over time. The problem with this approach is that having a perfect infrastructure does not guarantee mass adoption which is the truest measure of success. In the case of internally developed Web 2.0 applications or the utilization of open source, the architecture area could be expanded to encompass the entire System Development Lifecycle (SDLC).

We can define success (mass adoption) from two perspectives. In any knowledge type of application, you will have two key customer classes: the producer of the information and consumer of the information. The producer is the person, community, or application that creates a reusable asset in the form of information utilizing the Web 2.0 tools. This might include a wiki page or ownership to a specific weblog. The consumer is responsible for locating and accessing the information, assessing the ability to reuse the information, adapting to the information and integrating the information into the business. The consumer might not actually contribute to the environment in the form of comments or informational update. Production rates for Web 2.0 applications are still relatively low as compared to the number of users that consume the information. In one survey, only 11 percent of respondents would even consider contributing to Wikipedia while the actual number of contributors is less than 1 percent overall. When you have millions of consumers, a 1 percent contributor rate is pretty good. However, in an enterprise of 20,000 people that would indicate you will only have 20 contributors. This demonstrates the criticality of building up a producer community towards the long term goal of mass adoption. Not only do we need to focus on the information contained within the environment but also with the utilization of that information. The content must be used and to a greater degree, the utilization of the content drives the return on investment. This is not a trivial point; organizations must focus on the components of success and understand that having great hardware, software and functionality is simply the price of entry into the Web 2.0 environment.

Business Development

Located at the bottom right side of Figure 2 is the representation of the customer which could either be the producer or a consumer of the information. The overall framework centers around the customer behavior. The box on the left represents the functions of business development which we can define as any activity that impacts awareness and education before someone becomes a customer. Like the business functions of marketing, branding, and selling, business development strives to encourage participation. The ultimate goal is to ensure that every person in the enterprise knows these tools exist, understand how they can be used, and knows where to go to get engaged.

Client-Support

The vast majority of information workers are not familiar with virtual solutions and need guidance on how to best utilize and integrate this technology into their day to day operations. End users will ask five basic questions of the collaborative environment:

- What collaborative products and services are available to me?
- How can I utilize these products and services within my environment?
- Who can help me in case I need some professional guidance?
- Are the collaborative applications ready for enterprise usage?
- How am I doing in comparison to others or against best practices?

In order to address these questions, organizations should look toward developing a support group that can enable the end user rather than hindering their understanding of a collaborative environment. Meeting the needs of the customer may vary depending on the level of knowledge the user brings to the environment. Customers who are new to technology expect a high level of reliability and support in order to gain the greatest value possible (Johnston & Supra, 1997). Customer service should not be homogeneous and both the online and physical support environments need to take into account the experience level of the end user (Dutta & Roy, 2006). The customer wants to know what products, services, and documentation are available to them within the collaborative environment. The content of an online environment is not limited to the product or services provided. Rather, content includes the solutions and strategies employed to make it easy for the user to accomplish important tasks, such as information retrieval, search, and obtaining feedback (Calongne, 2001). Support information or content should include the product and service quantity, quality, and relevance to the customer (Palmer, 2002). Technologists often make the mistake of assuming a certain level of expertise with the user community. Unlike e-mail or Office products (Word Processing or Spreadsheet), collaborative tools are fairly unknown to the end user. A Client-Support environment would include many of the following components:

- Training and Education
- Subject Matter Expert Information
- Technical Supports and Operational Information
- Best Practices and “How To” Documentation
- Community of Practice
- Ordering Processes
- Product, Service and Solution Overviews
- End User Metrics of Content and Usage

Product and Service Development

Normally when you discuss the concepts around a product, you think of products like Microsoft’s Sharepoint, Confluence, Social Text, or IBM’s Connections. These could be considered products from an architecture or operations point of view. However, from the customer perspective these are tools. A product solves a problem or generates value based on the consumption or utilization. In other words, the product is what you can do with the tool. Products in the Web 2.0 space include weblogs, wikis, collaborative intranets, virtual workspaces, RSS feed readers, book marking, and professional profiles. Each of these can be categorized as a product regardless of the tool selected to perform the function. Services would include both tangible and intangible value-add activities that go along with the products. Services might include templates, user guides, editing, PDF conversion, education, and training. Services must be delivered to the customer and cannot be inventoried for later use. Taken together, products and services provide the customer experience that encourages participation and end user involvement.

Portfolios

Not counting the infrastructure, the model identifies three additional portfolios including the product portfolio, the service portfolio, and the business processes. The product portfolio would include the various products developed in the prior section. Since most products are meta-physical in nature, they must be demonstrated in the online support environment. The service portfolio will describe the services available to the end user. The idea is that products must be demonstrated while services must be described. Finally, the business process portfolio defines the business processes required to engage in the environment. Taken together, a customer may engage with several products, services and business solutions which constitute a solution offering. Assuming the ultimate goal of any Web 2.0 application is the mass adoption of the customer base then you want to move up the value chain. The value-chain has been well documented by Joseph Pine II and James H. Gilmore.

Pine and Gilmore (1997) discuss "the experience economy" by tracing the value added to the coffee bean in its various iterations from pure "commodity" to pure "experience." In their evolutionary construct there are four stages, in ascending order of sophistication the stages are commodity, good, service, experience. They point out that coffee is traded on the futures market at roughly $1 a pound (thus, about 2 cents a cup at the "commodity" level). After manufacturers roast, grind, package and distribute the bean for retail, the price jumps to between 5 and 25 cents a cup (the "goods" level). At a "run-of-the-mill" diner a cup might run from 50 cents to $1 a cup (the traditional "service" level). The authors contend that one can "Serve that same coffee in a five-star restaurant or espresso bar, where the ordering, creation, and consumption of the cup embodies a heightened ambience or sense of theatre, and consumers gladly pay anywhere from $2 to $5 for each cup." Thus, by creating value at the "experience" level, the seller is able to charge an extremely high premium over that charged by the "service" provider. In defining their terms they argue that, "When a person buys a service, he purchases a set of intangible activities carried out on his behalf. But when he buys an experience, he pays to spend time enjoying a series of memorable events that a company stages, as in a theatrical play, to engage him in a personal way." The idea is that organizations that support the Web 2.0 implementation must move up the value chain in order to obtain the mass adoption

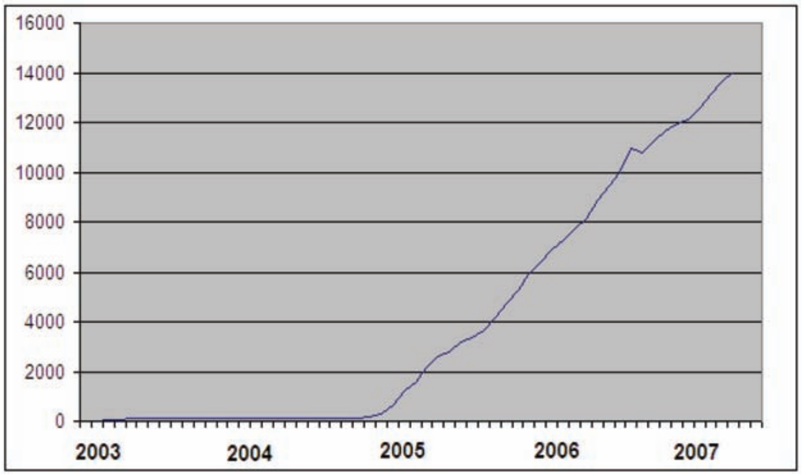
Field Trials

The first implementation of the framework occurred in 2004. As seen in Figure 1, the deployment had already reached the point of saturation when executive leadership contacted the author's team to see if the framework could be applied to the Web 2.0 area. In the past, the framework had been successfully implemented in knowledge management type implementations. These systems included prior research in metadata repositories, registries, and other knowledge applications. Focusing on the left most line (Figure 1), demand had leveled off at around 100 collaborative environments with an average monthly variance of +/-5 percent. The initial review and prioritization of activities focused on providing the online support environment, automating the procurement process, and developing marketing plans. These were seen as obvious gaps in the prior implementation model based on the author's observations. The author was also able contact various user communities in order to ascertain if these gap assumptions were true. No official survey was used to collect the information other than informal conversations. Figure 3 provides the results of implementing the framework over a 36 month period of time.

By the fall of 2007, the studied organization had over 13,000 collaborative sites with an average monthly growth rate of 423.62 percent. As the implementation matured, less focus was placed on the business opportunity area and more on the client-support area. This would make sense, since the number of employees that utilizing the collaborative environment was around 95 percent of the employee population. That is to say that 95 percent of the employees had heard of and used some form of the application which was determined by the unique user id logged into the system.

In 2006, the studied organization was purchased by a larger telecommunications company.

Figure 3. Collaborative sites after the framework application



This purchase brought together three different companies, all of which had an implementation of the collaborative suite. The adoption rates were similar to those found in Figure 1, represented by the two lines located on the right side of the chart. Executive leadership reviewed the implementations and determined that the framework needed to be applied in the other two companies as well. The initial step would be to survey the new organizations to see where the issues lie in adoption and determine which parts of the framework should be emphasized. Unlike the first field trial, the author had no insight into the new companies prior to the integration.

Survey Tool

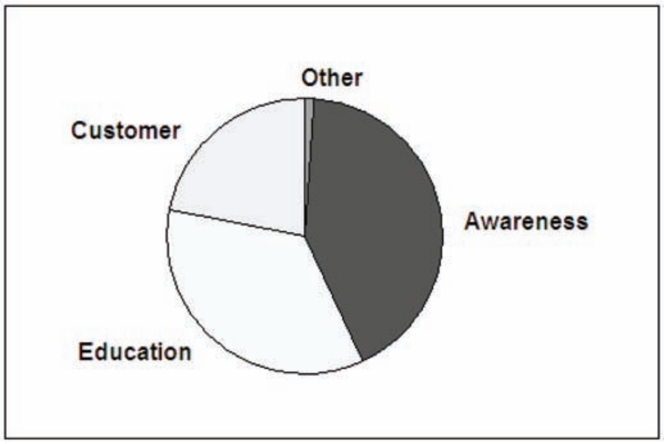
An employee survey was used to determine which of the adoption barriers were prevalent. Based upon some initial interviews, the survey was distributed via e-mail, and incentives were used to encourage participation. One hundred employees were randomly selected from the corporate directory. Two follow up e-mails were used to encourage users to participate and the average return rate was 72 percent. The questions on the survey included the following:

1. Have you ever heard of Collaborative or Social Software? (Determine Awareness)
2. If you have heard of these, does your organization utilize them? (Cultural, Political, or Social)
3. What is the primary use of the Collaborative and Social Tools within your organization? (Education)
4. Please describe your overall impression (use, purpose) of the products in a few sentences?
5. Please describe your experience with these tools in a few sentences (accessed infrequently, heavy user, loved it, etc.).

Notice that no specific questions were asked to differentiate between the cultural, political or social barriers. The reason for this was that the percentage was expected to be low as compared to the other issues of awareness and education. In order to ensure the survey group didn't know Collaborative and Social Software by other names, a product matrix was included that showed the specific products like Sharepoint, Confluence, or Open Source. The results of the surveys are presented in Figure 4.

The surveys were issued in January of 2007

Figure 4. Survey results (percentage)



and September of 2007. The first response to the survey showed an enormous awareness issue in which the vast majority of end users had not heard of the social offering or did not understand how the tool could be used. Over 77 percent of the respondents indicated that the lack of awareness or education was the primary reason for not implementing collaborative solutions in their environment. Only 2 percent acknowledged the actual use of the tools during the few years and the remaining 21 percent focused on the cultural, social and political issues. This result would indicate the need to address the awareness and educational issues first.

Field Trials II, III and IV

Based upon the results of the survey, awareness and education were identified as the two main issues. The user community simply did not know the tools existed or how they could be used in a business environment. The current deployment could easily be expanded to the new organizations will very little effort. For example, the community of practice was expanded to include the new organizations as was the ordering process replicated across all three companies. The client-support was centralized as a one stop location for the end user community to obtain information and support in a self service fashion. The results of these field trials showed a dramatic increase in demand in just a few months. On average, the new organizations increased by 1,608 percent and 78 percent respectively. Key actions by the implementation team included the following:

- Expanded the Client-Support environment to include all three organizations
- Expanded the product offering (new releases, Web components, and applications)
- Expanded the service offering for additional training, education, and consulting
- Developed new reward and recognition programs
- Published articles in various organizational newsletters
- Posted to the corporate weblog and wiki daily
- Developed audio and video training programs
- Consolidated business process into a single customer experience
- Expanded the Community of Practice (CoP)
- Developed additional audio and video programs for awareness and education

- Integrated with other Collaborative tools like Podcasting, Audio Conferencing, Instant Messaging, and Social Software
- Celebrated Key Milestones like the 25,000th Collaborative Environment

At the time of this writing, the total number of collaborative sites had grown to over 28,000 with 2-3 million page views per month. The number of document objects exceeded 2 million. An additional survey was conducted in late 2007 which was similar to the one described earlier. The awareness and education problem of 77 percent had dropped to 33 percent within the nine months of implementing the framework.

The final field trial occurred in mid 2007 with the deployment of Social Software which included weblog, wiki, and book marking applications. The major difference in this effort was that the framework was applied from the beginning. Similar activities were performed including adding an online client-support environment, training, education, marketing, and extending the offering with additional products and services. Within 4 months, the total number of information points exceeded 5,500.

CONCLUSION

In this chapter, we have laid a framework to support the implementation of collaborative and social software. As discussed, users come to this technology with a wide variety of experience levels which cannot be assumed by the technology community. Client-support must be implemented if the business wants universal adoption and a high degree of business value. The different components of the framework address the concerns of the end user which can put them at ease and create a more open environment for integrating this new technology. The framework focused on the issues with awareness, education, and the offering. By addressing these components, organizations will focus their resources on the components that define a successful implementation. The results of the field trials indicate that within any environment or culture, mass adoption can be obtained. With mass adoption, the return on investment of these technologies will be high.

REFERENCES

Ahn, L., Davis, M., Fake, C., Fox, K., Furnas, G., Golder, S., et al. (2006). Why do tagging systems work? *Proceedings of the SIGCHI conference on Human Factors in computing systems*. Montreal, Canada: The Association of Computing Machinery.

Baoill, A. (2004). Conceptualizing the weblog: Understanding what it is in order to imagine what it can be. *Interfacings: Journal of Contemporary Media Studies*, *5*(2), 1–8.

Blood, R. (2004). How blogging software reshapes the online community. *Communications of the ACM*, *47*(12), 53–55. doi:10.1145/1035134.1035165

Boyd, D., Davis, M., Marlow, C., & Naaman, M. (2006). Social networks, networking & virtual communities: HT06, tagging paper, taxonomy, Flickr, academic article, to read. *Proceedings of the seventeenth conference on Hypertext and hypermedia*. Odense, Denmark: The Association of Computing Machinery.

Calongne, C. (2001, March). Designing for website usability. *Journal of Computing in Small Colleges*, *16*(3), 39–45.

Carr, N. (2003). *Does IT Matter? Information technology and the corrosion of competitive advantage*. Boston: Harvard Business School Press.

Chellappa, R., & Gupta, A. (2002). Managing computing resources in active intranets. *International Journal of Network Management*, *12*(2), 117–128. doi:10.1002/nem.427

Christodorescu, M., Ganapathy, V., Giffin, J., Kruger, L., Rubin, S., & Wang, H. (2005). An auctioning reputation system based on anomaly detection. *Proceedings of the 12th ACM conference on Computer and communications security*. Alexandria, VA: The Association of Computing Machinery.

Cold, S. (2006). Using Really Simple Syndication (RSS) to enhance student research. *ACM SIGITE Newsletter*, *3*(1), 6–9. doi:10.1145/1113378.1113379

Davis, J., Farnham, S., & Jensen, C. (2002). Finding others online: Reputation systems for social online spaces. *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*. Minneapolis, MN: The Association of Computing Machinery.

Dutta, A., & Roy, R. (2006). Managing customer service levels and sustainable growth: A model for decision support. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*. Kona, HI: Institute of Electrical and Electronics Engineers, Inc.

Erickson, T., & Gratton, L. (2007). Eight Ways to Build Collaborative Teams. *Harvard Business Review*, *10*(7).

Gibson, F., Teasley, S., & Yew, J. (2006). Learning by tagging: group knowledge formation in a self-organizing learning community. *Proceedings of the 7th international conference on learning sciences*. Bloomington, IA: The Association of Computing Machinery.

Gilmore, J., & Pine, J. (2001). *The experience economy: Work is theater and every business a stage*. Boston: Harvard Business School Press.

Hof, R. (2005). Mix, Match, And Mutate. Business Week Online. Retrieved October 1, 2006 from http://www.businessweek.com/@@76IH*ocQ34AvyQMA/magazine/ content/05_30/b3944108_mz063.htm

Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. *Proceedings of the 10th Conference on Knowledge Discovery and Data Mining*. Seattle, WA: The Association of Computing Machinery.

Hu, N., Pavlou, P., & Zhang, J. (2006). Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication. *Proceedings of the 7th ACM conference on Electronic commerce*. Ann Arbor, MI: The Association of Computing Machinery.

Jhingran, A. (2006). Enterprise information mashups: Integrating information simply. *Proceedings of the 32nd international conference on Very large data bases*. Seoul, Korea: The Association of Computing Machinery.

Johnston, B., & Supra, J. (1997). Toward an integrated approach to information services support. *Proceedings of the 25th annual ACM SIGUCCS conference on User services*. Monterey, CA: The Association of Computing Machinery.

Kittl, C., Petrovic, O., & Teksten, R. (2001). Developing Business Models for eBusiness. *International Conference on Electronic Commerce 2001*. Vienna, Austria: International Center for Electronic Commerce.

Lerner, R. (2006). At the forge: Creating mashups. *Linux Journal*, *147*, 10.

McAfee, A. (2006). Enterprise 2.0: The dawn of emergent collaboration. *Sloan Management Review*, *47*(3), 21–28.

McNay, H. E. (2000). Corporate Intranets: Building Communities with Data. *IEEE Technology & Teamwork*, 197-201.

Millard, D., & Ross, M. (2006). Blogs, wikis & rss: Web 2.0: hypertext by any other name? *Proceedings of the seventeenth conference on Hypertext and hypermedia*. Odense, Denmark: The Association of Computing Machinery.

O'Neill, M. (2005). Automated use of a wiki for collaborative lecture notes. *Proceedings of the 36th SIGCSE technical symposium on Computer science education SIGCSE '05*. St. Louis, MO: The Association of Computing Machinery.

O'Reilly, T. (2005). *What Is Web 2.0: Design patterns and business models for the next generation of software*. Retrieved July 17, 2006 from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Palmer, J. (2002). Designing for Web site usability. *Computer*, *35*(7), 102–103. doi:10.1109/MC.2002.1016906

Pereira, C., & Sousa, P. (2004). A method to define an enterprise architecture using the Zachman framework. *Proceedings of the 2004 ACM symposium on Applied computing*. Nicosia, Cyprus: The Association of Computing Machinery.

Reinhold, S. (2006). Wikitrails: Augmenting wiki structure for collaborative, interdisciplinary learning. *Proceedings of the 2006 international symposium on Wikis WikiSym '06*. Odense, Denmark: The Association of Computing Machinery.

Riehle, D. (2006). How and why wikipedia works: An interview with Angela Beesley, Elisabeth Bauer, and Kizu Naoko. *Proceedings of the 2006 international symposium on Wikis WikiSym ,06*. Odense, Denmark: The Association of Computing Machinery.

Smith, D., & Valdes, R. (2005). *Web 2.0: Get ready for the next old thing*. Gartner Research Paper. Stamford, CT.

Timmers, P. (1998). Business models for electronic markets. *Electronic Markets*, *8*(2), 3–8. doi:10.1080/10196789800000016

Weiss, A. (2005). The power of collective intelligence. *netWorker, 9*(3), 16-23.

KEY TERMS

Client-Support: Client-Support is a term used to describe the various efforts to ensure the success of an environment. These efforts would include education, training, communities of practice, online documentation and automated business processes for procurement.

Collaboration: Collaboration is defined as people working together on non-routine cognitive work. This activity is about behavior, work habits, culture, management, and business goals and value generated we people from diverse backgrounds come together.

Information Worker: The information worker is a label placed on individuals that primarily work with information and data. Information workers perform non-routine, cognitive, or creative work that often requires both structured and unstructured information inputs from multiple sources.

RSS: In the simplest form, RSS shares the metadata about the content without actually delivering the entire information source. An author might publish the title, description, publish date, and copyrights to anyone that subscribes to the feed. A feed reader application is required just as an e-mail client is required to read e-mail.

Social Tagging: Social tagging describes the collaborative activity of marking shared online content with keywords or tags as a way to organize content for future navigation, filtering, or search.

Weblog: A blog (short for weblog) is a personal online journal that is frequently updated and intended for general public consumption. Blogs are a series of entries posted to a single page in reverse-chronological order. These original entries cannot be edit by others but can be commented on by anyone.

Web 2.0: Web 2.0 is a term used to decribe the next generation of Web applications where information flows both from the producer as well as from the consumer. Additionally, Web 2.0 embraces more of a thin client architecture which allows for the assembly of various components. Together, end user conent and thin client applications make the Web 2.0 environment.

Wiki: A wiki is software that allows users to easily create, edit, and link pages together. Unlike a blog, the end user can actually update the original authors information.

This work was previously published in Handbook of Research on Electronic Collaboration and Organizational Synergy, edited by J. Salmons; L. Wilson, pp. 560-577, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 1.14
Security in a Web 2.0 World

Richard T. Barnes
SunGard Higher Education, USA

ABSTRACT

Web 2.0 has brought enumerable benefits as well as daunting problems of securing transactions, computers, and identities. Powerful hacker techniques, including cross-site scripting (XSS) and cross-site request forgery (CSRF), are used to exploit applications to reveal and steal, at the worst, confidential information and money, or, at the least, cause trouble and waste time and money for reasons that may be best described as fun or simply possible to do. The people interested in transgressing Web 2.0 applications do so for money, prestige, or for the challenge. An infamous hacker from the early days of the Internet now heads his own Internet security company. A more recent hacker of some infamy has created a stir of concern and consternation as to how pervasive and potentially destructive hacker attacks can be. Securing Web 2.0 applications requires a multifaceted approach involving improved code development standards, organizational policy changes, protected servers and workstations, and aggressive law enforcement.

DOI: 10.4018/978-1-60566-122-3.ch005

INTRODUCTION

With the multitude of benefits derived from the various Web 2.0 technologies, it is unfortunate that this book needs a chapter on security. Although the collaborations, synergies, and transformations of the collective Web technologies (known as Web 2.0) have immeasurably changed society in a good way, there is a bad element that we must recognize, understand, and defend against.

The relatively open and participative nature of Web 2.0 is, at once, a strength and weakness. Opening sites to user content and comment creates synergies that would not exist had the sites been restricted to a select few. However, it is difficult to restrict user input to only positive discourse; various motivations compel some to poison this well we know as Web 2.0.

Collectively, the responsibility and burden falls on organizations and individuals to share in constraining the enablers to minimize the damage to our 2.0 Web sites. Although it is helpful to understand the motivations behind the various (and growing) attacks, it is more important to follow best practices in code development and security design (Evers, 2007). The adage "the best defense is a good offense" does not apply well to Web 2.0 security. We cannot proactively prosecute and punish someone before they commit a Web attack; we may be on the road to a changing world, but constitutional rights cannot be trampled upon.

It is likely that some are dissuaded by the possibility of punishment if caught; but if only a few carry out Web attacks, our best approach, still, is to mount our best defense. It is of course equally important to prosecute security offenses. The threat of punishment has to be more than theoretical: Offenders must know that if they are caught, there will be consequences.

This chapter will explore the motivations, methods, and defenses against the malicious behaviors that cost time and money, and lessen the positives that can come from these technologies. There have been notable attacks to prominent Web sites; a few of these will be examined for their causes and associated effects. The evolution of the World Wide Web into version 2.0 has had social impacts, too. What are these impacts, and are there trends evident that may help us predict where security attacks and defense strategies will go in the future? Some possibilities are explored here and in subsequent chapters.

There is an old adage that says those who forget the past are condemned to repeat it. This idea cannot be forgotten in Web 2.0 security. We must remember how attacks happened before so we can avoid similar attacks in the future. By examining the trends, analyzing our mistakes, and understanding our needs, we can improve on Web 2.0 and make it better. That is how we got to version 2.0 from 1.0. Perhaps, as the Web evolves into what some in the community are calling Web 3.0, the lessons learned here will not be forgotten.

BACKGROUND

It is perhaps ironic that the following definition for application security comes from one of the best known wikis, Wikipedia. Application security encompasses measures taken to prevent exceptions in the security policy of an application or the underlying system through flaws in the design, development, or deployment of the application. This definition is an excellent start in addressing a very large problem. However, it does not really tell us why; that is, why is it necessary to prevent exceptions to security policy?

A broader definition may help. There are several definitions of the word *security*: The freedom from danger or the freedom from fear and anxiety are two variants that tell us why application security is so important to Web 2.0 applications. Identity theft, corporate espionage or sabotage, and/or simple maliciousness are certainly enough to give most of us some pause or anxiety. Application security, as it relates to Web 2.0, is now an area of great attention because of our collective need to be free of these dangers.

A confluence of factors has complicated our lives as Web 2.0 becomes a more significant presence. The graphics-rich functionality, collaboration, and opportunities have not only yielded "serendipitous innovation" (Tapscott & Williams, 2006), but less desirable consequences, too.

Consequences such as cross-site scripting (XSS) and cross-site request forgeries (CSRFs) were not anticipated when foundational Web 2.0 technologies were created. Asynchronous JavaScript and XML (extensible markup language), or AJAX, is a set of Web development techniques that enable Web sites to be interactive and rich with features that make the static Web pages of a few years ago seem, well, static. However, it is through AJAX and other technologies that Web attackers have created innovative ways to, at the least, cause mischief, and at the worst, cause severe harm.

The subject of Web 2.0 security has a number of important acronyms; AJAX is one. Others include XML, RSS (really simple syndication), and SOAP (simple object access protocol; Shah, 2006). This chapter is not necessarily intended to define and explain these terms, but more to put them in context with the larger problem (and challenge) of Web 2.0 security. Much of the recent literature talks of where the burden lies: with the developers, with the companies, and with the users. The reality, however, is that at least some of the burden falls on each of us.

Developers can build security into the applications they write instead of adding security later as an afterthought. Companies can elevate security to an enterprise-level initiative and build it into the products they sell. Although companies have been victims of security incursions, especially of late, the user has often been victimized, and it is often at the user level where good security hygiene is easiest and best applied.

A comprehensive approach to securing Web 2.0 applications stands the greatest chance for success. It is not enough to say that developers need to do a better job or companies cannot rush products to market when everyone can contribute to better security. How ironic that one of the great strengths of Web 2.0, collaboration, can, in effect, help solve one of its greatest weaknesses!

SECURITY PROBLEMS AND WEAKNESSES

AJAX in this context is not a cleanser made famous by a television jingle, nor is it a reference to the Greek who out of madness quite literally fell on his sword. AJAX in this context is a programming technique that employs a few other successful technologies including, as the acronym implies, XML and JavaScript (O'Reilly, 2005). Seen as an essential enabler of the Web 2.0 era, AJAX serves as the technical nexus for the development of the rich, interactive Web sites that we now take for granted. Capable of making direct communication with the Web server, AJAX Web sites can request server data without reloading the Web page. These behind-the-scenes data exchanges occur without the user's knowledge and provide a transparency to applications that make them distinctly more advanced than their Web 1.0 predecessors.

This direct communication between the Web server and Web application also provides an opening for application security transgressions (Enright, 2007). One method of attack, known as XSS, involves the injection of malicious code into a Web page viewed by others. The silent (i.e., hidden) execution of the AJAX code is perhaps one of the biggest areas of concern for IT security experts. Because transactions are occurring in the background without the user's knowledge and input, the potential for the execution of malicious code running unimpeded is large. This code may run on the computers of unsuspecting Web site visitors, exposing the visitors to the possible theft of sensitive information (e.g., banking information, authentication credentials, etc.). There are several types of XSS attack methods; however, the common denominator in the end result is that the user's browser always executes code that is, at least, not authorized, and at worst, destructive or compromising (McMillan, 2007).

Attacks using the CSRF method have been less common, but a large unrealized potential exists for more attacks (and more harm) in the future. The *cross-site* in CSRF is derived from the XSS method explained above and denotes some of the similarities between the two methods. Where the two methods are most different can be denoted in the *forgery* part of the name. According to Merriam-Webster Online, a forgery is defined as an imitation passed off as genuine. This is quite literally what the CSRF attack method does: Unauthorized (or forged) commands from a user are transmitted to or from a trusted Web site. The reliance and trust on an established (authenticated) identity is how the Web has worked, and the CSRF method exploits this common characteristic by imitating the authenticated user (Waters, 2007).

XML poisoning is another method of attack. Web 2.0 applications transmit XML data between the client and server as part of their normal operations. Although poisoning may sound a bit overstated, the method will corrupt (or poison) the XML data in such a way as to disrupt the processing of the information. The effect of this poisoning can range from denial of service (DoS), where the targeted server is bombarded with spurious requests that will, in effect, bring the server down or render it nonfunctional, to compromised confidential information.

RSS is a method and an XML 1.0-compliant format that aggregates disparate sources of Web information into a feed for the user. The user subscribes to the feed and sets the criteria for what information should be included (or excluded) in the feed. While this may sound like an incredible time saver for the user, what it also can do is expose the user to malicious code embedded in the source information. This code may also be aggregated along with the legitimate information, and the user will not necessarily know anything is amiss. This RSS injection method, as it is known, can install and launch software on the user's computer, potentially compromising confidential information.

Web services can be a fertile ground for Web 2.0 security intrusions. SOAP is the common protocol used for Web services remote procedure calls, and is therefore an often-used method for attack. Parameter manipulation and node exploitation are common techniques used in Web-services-related attacks. In parameter manipulation, the variables that are passed in services calls can be manipulated to suit the attacker's needs. If a site has insufficient validation of received parameters, the site is open to attack and compromise. Nodes on the Internet cannot be assumed to be secure: If a node is compromised, a SOAP message en route can be intercepted and modified as part of an attack. Although there are several variations on the Web services theme, these types of attacks are less prevalent than the JavaScript-based attacks.

WHO ARE THE ATTACKERS, AND WHAT ARE THEIR MOTIVATIONS?

Just as Web applications have evolved from the static, even boring presentation of webmaster-defined content, so too have the people who stage attacks on the Web. Because *evolution* may imply betterment, perhaps it is incorrect to say that attackers have evolved; what has really happened is that their methods have become more sophisticated as the Web infrastructure has become more complex. But, who are the attackers and why do they do what they do?

There are no clear definitions or labels for these Web attackers. While the word *hacker* has been generally used in a pejorative sense by the media, there are other words that may be more precise. The word *cracker*, also known as a black hat, refers to a person who illegally compromises the security of a computer system or network to reach a malicious end. A white hat will (attempt to) compromise security, but will also have valid permission from the system or network owners; the objective, of course, is to locate security holes and plug them before they are exploited. Not to imply a standard curriculum has been established, but to acquire the necessary skills to be a cracker, one should first be a hacker (Walker, 2005). The usage of the word *hacker* has changed over the years. Its etymology has mostly had a creative connotation. However, since the movie *Wargames* in 1983, the general (public) meaning has been to creatively manipulate computer code or procedures to achieve a desired outcome, generally a mischievous or malicious one. For purposes of brevity, the word *hacker* will be used in this chapter not to be imprecise, but more specifically to be consistent with the general connotation of the word.

The rebellious, antiauthority teenager may still exist, but the harm that erstwhile person can do is now less than before, mostly thanks to much higher numbers of installed security software. While it is certainly not ubiquitous, the awareness of computer viruses and worms is higher than what it was, therefore leading to a greater number of protected computers. There are exceptions, of course, but the days are mostly passed when a lone and unassisted teenager can cause significant harm (without a significant response) from his basement computer. However, the problem will never go away. As new, more complex applications are deployed, new exploitable vulnerabilities are found.

Perhaps one of the earliest and best known hackers who had the mind-set, skills, and intelligence for finding vulnerabilities is now out of jail. Having satisfied all penalties levied against him, he now has his own security consulting company and has published more than a few books. According to a recent interview, Kevin Mitnick's original motivation was "fun and entertainment." There was a thrill to gaining access to confidential or proprietary information (Brandon, 2007).

A hacker of more recent fame, the creator of the infamous Samy worm that forced MySpace off line after an XSS attack, created it as a prank according to an interview shortly after the incident. Despite the greater prevalence of security software, the Samy worm raised awareness to the increased risks associated with applications under the Web 2.0 umbrella. When it became known that Samy's very first AJAX project created the Samy worm, business and security leaders took note; if this is his first, what could be next? It may seem a bit perverse that Samy, the perpetrator of the attack, did not fault himself or MySpace for the worm's success, but instead blamed the client browsers for making it easy to run unauthorized JavaScript code on them (Lenssen, 2005).

Like Mitnick who preceded him, but perhaps to a lesser degree, Samy may have gained more than he has lost from his exploit. Although Samy Kamkar was sentenced to probation, community service, and a certain amount of restitution, his reputation has been indelibly etched into Web history. One online news source even referred to his exploit as "cunning" (Orlowski, 2005).

The motivations behind Web 2.0 attacks have changed. While it may have been a prank or an adventure a few years ago, the development of new Web exploits is now the stuff of nationalism, social castes, and real money. A denial-of-service attack that hit Estonia in early 2007 is suspected to be rooted in Russian nationalism, although some security experts dispute the nationalism claim (Brenner, 2007), saying it was motivated by anger over a government decision to move a revered monument. Despite the lingering issue of why the Estonian attack happened, Russia is thought to be a proverbial hotbed of malware (malicious software) engineering (Brenner). Other than the booming petro economy, the Russian "other" economy is bleak, and programmers not only see malware as a money opportunity, they also do not believe developing malware is wrong. Still, the nationalist bent cannot be discounted; a small Russian newspaper reportedly praised the local perpetrators of a hacking attack for their accomplishment at the capitalists' expense (Claburn, 2007). With a report that a Russian Web server was found to be hosting approximately 400 malware applications (Brenner), and a report that China's production of malware exceeds Russia, how long will it be before nationalist Web attacks are directed to the Western hemisphere?

In the West, a social order has developed among the people who develop and/or deploy malware on the Web (Claburn, 2007). Even though Kevin Mitnick has a legitimate career now after paying his dues to society, he is held in reverence and awe by young and aspiring hackers. Samy Kamkar is substantially younger, but his "Samy is my hero" prank on MySpace is still admired by people in their teens, 20s, and even 30s. Even though they are legitimate now because of their admiring fans, both Mitnick and Kamkar are certainly near the high end of the hacker social order.

The lure of an income, perhaps a large income, is a major motivator in both the West and in Asia. No longer is malware just written by and for the exclusive use of its writer. Malware kits, that is, kits for developing malware, are readily available for sale on the Internet. Originators of malware programs or kits are at or near the top of the hacker social order. These are the people who stand to earn the greatest amount of money from their actions, and the money can be considerable. Lower in the hacker social order are the people who buy these kits to develop their own malware, but that does not mean they are of lesser significance; the availability and the use of malware kits are a significant factor in the proliferation of malware on the Internet. Though the road can be tough (Kevin Mitnick was arrested several times), the hope of a lucrative hacking career and the high status it may bring can be compelling.

Malware kits are one source of income. Another money-earning method involves renting botnets. Botnets are networks (or groups) of security-compromised computers known as zombies. Botnets are also used in spamming (mass junk e-mail distributions), focused attacks to steal data, and denial-of-service attacks as client-naïve instruments. They have even been used to extort money under threat of a DoS attack.

Depending on the size of the botnet, that is, how many zombie computers are part of the botnet, the weekly rental income can be several hundred dollars. More income can be earned with the sale of an unpublished security vulnerability, perhaps up to US$1,000 or more (Evers, 2007). What is less known, but feared more, are the markets for identity-related data such as credit card numbers, social security numbers, and company-related confidential information to be used in espionage.

Phishing scams are relatively new, but their impact cannot be understated. With Web sites that look legitimate, but are in fact facades developed to acquire credit card numbers, passwords, social security numbers, and other personal data, phishing is proving to be a lucrative criminal method of cyber attack. Security firms such as Symantec have included phishing protection in their software, and Microsoft has phishing filter functionality included in Internet Explorer 7 to help minimize the dangers phishing can pose to the unwitting user.

WHAT CAN WE DO?

What can be done to mitigate the security risks and minimize the problems that have evolved from Web 2.0? We cannot eliminate the problem of cyber attacks on Web 2.0 applications or otherwise. The best we can hope for is to minimize the probability of an attack, or if attacked, minimize the damage done.

The field is a rapidly changing one, where hackers find a weakness, companies or individuals respond, and society slowly incorporates and adapts to the changes over the long term. During the 1980s, few could have imagined what the Internet in general and Web 2.0 in particular could have become in 2007, but what was unimagined then is reality now: Society will continue to adapt. One constant, though, is that hackers will exploit a weakness, and the rest of us will respond. Society's innovation will always be challenged by the hacker's innovation.

Many blame the rush-to-market mentality of companies in getting their Web 2.0 applications up and running. In their haste to meet deadlines, developers forego what few best practices there are to complete their tasks. There are two things that can be done to address this part of the problem. In 1996, Alan Greenspan, the Federal Reserve chairman, used the term "irrational exuberance" to characterize the overinflated values of the stock market; perhaps companies venturing into Web 2.0 territory need to tamp down their exuberance to a level where their developers can properly include security into their Web 2.0 applications. Secondly, Web security is a rapidly changing field; to keep up with the changes, developers should be continually trained on threats, techniques, and best practices. As mentioned before, there is not one answer to the security problem, but as companies work toward a whole solution, they should concentrate on two of the biggest threat sources: cross-site scripting and cross-site request forgeries.

There is more that companies can do to address Web security issues. Widely acknowledged as a problem, company confidential or proprietary data can appear in Web 2.0 weblogs (blogs) that may eventually become known to the company's competitors. This is not caused by hackers, but more likely company employees, disgruntled or simply naïve. Companies need to develop (better) data policies: Their data need to be identified, managed, and controlled (Vijayan, 2007). Knowing what data the company has and where they are located is an essential first step. Also essential to know is who has access to data; that is, companies should know who should have access and put controls in place to enforce company policy on data.

A more problematic approach is determining if and how to limit employee access to blogs. Whether deliberate or accidental, when an employee puts company data on a blog, they become public and potentially compromise the position of the company. An outright ban on company computers is certainly possible. Web sites can be blocked, firewall rules can be stringently applied, and so forth (Fanning, 2007), but how can a ban be enforced on noncompany computers? The short answer is it cannot, so the likely answer for companies is to take a middle position; develop, apply, and enforce a comprehensive data policy that can extend beyond the company's parking lot.

The onus should not be completely on companies to address Web security problems. Organizations are generally more strident about security, and individuals should learn from their employers and become more security conscious. Personal data should be guarded as closely as company data. Home computers can be current on the latest security software just as work computers are (usually). Individuals should recognize the public nature of the Internet and the Web, and realize that *public* does not mean *safe*. Web 2.0 has enabled the Web to become more of a conversation than a billboard; that conversation will have questionable and nefarious participants, and individuals have to remember that.

WHAT IS ON THE HORIZON FOR WEB SECURITY?

As suggested in the previous section, companies and individuals are reacting to the Web security issues that have been exacerbated with Web 2.0 technologies. Some believe there really is one magic solution, while others embrace the idea that a comprehensive strategy and solution is the best approach.

One notable and positive trend is the telltale movement to thin-client applications and hardware. Thin clients are essentially input and output devices only; the applications, heavy processing, and data reside on a server. With the advent of fast corporate and broadband networks, thin-client computing is more possible than ever. In 1995, Oracle founder Larry Ellison presaged a future where network computers replaced the personal computer (Bock, 2006). Though his reasoning was based on the growing complexity of personal computers during the 1990s and not the security concerns raised in the 2000s, his prediction was nevertheless prophetic. Microsoft now has in beta a server-based version of Office that will further the trend of thin-client applications. With computing power and data moving to more protected Web servers, opportunities for intrusion are lessened and malware attacks will be reduced.

The best practices for Web security are developing and becoming better known. When Web 2.0 applications began to appear, the best practices had not caught up to the technology, but that is changing now. Control Objectives for Information and Related Technology (COBIT) is both a format and forum for governance of all things IT. As practices become established and accepted by the IT community, COBIT-codified practices are updated and disseminated through traditional channels such as conferences and published standards. Web 2.0, podcasts, blogs, and wikis, though perhaps nontraditional channels now, are becoming evermore important in communicating best practices to the people and organizations that need them.

There may be a developing trend with negative consequences for Web security, at least in the short term. An increasing number of acquisitions of Web security specialty firms have some security experts concerned. With IBM's acquisition of Watchfire and HP's recent acquisition of SPI Dynamics, will the products of those companies be less available to the general market? These smaller companies, before their acquisitions, were autonomous leaders in Web 2.0 security. However, with their independence gone, it is unclear whether their products will be directly available to the market at large, or limited as offerings from the acquiring companies. This should be a short-term problem because their technology, and technology from other similarly acquired companies, will eventually make its way to the market (Germain, 2007).

CONCLUSION

Where are we now with Web security? This chapter has examined the background and evolution of where we were. It is evident now that with the advent of Web 2.0, the security technology for Web applications lagged behind the technology for the social software embodied in Web 2.0.

This gap in technologies created an opportunity for makers of mischief, maliciousness, espionage, and profit. However, the market has responded and will continue to respond.

In the context of Web security, terms such as AJAX and SOAP may never become well-known outside IT circles, however the impact they have (and will) have on all of us cannot be understated.

This is certainly not to suggest or imply support of hackers, their social order, or the nationalistic motivations behind Web security incursions and attacks, but what is the underlying net effect of hackers, nationalists, IT, and security companies? Is the net effect positive, negative, or neutral? Besides the obvious inconveniences and annoyances, the negative side includes companies sabotaged, extorted, and coerced out of uncounted sums of money. IT departments have had to divert resources to the security problem in order to plug the leaks known and identify the holes not yet known, and individuals have had identities stolen and bank accounts drained.

Is there a plus side? Perhaps yes—more companies have been created that specialize in Web security, IT department budgets have been increased to augment their security staffs and tools, and while of questionable value, some (formerly unemployed) people in Asia and elsewhere are now working to help meet the burgeoning demand for malware.

From a dispassionate viewpoint, one might argue that the plusses outweigh the minuses; that is, the response to Web security problems has been more beneficial to the many than harmful to the few. A victim of identify theft or the CEO of a company that was impacted by an attack or extortion will surely disagree.

It should be clear that there will always be concerns over Web security. A certain trend is that knowledge of the security threats arising from Web 2.0 is spreading, and users, developers, and companies are responding with education, new products, and new strategies to mitigate the risks.

In this Web 2.0 world of podcasts, blogs, and wikis, there is an unfortunate but real repeating process that helps keep IT security managers employed. As security awareness increases, knowledge increases; as knowledge increases, the potential for new security threats increases. Although security is fleeting, the field of IT security is forever.

REFERENCES

Bock, W. (2006). *Larry Ellison and the network computer that wasn't*. Retrieved December 30, 2007, from http://www.mondaymemo.net/031103feature.htm

Brandon, J. (2007). *Q&A with the forefather of hacking*. Retrieved December 24, 2007, from http://www.pcmag.com/print_article2/0,1217,a=213916,00.asp

Brenner, B. (2007). *Black hat 2007: Estonian attacks were a cyber riot, not warfare*. Retrieved December 30, 2007, from http://searchsecurity.techtarget.com/originalContent/0,289142,sid14_gci1266728,00.html

Brenner, B. (2007). *How Russia became a malware hornet's next*. Retrieved December 30, 2007, from http://searchsecurity.techtarget.com/originalContent/0,289142,sid14_gci1275987,00.html

Claburn, T. (2007). *Hacker profile becomes more social, adds women*. Retrieved December 24, 2007, from http://www.informationweek.com/shared/printableArticle.jhtml?articleID=205101618

Enright, G. (2007). *Web 2.0 vulnerabilities to watch for*. Retrieved July 22, 2007, from http://www.computerworld.com/action/article=9027342

Evers, J. (2007a). *Hacking for dollars*. Retrieved December 24, 2007, from http://www.news.com/2102-7349_3-5772238.html

Evers, J. (2007b). *The security risk in Web 2.0*. Retrieved July 17, 2007, from http://news.com.com/2102-1002_3-6099228.html

Fanning, E. (2007). *Editor's note: Security for Web 2.0*. Retrieved July 17, 2007, from http://www.computerworld.com/action/article=283283

Germain, J. M. (2007). *IT security and the no good, very bad Web app nightmare*. Retrieved December 24, 2007, from http://www.technewsworld.com/story/60208.html

Lenssen, P. (2005). *Samy, their hero: Interview*. Retrieved December 29, 2007, from http:blogoscoped.com/archive/2005-10-14-n81.html

McMillan, R. (2007). *Researchers: Web 2.0 security seriously flawed*. Retrieved December 24, 2007, from http://www.pcworld.com/printable/article/id,131215

O'Reilly, T. (2005). *What is Web 2.0: Design patterns and business models for the next generation of software*. Retrieved July 1, 2007, from http:www.oreilly.net.com/lpt/a/6228

Orlowski, A. (2005). *Web 2.0 worm downs MySpace*. Retrieved July 17, 2007, from http://www.theregister.co.uk/2005/10/17/web20_worm_knocks_out_myspaces/print.html

Shah, S. (2006). *Top 10 Web 2.0 attack vectors*. Retrieved July 17, 2007, from http://www.net-security.org/article.php?id=949&p=1

Tapscott, D., & Williams, A. (2006). *Wikinomics: How mass collaboration changes everything*. New York: Portfolio.

Vijayan, J. (2007). *Six ways to stop data leaks*. Retrieved July 17, 2007, from http://computerworld.com/action/article=285138

Walker, A. (2005). *How and why hackers want to get inside your machine*. Retrieved December 24, 2007, from http://www.informit.com/articles/printerfriendly.aspx?p=425380

Waters, J. K. (2007). *Web 2.0 entails "sleeping giant" security risk*. Retrieved December 24, 2007, from http://www.adtmag.com/print.aspx?id=21499

This work was previously published in Social Software and Web 2.0 Technology Trends, edited by P. Deans, pp. 58-67, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 1.15
Web Site Localization Practices:
Some Insights into the Localization Industry

Nitish Singh
Saint Louis University, USA

John E. Spillan
University of North Carolina, USA

Joseph P. Little
Saint Louis University, USA

ABSTRACT

The e-commerce industry has experienced spectacular growth, change and development. This situation has initiated an enormous business revolution that has affected the process of globalization tremendously. The goal of this study was to analyze the Web sites of localization companies that provide localization and translation services to other companies and see if they themselves are practicing what they are preaching. The results suggest that localization companies are indeed not practicing what they are preaching. Analysis shows that localization company Web sites are less localized than the Web sites of their clients, the multinational companies. The findings provide some implications to domestic and international marketers who currently operate in or are planning to enter into the global markets in the near future. [Article copies are available for purchase from InfoSci-on-Demand.com]

INTRODUCTION

From the mid-1990s to the present day, the e-commerce industry has experienced spectacular growth, change and development. The global online population is estimated to reach 1.8 billion by 2012 (Jupiter Research, 2008). In 2008 North America accounted for only 17.5% of the online population and that percentage is in decline as countries such as China, Brazil, India, and Russia show the highest level of online population growth (Internet World Stats, 2008). This situation has initiated an enormous business revolution that has affected the process of globalization tremendously (Cyr & Lew, 2003). During the past several years an entire industry (the localization industry) has grown up around helping companies design multilingual Web sites and software applications for different countries. An industry report estimates the size of the worldwide translation and localization services market at US$ 8.8 billion (DePalma & Beninatto, 2006). According to this report the commercial market for localization services is estimated at US$6 billion and the government market at US$ 2.8 billion worldwide.

Localization is the process of adapting products and services (Web sites, manuals and software) in accordance to linguistic, cultural, technical and other locale-specific requirements of the target market (Localization Program at California State University, Chico, College of Business, 2008). Localization is now being seen by multinationals as a necessary process to develop multilingual and multicultural content to effectively tap global markets. Forester research estimates put the 2006 global e-commerce revenues at around $12.6 trillion. Furthermore, research has shown that consumers prefer Web sites in their native language and Web sites that reflect their local preferences (Singh, Furrer, & Ostinelli, 2004). Thus, companies around the world are creating multilingual Web sites to tap this vibrant online market. Companies like IBM, Oracle, Intel and other have almost 90 international sites to take advantage of the global online markets and communicate with their global customers. This surge in creating multilingual online content and software has also led to the growth of the localization industry which is helping these companies by effectively translating their Web sites, user interfaces, software, and manuals. Beyond translation the localization companies are also involved in the following (Esselink, 2000):

- Making visual or graphics, technical and textual modifications to the site content.
- Rewriting the text, translating the text, and ensuring translation, idiomatic, and conceptual equivalence of the translated text.
- Modifying graphics, data fields, tables, forms, layout, colors and tables etc.
- Modifying the cultural content of the site or software to be congruent to the local culture.
- International e-commerce readiness for multi-country transactions.
- Web site navigational modifications to meet local preferences.

This study analyzes the Web sites of localization service providers in order to understand to what extent these companies are translating and localizing their own sites, and modifying graphics, layout, colors, text, policies, navigation, and cultural content. The study then compares the localization efforts of the localization vendor company with their clients, who are generally multinational companies. This analysis will reveal if the companies that are preaching localization are also effectively implementing it on their own Web sites. After all, the localization service providers should set the benchmarks for their clients to follow. Moreover, the companies that will be the winners in this fast growing and consolidating localization industry will be the ones that are able to highlight and exemplify the need for Web site localization. What is a localization service provider telling their current clients along with potential future clients if their own Web sites are not sufficiently localized? The goal of this article is to gain understanding of the localization processes used in the localization industry and to focus on the current trends in the localization industry.

This article is composed of seven sections. The first section presents and introduces a review of the globalization and localization literature. Section two explains the research methodology. Section three provides a presentation of the analysis. Section four provides a discussion of the results, section five describes some managerial implications, section six explains some limitations along with future research ideas, and finally section seven provides a conclusion.

LITERATURE REVIEW

Globalization and the Localization Industry

Increased ownership of computers and Internet usage is growing every day. Throughout the world the Internet is rapidly becoming the main source for information, shopping and services. Furthermore, computer and Internet users are increasingly from non-English speaking countries. One estimate indicates that 32% of Internet users are non-native English speakers. This number is increasing. The result of this huge Internet expansion motivated businesses to recognize the value of Website localization (Kwintessential, 2009). Moreover, the unending process of globalization is fundamentally altering the manner in which enterprises do business. When businesses globalize their e-business, there is a great need recognize that language, cultural expectations and trust play a huge role when building online Web capabilities (Culnan & Armstrong, 1999; Jarvenpaa, Tractinsky, & Vitale, 2000; Singh & Pereira, 2005; Violino, 2001). For an organization to be successful in this demanding setting, they must adjust their offerings so that their products and services present the appearance and feel of being produced locally. The process of localization begins with an understanding of a wide range of linguistic, cultural, content, and technical issues. A product or service presentation has to be tailored to the local customs and practices of a country or region. For example, producing a Web site in only the English language is not sufficient because the majority of the world does not understand English. Furthermore, even if consumers do understand English research has shown that they prefer Web sites in their native language (Singh et al., 2004). While many companies use machine translations to adapt the language used on their Web sites (Singh & Boughton, 2002), this type of translation is not sufficient either. Languages not only differ in their use of characters or syntax, but also their use of rhetorical style and use of metaphors. Therefore, a simple machine translation, without an understanding of a culture and its language, may result in a cultural faux pas (Singh et al., 2004). Beyond linguistics companies also need to consider such culturally sensitive areas such as persuasion techniques, colors, icons, signs, Web page layout, and cultural values when localizing Web sites (Singh & Pereira, 2005).

Previous research has shown that culturally sensitive Web content enhances the site's usability (Fock, 2000; Luna, Peracchio, & de Juan, 2002; Singh & Pereira, 2005; Simon, 2001). So, in order to effectively communicate to foreign online consumers it is beneficial for a firm to adapt their Web sites to the targeted market. Furthermore, research has shown that not only does Web site localization enhance usability but also attitude towards the Web site, perception of the ease of site navigation, and ultimately purchase intention (Singh et al., 2004).

The cultural impact is substantial and Hall (1976) believes that it is very difficult to act or interact in any meaningful way if they do not understand language and culture. The consequence of not including language and culture, when considering global Web presence, is discarding profitable global online consumer. The Internet, similar to any other advertising document, is a replica of the culture of the country or locale (Cyr & Trevor-Smith, 2004; Hermeking, 2005; Singh & Matsuo, 2004). According to Mooij (1998) advertising mimics a society's values. It can only be effective when it is inseparably connected to the primary culture of the group for which it is targeted. Studies have demonstrated that advertising that is harmonious with local cultural values is significantly more compelling than standardized advertising. Several researchers, therefore, have emphasized the use of country-specific cultural values appeal when developing international advertising campaigns and communication material (Albers-Miller & Gelb, 1996; Han & Shavitt, 1994). Research indicates that Web site localization and cultural customization promotes a better opinion regarding the site ultimately influencing people's purchase intentions (Singh & Pereira, 2005). Luna et al. (2002) discovered that culturally harmonious Web content creates a more user friendly environment where the user has clear instructions and comes away from the Website with a better attitude about the content that is presented. Consequently, the localization of Web sites also necessitates culturally tailoring the Web sites to be congruent with the cultural requirements of the local environment. Miscommunications, in the international context, generally take place when the message is seemingly mismatched with the local culture and does not produce the response that was expected towards the businesses products or services. The foreign language, signs and symbols, and Web content that is culturally different, creates confusion, frustration, offensiveness and in the long run a loss of business (Luna et al., 2002).

While company Web sites provide a major opportunity to impart and promote a corporate image and to sell products and services, the effectiveness of the Website depends almost exclusively on the value of its content (Pollach, 2005). An effective Web site is the one where the consumers invest a considerable amount of time reviewing the content of interest, requesting more information, and buying the goods or services offered (Liu, Marchewka, & Ku, 2004). As such, the quality and value of a Web site will be influenced by how the Web site mirrors the culture of the nation for which it has been designed (Fletcher, 2006; Singh & Pereira, 2005).

Culture influences just about everything we do, say, read, hear and think. Web sites are not immune to the affect of culture (Kwintessential, 2009). Those companies that are able to develop, manage and customize their business Web sites to the culture of the country they are doing business, will generate more interest in their company and ultimately increase the sales of their products and services.

The differences in cultures require international businesses to find ways to make their Web sites communicate with different cultures in different parts of the world. As such, the key acronym that has emerged in this new arena of business operation is GILT or Globalization, Internationalization, Localization and Translation (Lommel, 2003). Globalization addresses the enterprise issues associated with making a company truly global. So, for products and services this means integrating the internal and external business functions with marketing, sales, and customer support in the world market (The Localization Industry Standards Association, 2008). More specifically, Web site globalization includes two complementary processes: Internationalization and Localization.

Internationalization is the process of generalizing a product so that it can handle multiple languages and cultural conventions without the need for redesign. In more technical terms, it is the process through which back-end technologies are used to create modular, extendible, and accessible global Web site templates that support front-end customization (Singh & Boughton, 2005). This process enables company Web sites to be locally responsive to the end-user through front-end customization. Internationalization takes place at the level of program design and Web document development (Singh & Little, 2009).

Localization and translation is the process of adjusting a product or service and making it linguistically and culturally appropriate to the target locale. More specific to the current study, Web site localization is the process of the front-end customization, whereby Web sites are adapted to meet the needs of an international target market (Singh & Boughton, 2005; Singh & Little, 2009).

The localization industry can trace back its roots to early 1980's when the software industry was emerging as an upcoming sector of the US economy, and felt a need to translate software products in multiple languages (Globalization

Industry Primer (LISA), 2007). As the application of software grew across a cross-section of industries and with the growth of the Internet, the localization industry also saw sustained growth. Now the localization industry is seeing a growth phase with the need for translation and localization of software, manuals, packaging, and most importantly multilingual Web sites. The industry is also undergoing considerable consolidation. During the 1990's the trend toward industry consolidation started with small vendors joining hands to offer "one-stop shopping" for large software developers like Microsoft, Oracle, and IBM who needed translation and localization services in multiple languages (Cyr & Lew, 2003). The industry consolidation leads to the emergence of multi-language vendors (MLVs) which specialized in completing multi-language, multi-service localization/translation projects. These MLVs also used an outsourcing model where they outsourced the core translation services to single-language vendors (SLVs); Single Language Vendors normally work into one target language only, from one or more source languages (Esselink, 2000). The acquisition of Bowne Global Solutions by Lionbridge Technologies in 2005 lead to the emergence of Lionbridge as one of the largest Globalization and off-shoring companies in the industry. Similarly, SDL International, another major player in the localization industry, enhanced its portfolio by acquiring Trados Inc., which was a major translation technology solution provider. As the localization industry grows, and serves new and bigger clients across a cross-section of industries, it will need localization vendor companies to invest in process and product innovations and R&D to be competitive. In order for these large investments and comprehensive solutions to be provided, vendors will need to be backed by substantial capital investments. However, Leon Z. Lee (2005), an industry expert, warns that the current focus of large and small localization companies toward primarily cost leadership, automated enterprise workflows, and technology integration from corporate consolidations is not a recipe for long term growth and sustainability of this industry. Lee (2005) recommends that for the localization industry to be viable it needs to expand its role from just a translation or technology-solution provider to truly embracing the wider concept of localization by providing international marketing expertise. This international marketing orientation will then help the localization companies to expand their offerings by delivering localized information and comprehensive resident knowledge in designing marketing campaigns for geopolitical and ethnographic regions in areas of print advertisement, online brand valuation, and Website usability analysis (Lee, 2005).

The next sections of this article will detail the methodology, sample, and the analyses used in this study. Additionally, insights into the current level of localization practiced by localization vendors and their multinational clients are presented.

METHODOLOGY

To analyze the quality and extent of localization depicted on the localization vendor Web sites, the study conducted a content analysis of the vendor Web sites and Web sites of multinational companies. More specifically, content analysis methodology was used and a coding system was developed to measure various facets of the localization efforts. The coding system used in this study was adapted from Singh, Toy and Wright (2009). The coding sheet included items like:

- Ease of finding global gateway on the Web site
- Use of country code domain names of ccTLD
- Translation depth
- Local customer support
- E-commerce information and policies
- Navigational outlay

- Web site page structure/layout
- Use of Locale-specific graphics, colors and values.

To perform the content analysis two coders were trained in the coding scheme and jointly coded several Web sites. The inter-coder reliability on the sample of vendor and multinational client Web sites ranged from .82 to .86. It is suggested that inter-coder reliability needs to be above .80 in order to be acceptable (Grant & Davis, 1997). Thus, the coder reliability exceeded the suggested threshold.

SAMPLE

Analyzing all country sites (which can range from 10-90 and may include more than 2000 pages) to measure localization efforts was beyond the scope of this study. Thus, the study measured the localization efforts on the German and Spanish Web sites of each vendor company. German and Spanish have been forecasted as some of the top languages in which multinationals are localizing their sites.

To find a sample of localization vendor company Web sites the study used the vendor company data base provided at the Globalization and Localization Association Web site. In total the study was able to include only 53 localization vendor company Web sites in the sample, as these were the only companies we found having international Web sites for Germany and Spain. Thus, 53 companies and their German and Spanish sites served as the final sample, which included almost 106 Web sites and more than a thousand Web pages. The study also analyzed the company home site (mostly in English) to see the structure of the global gateway, Web page structure, and Web content depth and navigation. The sample of multinational company Web sites was selected from Forbes top 500 international company list. 100 multinationals were identified that had international Web sites and Web sites specifically for both Germany and Spain. Thus, a total of 100 multinationals with 300 country sites (U.S. English, Germany and Spain) were analyzed for this study.

RESULTS AND ANALYSIS

Number Unique Languages Supported

The purpose of finding the number of unique languages was to understand how many languages is the company providing its services in. The results show that on average a vendor company site had about 7 unique languages depicted. On the other hand the Vendor clients, such as Multinational company Web sites had on average 19 unique languages supported. This shows that Localization vendors are far behind their clients in terms of languages supported on the site (see Table 1). In fact the mean number of languages depicted by multinational Web sites (19.38) exceeded the maximum depicted (16) by the vendor sites. An independent sample t-test indicates the means are significantly different ($F = 37.708, p = .000$).

Ease of Finding Global Gateway

The aim here is measure how visible the link for international sites is from the company's U.S. English home page. Based on the comparison data between the vendor and client Web site, it seems 30 percent of client (Multinational) sites have a dedicated global gateway page compared to which only about 9 percent of vendor sites have a dedicated gateway page (see Table 2). The results of a chi-square test for two independent samples indicates there is a significant difference in the presentation of a global gateway page between vendor and multinational sites ($\chi^2 = 22.191$, $p = .000$). However, it seems both vendor and client sites are lacking quality gateway pages, which

Table 1. Number of languages used

Number of Languages Used

	N	Minimum	Maximum	Mean	Std. Dev.
Vendor No. of Languages	53	3	16	7.11	3.06
Multinational No. Of Languages	102	5	38	19.38	7.44

Table 2. Web site global gateway page

	Vendor Frequency	Multinational Frequency	Vendor %	Multinational %
No Link	0	7	0	6.80
Not Easy to Locate	6	20	11.32	19.42
Located at Middle Third	18	12	33.96	11.65
Located at Upper Right Corner	24	33	45.28	32.04
A Dedicated Global Gateway Page	5	31	9.43	30.10
Total	53	103	100.00	100.00

are crucial to drive international online traffic to country-specific sites.

Use of Country-Specific Domain

The goal here is to see if the company has invested in buying the country code top level domains also called cctld for the country. The use of ccTLD helps in international search engine optimization and also shows commitment of the company to that country market (see Table 3). The analysis shows that no vendor site was using ccTLD exclusively to create international sites. Most vendors were using some extension of .com/Spain of .com/Germany. The client multinational Web sites did relatively better in terms of use of ccTLD. About 26 percent of multinational sites were fully using ccTLD for their international Web sites. The results of a chi-square test for two independent samples indicates there is a significant difference in the use of ccTLD between vendor and multinational sites ($\chi^2 = 20.346$, $p = .000$).

Localization Assessment of Country-Specific Web Sites (Germany and Spain)

- Translation Depth: Translation depth was measured to see to what extent are the companies translating their Web pages relative to U.S. English Web pages. To measure translation depth the study counted the number of English page and local language primary links or main links on the home page of English and local language site. The results, in table 4, show that on vendor Web sites about 85 percent of English pages links were translated. On the other hand in terms of Multinational sites, on average about 67 percent of English page links were translated.
- Content Localization: This Category measures to what extent the company has localized its Web site content in terms of local

Table 3. Web site use of country-specific domain

	Vendor Frequency	Multinational Frequency	Vendor %	Multinational %
No ccTLD	22	20	42.31	19.80
Not Fully Using ccTLD	30	54	57.69	53.47
ccTLD Used	0	27	0.00	26.73
Total	52	101	100.00	100.00

Table 4. Web site English vs local language links

	N	Minimum	Maximum	Mean	Std. Dev.
Vendor - English	53	5	92	28.51	20.04
Multinational - English	101	4	299	56.08	47.10
Vendor - Local Language	102	2	96	24.23	19.01
Multinational - Local Language	202	3	177	39.73	30.55

support, e-commerce related information, and navigational ease.

- Local Customer Support and Contact: By analyzing the level of local customer support it can be measured to what extent is the company localizing its customer service efforts for a specific-locale (see Table 5). The results show that while both the vendors and the client multinationals are not fully localizing their sites, the vendor sites depict far less degree of localization efforts. Only about 4 percent of vendor sites had local support pages which were equivalent to their US Web site, compared to 24 percent by multinationals. The results of a chi-square test for two independent samples indicates there is a significant difference in the use of local customer support between vendor and multinational sites ($\chi^2 = 21.200$, p = .000).
- Availability of all policies and e-commerce information such as shipping policy, return, privacy, terms of use, copyright etc.

Table 5. Web site level of customer support

	Vendor Frequency	Multinational Frequency	Vendor %	Multinational %
No Online Support	6	14	5.94	6.93
Basic Support	15	20	14.85	9.90
Basic Support - Customer Contact	38	68	37.62	33.66
Several Pages of Support	38	51	37.62	25.25
Equivalent to English Site	4	49	3.96	24.26
Total	101	202	100.00	100.00

Table 6. Web site e-commerce and information use policies

	Vendor Frequency	Multinational Frequency	Vendor %	Multinational %
Not Available	70	26	69.31	12.87
Only One or Two of the Policies	2	30	1.98	14.85
2-3 Policies Available	3	18	2.97	8.91
Most Policies Available	4	32	3.96	15.84
All Policies Available	22	96	21.78	47.52
Total	101	202	100.00	100.00

Table 7. Web site navigational ease

	Vendor Frequency	Multinational Frequency	Vendor %	Multinational %
Very Poor	6	7	5.94	3.47
Poor	24	27	23.76	13.37
Few Navigational Elements	35	54	34.65	26.73
Navigational Elements Seen	28	61	27.72	30.20
Navigation Elements Equivalent to English Site	8	53	7.92	26.24
Total	101	202	100.00	100.00

Under this category the results for vendor and client sides were very different. Almost 70 percent of vendor sites did not have policies related to e-commerce and information use (see Table 6). This shows that most vendors are not very global in terms of conducting e-commerce. On the other hand, almost 47 percent of multinational sites had all the policies available. These results suggest that at least half the multinationals are localizing their site in terms of e-commerce readiness. The results of a chi-square test for two independent samples indicates there is a significant difference in the availability of policies between vendor and multinational sites (χ^2 = 101.136, *p* = *.000*).

- Navigational ease in terms of sitemap, local search, navigation buttons etc (see Table 7): Analysis of navigation revealed that only about 8 percent of local vendor sites had navigational elements that were equivalent to their US home pages. On the other hand the 26 percent of the multinational client site had navigational elements equivalent to the US English pages. In general, both Vendor and Client Web sites were not highly localized in terms of navigation. However, the results of a chi-square test for two independent samples indicates there is a significant difference in the navigational ease between vendor and multinational sites (χ^2 = 18.085, *p* <= *.001*).
- Layout and cultural adaptation
  - Web Page Structure (see Table 8): The Overall Look of the Site and Design: Under this category the objective is to measure to what extent the look and the layout of the Web site has been localized for a specific-locale. Surprisingly, almost 92 percent of vendor international sites for Spain and Germany were basically standardized templates of their US site.

Table 8. Web site page structure

	Vendor Frequency	Multinational Frequency	Vendor %	Multinational %
Standardized	96	92	92.31	45.10
Mostly Standardized	5	27	4.81	13.24
Some Differences	3	37	2.88	18.14
Localized	0	28	0.00	13.73
Highly Localized	0	20	0.00	9.80
Total	104	204	100.00	100.00

Table 9. Web site local culture

	Vendor Frequency	Multinational Frequency	Vendor %	Multinational %
Standardized	97	72	93.27	35.29
Mostly Standardized	4	52	3.85	25.49
Some Differences	1	38	0.96	18.63
Localized	0	34	0.00	16.67
Highly Localized	2	8	1.92	3.92
Total	104	204	100.00	100.00

Multinational sites also did not seem to achieve much localization under this category with almost 45 percent international sites being standardized. The results of a chi-square test for two independent samples indicates there is a significant difference in the structure of the Web pages between vendor and multinational sites (χ^2 = 66.671, *p* = *.000*) **(see** Table 9).

- Use of local models, graphics, colors and other cultural markers.Under this category the study measures if the site uses local models, different colors more appropriate for the country, and cultural symbols. The results show that only about 2 percent of vendor sites are localized or highly localized, compared to 20 percent of multinational sites. It seems neither vendors or multinational clients are truly focusing their efforts to culturally customize their sites, even though a growing body of evidence is suggesting that cultural customization of sites leads to better attitude and intentions to buy online (Singh & Pereira, 2005). The results of a chi-square test for two independent samples indicates there is a significant difference in the use of local culture between vendor and multinational sites (χ^2 = 95.101, *p* = *.000*).

DISCUSSION

The results clearly show that companies (localization vendors and multinationals) are currently not fully localizing their sites in terms of using country code top level domains, global gateway pages, customer support, e-commerce and information use policies, navigation, Web site structure, layout, colors, and graphics. It is even more concerning that companies selling localization services are actually localizing their own sites to a much lesser extent than multinational companies, the firms that tend to be their clients. The localization industry is not practicing what they preach.

In every single category examined in this study the multinational Web sites were shown to be more localized than the vendor Web sites. A telling comparison is the number of distinct languages used. Vendor sites average using seven different languages compared to multi national sites that average nineteen different languages. This result exemplifies the lack of localization practices being used by the vendors themselves. Furthermore, over 93 percent of vendor sites are culturally standardized.

None of the vendor sites were found to be using a ccTLD which is surprising as international domains are crucial for international search engine optimization. This may be due to the fragmented nature of the localization industry wherein small localization vendors from a specific country tend to serve their own local market and are content with their local customer base. However, large localization vendor Web sites also seemed to show lack of ccTLD use and an overall low level of localization on various parameters we used in this study. So, do these vendors really believe that Web site localization practices are important? From this study's results, the picture we get is that vendor sites are lacking commitment toward localization. However, before reaching any conclusion we should consider a bigger picture and understand what are the reasons for localization vendors to not sufficiently localize their sites? In the limitations and future research section we discuss some of these issues.

MANAGERIAL IMPLICATIONS

With few large localization vendors like Lionbridge Technologies, SDL International, and Translations.com holding the top positions there seems to be a large segment of niche markets that small localization vendors or single language vendors are able to serve without much competition. However, as the localization industry consolidates and matures, it will be difficult for small localization vendors to remain competitive and profitable by just providing generic translation/localization services to niche markets or local country markets. Even large sized localization vendors risk losing their competitive position due to industry consolidation, over-reliance on generic translation services, and cut throat price-based competition. If the localization industry wants to keep its competitive position and provide a healthy industry environment for both small and large localization vendors to grow, it must go beyond generic product offerings in the form of translation services, and expand the definition of localization to include not just translation but to also offer:

- Localization of the Website lay out and navigation based on locale-specific requirements.
- Cross-cultural Web site and user interface usability research.
- Country-market analysis
- Cultural customization of Web sites to specific locales
- International online business strategy and marketing expertise
- Online branding and advertising localization
- International search engine optimization and search engine marketing

These are just some ways the industry can expand the definition of localization and offer complementary services that can help companies differentiate their offerings from their competitors and stay profitable.

Localization services firms, in order to be profitable in the future, must practice what they preach. They must practice not only to exemplify their services being sold, but also to appeal to an international market. Research has shown that individuals prefer Web sites that are localized to their own language and culture (Singh et al., 2004).

LIMITATIONS AND FUTURE RESEARCH

The current study is an exploratory study focusing on the amount of Web site localization used by vendors offering Web site localization services compared to their clients level of Web site localization (multinational company Web sites). So, the data analysis here is a simple frequency examination. The goal of this study was to examine the use of localization practices frequency and that goal was met. However, further, more in-depth data analysis could be used in the future. Another limitation to this study was the sample used. The vendor Web sites may not be intended for an international audience. As stated earlier, many of the smaller localization firms are serving single, niche markets. On the other hand, multinational company Web sites, by their nature, are meant for international consumption. Therefore, it is expected that vendor sites may be less localized than their client's sites. However, the results show that even the large localization services firms lack localized content on their Web sites to meet the needs of an international market. Furthermore, even the smaller firms should aspire to exemplify the practices that they preach.

Future research directions should include a closer examination of localization practices within the localization industry. A comparison between the large localization service firms Web sites and their smaller, niche market, counterparts is needed. Also, a longitudinal study examining the increased amount of localization used on the internet is warranted. Is the trend to localize Web sites to a greater extent, or are more Websites trying to serve a smaller, local niche market instead of an international market? If the trend is to serve international markets then are the localization practices keeping up with international expansion?

To further shed light into why localization vendors are not actively localizing their sites we need further research to investigate their overall globalization strategy by asking questions such as:

- Are the localization vendors just targeting some large multinational companies from predominantly English speaking countries (U.S., U.K., Australia etc.) and a few other non-English speaking countries?
- What resource and marketing constraints do these vendor companies face?
- Another interesting question to investigate is to understand the top management willingness to globalization and their vision for globalization. It seems several small and medium sized localization companies are run by top management which has primarily a translation background. So is it the lack of business education background that is restricting the global expansion of localization vendors?

Thus, to get a full picture of Web globalization efforts of localization vendors, we should not lonely study localization vendors Web sites but also understand their overall globalization strategy and how it has evolved over time.

CONCLUSION

The findings of this research suggest that both localization vendor and multinationals are barely localizing their Web site offerings. This may not be all bad news, as more multinationals seek to tap online markets and compete for them, the winner will be the multinational sites that are truly localized and speak to their international customers in their language and culture. Localization service sales may be increased by simply practicing what they preach.

Not only has research shown that consumers prefer localized Web site content, but research has also shown that by localizing you can increase traffic to Web sites (Ferranti, 1999), and increase willingness to purchase (Singh et al., 2004). Localization services vendors may be missing out on increased sales by simply preaching and not practicing. Actually, applying what they preach to their own Web sites, according to research, should increase the amount of traffic to their sites while also increasing the willingness to purchase their localization services. This means that the localization industry can look forward to significant growth, but only if it can educate its multinational clients about the importance of localization and the best way to do that is to practice what they preach.

ACKNOWLEDGMENT

The data collection for this study was supported by students and Gary Muddyman, CEO of Conversis Global.

REFERENCES

Albers-Miller, N., & Gelb, B. (1996). Business advertising appeals as mirror of cultural dimensions: A study of eleven countries. *Journal of Advertising, 25*(Winter), 57-70.

Culnan, M. J., & Armstrong, P. K. (1999). Information privacy concerns, procedural fairness and impersonal trust: An empirical investigation. *Organization Science, 10,* 104-115.

Cyr, D., & Lew, R. (2003). Emerging challenges in the software localization industry. *Thunderbird International Business Review, 45*(3), 337-358.

Cyr, D., & Trevor-Smith, H. (2004). Localization of Web design: An empirical comparison of German, Japanese and United Stats Web site characteristics. *Journal of the American Society for Information Science and Technology, 55*(13), 1199-1208.

DePalma, D., & Beninatto, R. (2006). Language services 2006: Supply-side outlook. *Research Report by Common Sense Advisory.* Retrieved July 28 2008, from www.commonsenseadvisory.com

Esselink, B. (2000). *A practical guide to localization (Language International world directory).* Philadelphia: John Benjamins Publishing Co.

Ferranti, M. (1999). From global to local. *Infoworld, 21*(41), 36-37.

Fletcher, R. (2006). The impact of culture on Web site content, designs, and structure: An international and a multicultural perspective. *Journal of Communication Management, 10*(3), 259-273.

Fock, H. (2000, September). *Cultural influences on marketing communication on the World Wide Web.* Paper presented at the Multicultural Marketing Conference, Hong Kong.

Globalization Industry Primer (LISA) (2007). Globalization industry primer. *LISA.* Retrieved April 3, 2009, from http://www.lisa.org/Globalization-indust.468.0.html#c261.

Grant, J., & Davis, L. (1997). Selection and use of content experts for instrument development. *Research in Nursing and Health, 20,* 269-274.

Hall, E. T. (1976). *Beyond culture.* Garden City, NY: Doubleday & Company.

Han, S. P., & Shavitt, S. (1994). Persuasion and culture: Advertising appeals in individualistic and collectivistic societies. *Journal of Experimental Psychology, 30,* 8-18.

Hermeking, M. (2005). Culture and Internet consumption: Contributions from cross-cultural marketing and advertising research. *Journal of Computer-Mediated Communication, 11*(1). Retrieved April 3, 2009, from http://jcmc.indiana.edu/vol11/issue1/hermeking.html.

Internet World Stats (2008). World Internet users and population stats. *Internet World Stats*. Retrieved July 28, 2008, from http://www.internetworldstats.com/stats.htm.

Jarvenpaa, S. L., Tractinsky, N., Vitale, M. (2000). Consumer trust in an Internet store. *Information Technology and Management*, *1*(1-2), 45-71.

Jupiter Research (2008). Jupiter research. *Jupiter Research*. Retrieved July 28, 2008, from http://www.jupiterresearch.com/bin/item.pl/home/.

Kwintessential (2009). Culture and Web site localization. *Kwintessential*. Retrieved April 3, 2009, from http://www.kwintessential.co.uk/translation/articles/culture-Website-localization.html.

Lee, L. Z. (2005). Evolving localization and its brand extension. *Galaxy Newsletter, Q4*. Retrieved April 3, 2009, from http://www.gala-global.org/newsletters/newsletter_3516.html.

Liu, C., Marchewka, J., & Ku, C. (2004). American and Taiwanese perceptions concerning privacy, trust, and behavioral intentions in electronic commerce. *Journal of Global Information Management*, *12*(1), 18-40.

Localization Program at California State University, Chico, College of Business (2008). What is localization?. *The Localization Program*. Retrieved November 26, 2008, from http://www.csuchico.edu/localize/whatislocalization.html.

Lommel, A. (2003). LISA, The localization industry primer, 2nd edition. *LISA*. Retrieved July 28, 2008, from http://www.lisa.org/Globalization-Indust.468.0.html?&no_cache=1&sword_list[]=industry&sword_list[]=localizatio&sword_list[]=primer.

Luna, D., Peracchio. L. A., & de Juan, M. D. (2002). Cross-cultural and cognitive aspects of Web site navigation. *Journal of the Academy of Marketing Science, 30*(4), 397-410.

Mooij, M. D. (1998). *Global marketing and advertising. Understanding cultural paradox*. Thousand Oaks, CA: Sage Publications.

Pollach, I. (2005). Corporate self-presentation on the WWW: Strategies for enhancing usability, credibility and utility. *Corporate Communications*, *10*(4), 285-301.

Simon, S. J. (2001). The impact of culture and gender on Web sites: An empirical study. *Database for Advances in Information Systems, 32*(1), 18-37.

Singh, N., & Boughton, P. (2002). Measuring Web site globalization: A cross-sectional country and industry level analysis. *Proceedings from American Marketing Association Educators' Conference* (Winter), Austin, TX. Chicago: American Marketing Association.

Singh, N., & Boughton, P. (2005). Measuring Web site globalization: A cross-sectional country and industry level analysis. *Journal of Web Site Promotion*, *1*(3), 3-20.

Singh, N., Furrer, O., & Ostinelli, M. (2004). To localize or to standardize on the Web: Empirical evidence from Italy, India, Netherlands, Spain, and Switzerland. *Multinational Business Review, 12*(1), 69-88.

Singh, N., & Little, J. (2009). Culturally customizing international Web sites. In M. A. Shareef, Y. K. Dwivedi, M. D. Williams, & N. Singh (Eds.), *Proliferation of the Internet economy: E-commerce for global adoption, resistance, and cultural evolution*. Hershey, PA: IGI Global.

Singh, N., & Matsuo, H. (2004). Measuring cultural adaptation on the Web: A study of U.S. and Japanese Web sites. *Journal of Business Research, 57*(8), 864-872.

Singh, N., & Pereira, A. (2005). *The culturally customized Web site: Customizing Web sites for the global marketplace*. Burlington, MA: Elsevier.

Singh, N., Toy, D. R., & Wright, L. K. (2009). A diagnostic framework for measuring Web site localization. *Thunderbird International Business Review, 51*(3), 281-295.

Violino, B. (2001). E-business lurches abroad. *Internet Week, March 19th*. Retrieved July 28, 2008, from http://www.internetweek.com.

The Localization Industry Standards Association (2008). What is globalization? *The Localization Industry Standards Association*. Retrieved July 28, 2008, from http://lisa.org/What-Is-Globalization.48.0.html.

This work was previously published in International Journal of E-Adoption, edited by S. Sharma, pp. 36-54, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Section II
Development and Design Methodologies

This section provides in-depth coverage of conceptual architectures, frameworks and methodologies related to the design and implementation of Web technologies. Throughout these contributions, research fundamentals in the discipline are presented and discussed. From broad examinations to specific discussions on particular frameworks and infrastructures, the research found within this section spans the discipline while also offering detailed, specific discussions. Basic designs, as well as abstract developments, are explained within these chapters, and frameworks for designing successful Web sites, Web-based applications, and Web portals are provided.

Chapter 2.1
Perceptions of Mobile Device Website Design:
Culture, Gender and Age Comparisons

Dianne Cyr
Simon Fraser University, Canada

Milena Head
McMaster University, Canada

Alex Ivanov
Simon Fraser University, Canada

ABSTRACT

Anytime anywhere services offered through mobile commerce hold great potential to serve customers in wireless environments. However, there is limited understanding of how mobile Web site design is perceived by diverse users. This chapter explores how users who differ by culture, age, and gender perceive the design of a mobile device and their subsequent level of satisfaction with the device. Sixty subjects were tested in a controlled laboratory experiment on an Internet enabled phone. The results of a quantitative analysis were statistically inconclusive in terms of cultural and gender differences, but significant differences were found between older and younger users. However, an in-depth qualitative analysis of interview transcripts revealed some interesting differences among cultural, gender and age groups. Consistent with findings in the stationary Internet domain, design elements were found to impact satisfaction with mobile services.

DOI: 10.4018/978-1-60566-366-1.ch010

INTRODUCTION

Organisations will be well served to not be complacent with their wireless site design efforts… [But] before wireless site designers can address the usability challenge, and before organisations can leverage the commercial benefits of m-commerce, a deeper understanding of what aspects of usability is important to users and how they may differ in a wireless context is required. (Venkatesh et al., 2003, p.56)

Mobile commerce[1] (or m-commerce) has huge potential to serve customers in wireless environments. The adoption of m-commerce is dependent on consumer acceptance of new and well-designed technologies (Ancker & D'Incau, 2002; Coursaris et al., 2003: Kim et al., 2002; Kumar & Zahn 2003; Nysveen et al., 2005; Perry et al., 2001; Schrott & Gluckler, 2004; Yang, 2005). It is expected design characteristics may influence user perceptions towards a mobile device. Congruent with work done by Cyr (2008), Information Design (ID), Navigation Design (ND), and Visual Design (VD) may all contribute to user adoption of a technology, as well as to satisfaction with a mobile technology.

Concerning user attitudes of handheld devices, it is also expected that diverse categories of users based on culture, gender, or age may react differently to using the device. There is growing literature on cross-cultural website design, mostly evaluated within the context of the stationary Internet (Becker, 2002; Chau et al., 2000; Cyr, 2008; Cyr et al., 2006; Cyr & Trevor-Smith, 2004; Marcus & Gould, 2000; Sun, 2001). More recently, research has examined culture and mobile data services (Choi et al., 2006). Investigations have likewise examined gender and design in the context of mobile devices, but research results are mixed (Anckar & D'Incau, 2002: Kwon & Chidambaram, 2000; Teo et al., 1999). Finally, research that examines cohorts by age (younger versus older users) is practically nonexistent when design is considered, although some studies have focused on special needs and preferences of older users (Goodman et al., 2005), or features of a mobile device perceived by user groups as adding value (Anckar & D'Incau 2002).

To explore the role of user differences concerning the perception of the design of a mobile device, users who differ on cultural, gender and age dimensions were tested on an Internet enabled cellular phone. Related to culture, Canadian and Chinese cultures were chosen due to acknowledged diversity (Hofstede, 1980). Between-group comparisons were conducted with respect to screen design (including information design, navigation design, and visual design), and satisfaction with the mobile device. In an exploration of these topics, the paper provides a review of relevant literature leading to the hypotheses for testing, the methodology used, an elaboration of results, and discussion of the findings. Given the increased diversity of mobile users, developing an expanded understanding of user perceptions and preferences not only has theoretical importance, but also serves to enhance the reengineering of devices to best meet consumer requirements.

MOBILITY AND DESIGN

In the realm of the stationary Internet, effective website design engages and attracts online consumers (Agarwal & Venkatesh, 2002; Fogg & Tzeng, 1999; Hui & Triandis, 1985; Morgan & Hunt, 1994; Schultz, 2003). According to Gommans et al. (2001), 'A website has to be designed for a targeted customer segment…' Chau et al. (2000) argue the modes of information presented on the Internet, and the quality of graphics has a significant impact on user experience. Research in design suggests various guidelines for effective Web navigation (Childers et al., 2001; Farkas &

Farkas, 2000), criteria for optimal Web design (Bernard, 2002; Egger 2001), and how aesthetics and usability might be linked (Tractinsky, 1997). The sensory experience of the website can help to determine if a user stays and shops at a site (Rosen & Purinton, 2004; Yoon, 2002).

The quality of handheld displays that favor enhanced information design and visual design is steadily increasing and affects user perceived effectiveness of the presentation (Rau et al., 2006). While mobile screens are much smaller than those available on the stationary Internet (Schmidt & Frick, 2000), various studies demonstrate comprehension rates on smaller screens are generally equivalent to their larger counterparts (Dillon & McKnight, 1990; Duchnicky & Kolers, 1983; Resiel & Shneiderman, 1987). Other researchers consider the smaller screens of mobile devices a 'serious obstacle to usability of the mobile Internet' (Chae & Kim, 2004, p. 165). Sarker & Wells (2003) examined interface characteristics and network capabilities that affect the implementation and acceptance of wireless phones. They discovered while users were 'quite forgiving of physical limitations of the device due to technological constraints, they were bothered by flaws in the interface of the devices' (p. 37).

In recent years there has been increased attention to mobile usage and in 2004 the *International Journal of Human-Computer Studies* devoted a special issue to this topic. The issue addressed mobile use in a variety of contexts including human characteristics and interface systems, although design aspects of the mobile website were not specifically considered. Further, in 2005 *Behaviour & Information Technology* devoted a special issue to mobility from a human computer interaction (HCI) perspective. Topics included navigation support, user acceptance and trust, and user evaluation of usability of mobile devices. In research in which interface design and usability are examined for wireless devices in m-commerce, Tarasewich (2003) suggests many current principles of interface design can be transferred to mobile devices. He examines various issues such as content, user interaction with the device, issues of reading text on small screens, rapid serial visual presentation, and browser types.

Relevant to the current research, design categories for information design, navigation design, and visual design as suggested by Garrett (2003) were selected for systematic examination. A definition of each category follows.

a. **Information design:** Elements of the site that convey accurate or inaccurate information to a user. For instance, the location of an icon on the screen would be the domain of information architecture, but whether or not that icon conveys the right information to a user is the domain of information design. Clear and logical presentation of information about services or products is also a component of information design.
b. **Navigation design:** The navigational scheme used to help or hinder users as they access different sections of the site, such as the location and format of navigation aids.
c. **Visual Design:** Elements that deal with the balance, emotional appeal, aesthetics, and uniformity of the website overall graphical look. This includes colors, photographs, shapes, or font type.

These categories are represented in other work in design (Agarwal & Venkatesh, 2002; Yoon, 2002; Flavian et al., 2005; Palmer, 2002; Simon, 2001), and while not exhaustive are representative of key elements of website usability. The same categories were used by Cyr (2008) to study website design across cultures, and by Cyr and Bonanni (2005) regarding website design and gender. Further, in a study of mobile services Choi et al. (2006) consider three categories of user experience across cultures. These are content (similar to information design), information architecture (which includes navigation design), and graphical user interface (which is similar to visual design). In this research, information design, navigation design, and visual design are considered in the specific context of a mobile device.

Culture and Design

User preferences for website design features are known to vary across cultures (Barber and Badre, 2001; Cyr, 2008; Evers and Day, 1997; Nielsen and DelGaldo, 1996; Sun, 2001). Cyr and Trevor-Smith (2004) examined design elements for 30 municipal websites in each of Germany, Japan, and the United States. Significant differences were found across countries for use of symbols and graphics, color preferences, site features, language, and content. In a study in which user impressions were evaluated toward eight website design features, numerous differences were detected between collectivist Japanese and Chinese users with individualist British users (Hu et al., 2004). In the current investigation participants are tested who are either Canadian or Chinese. With respect to these cultures, Singh et al. (2003) compared domestic and Chinese versions of websites for 40 American-based companies and found differences in all the cultural categories examined.

Studies of m-commerce in different countries and considering different cultures are rare, although 'an understanding of the cultural dimensions of a market can aid marketers immensely in developing appropriate m-commerce services…' (Harris et al., 2005). In response, research in the area of culture and mobility is beginning to emerge. Cross-country differences were found for adoption of mobile applications in Hong Kong, Japan, and Korea (Kim et al., 2004), between the UK and Hong Kong (Harris et al., 2005), and between France and the USA (Carlson et al., 1999). Lee et al. (2002) compared Japan and South Korea in a study of m-commerce usage, and found significant cultural differences regarding value structures of the mobile Internet and their effect on users' satisfaction.

More specifically, and with respect to information design, research comparing user preferences in Canada, the U.S., Germany and Japan for perceived access and presentation of product information on a stationary computer uncovered few significant differences between the U.S., Canada, and Germany, but significant differences ($p<.01$) between these countries and a highly collectivist[2] culture like Japan (Cyr et al., 2005). Based on qualitative comments from the study, there appeared a desire on the part of Canadians, Americans, and Germans for utility - at least as far as obtaining site information is concerned. Choi et al. (2006) examined cultural characteristics and user experience attributes in mobile data services in Korea, Japan, and Finland. Based on qualitative findings, the authors found user experience attributes correlated to the user's culture and 'Finnish participants showed a cultural profile opposite to that of the Koreans' (p. 192). For example, in the area of information design Koreans and Japanese (both collectivists), preferred large amounts of information on a single screen while Finns (individualists) preferred direct, explicit communication and reacted negatively to large amounts of content. Related to the preceding studies, like the Finns, Canadians are individualists. Alternately, Chinese are collectivists in alignment with Koreans and Japanese.

H1a: *There will be differences between Chinese and Canadian users in the perception of information design of a mobile device. Canadian users will prefer utility in information design, while Chinese prefer more detail and depth of information presented in a mobile medium.*

Regardless of culture, users prefer easy to navigate websites. In an experiment using a stationary website, Simon (2001) found that North Americans prefer navigation that enhances movement and makes the interface simpler to use. In the study by Choi et al. (2006) in mobile data services as already mentioned, Koreans and Japanese liked clear and logical ordering of menu items, while Finns mentioned they most liked search facilities. Despite the paucity of prior research on navigation design across culture, and particularly in the context of mobile devices the following exploratory hypothesis is offered.

H1b. *There will be no differences between Chinese and Canadian users in the perception of navigation design of a mobile device. Both Canadians and Chinese will prefer simple and logical navigation formats.*

User preferences vary by culture with respect to visual design of the interface. Color varies by culture. Red means happiness in China, but danger in the US (Barber and Badre 2001). When applied to Web design, color may impact user expectations of the interface as well as overall satisfaction (Barber & Badre, 2001). In a cross-cultural study on website design, a Japanese respondent indicated a preference for more pictures and an "emotional approach" (Cyr et al., 2005). In other work specifically focused on images used in website design, in qualitative analyses Canadians perceived images to have aesthetic, affective and functional qualities while Japanese respondents focused only on affective qualities (Cyr et al., 2006). Sun (2001) found that users from cultures such as China or Japan have a strong preference for visuals and aesthetic beauty of the interface. In a mobile context, Koreans and Japanese preferred colorful screen design and Finns preferred simple screen design with less emphasis on color (Choi et al., 2006).

H1c. *There will be differences between Chinese and Canadian users in the perception of visual design of a mobile device. Chinese will be more concerned than Canadians with aesthetic beauty in visual design.*

Gender and Design

Gender is frequently used as a basis for segmentation, and researchers have attempted to understand the fundamental similarities and differences between the men and women for decades (Deaux & Kite, 1987; Putrevu, 2001). Past empirical studies have shown significant gender differences across a variety of tasks and domains. For example, men often perform better than women on spatial orientation tasks, whereas females tend to score better on verbal or linguistic tasks (Simon, 2001; Deaux & Kite, 1987). Similarly, men and women differ in their reactions to visual images, affecting recall and recognition (Jones et al., 1998).

There are also gender differences in computer usage. Men and women diverge in Web acceptance, with perceived usefulness found to positively influence intention to use the Web more in men than women (Sanchez-Franco, 2006). Women use computers for collaboration and networking, while men view computers as a tool for obtaining and evaluating content (Gefen & Straub, 1997). Pearson et al. (2003) examined gender as a moderating variable to end-user computer efficacy, and found no differences between men and women although women were somewhat less confident to learn new computer applications. In other research, narrowing of differences between men and women has occurred concerning software use, anxiety, and enthusiasm (Rainer et al., 2003). Specific to mobile commerce, studies indicate that among Internet users men are predisposed to mobile adoption more than women (Yang, 2005; Brennan, 2000; Joines et al., 2003; Park & Jun, 2003; Rohm & Swaminathan, 2004). Contrary to these studies, gender differences were not detected in Spanish users related to shopping patterns and m-commerce adoption (Bigne et al., 2005).

It is anticipated the design of a website will impact user preferences, which in turn may produce different reactions between men and women (Chen & Dhillon, 2003). In one investigation in which gender and design are considered, Simon (2001) tests users' perceptions of a site, which refer to information richness, communication effectiveness, and communication interface. Women were found to have a less satisfied perception of the websites than men. Other work demonstrates differences between men and women for content and navigation (Maltby et al., 2003) and preference for color or graphics (Rodgers & Harris, 2003). In a study of website design, Cyr and Bonanni (2005) found specific information design elements (such as site organization and presentation of product information) were perceived more favorably by men than women. Further, men found the sites easier to navigate, and liked certain visual design aspects such as degree of interaction and animations more than women.

The above findings on a stationary website suggest there will be differences between the preferences of men and women for interface design in a mobile context. This assumption will be explored in the following hypotheses.

H2a. *There will be differences between men and women in the perception of information design of a mobile device.*

H2b. *There will be differences between men and women in the perception of navigation design of a mobile device.*

H2c. *There will be differences between men and women in the perception of visual design of a mobile device.*

Age and Design

Age is another common dimension used to segment consumer and user groups. There has been a growing level of interest and research in issues relating human computer interaction and age groups (Goodman & Lundell, 2005). Some researchers have focused on the unique design characteristics posed by older adults. Hawthorn (2000) lists various physical, sensory and cognitive limitations that may alter with increased age and their implications for interface design. Physical limitations, such as reduced dexterity and precision, can make the use of small and delicate input devices (as found on various mobile devices) more difficult. Sensory limitations can create limitations for the design of computer output and cognitive limitations can affect the design of the interface itself. For example, cognitive spatial ability has predicted computer performance (Kelley and Charness 1995), and was demonstrated to decline with increasing age (Salthouse, 1992). In other research, age affected the retention of computer training (Brown, 2001), and confidence in learning new applications (Crosby et al., 2003). Older employees generally exhibited a less positive attitude towards computers (Brown, 2001), and were less satisfied users (Simmers & Anandarajan, 2001) than younger employees. It is expected that well designed visual cues such as text links and icons are able to support the needs of older users. The format for organizing Web contents and the amount of information appearing on a screen enable higher performance for older users as their visual search skills and selective attention diminish (Ellis & Kurniawan, 2000).

As with gender, studies that investigate mobile use or m-commerce with consideration of age or design are few. Stroetmann et al. (2002) found that 43% of elderly people surveyed had at least some difficulty with mobile devices and 21% had considerable difficulty because of some physical or cognitive impairment. In a study of Swedish teenagers, Weilenmann and Larsson (2000) reported that young people use a mobile phone in radically different ways from more mature adults. Younger users use a mobile device more for expression than for information, and for social purposes rather than for coordination or efficiency. In a study with Spanish users, younger people are more predisposed to m-commerce adoption than older Internet users (Bigne et al., 2005). In a report on mobile use in India (MACRO 2004), limited adoption of mobile devices among older users resulted from small buttons on the handset and tiny screens that impede user visibility. Goodman et al. (2005) found that text, speech, and photographs were all effective ways to present landmark information to older users using a mobile navigation aid.

In this investigation we are interested to examine design characteristics (such information design, navigation design, and visual design) as outlined by Garrett (2003) for a mobile device, in this case an Internet enabled phone. To our knowledge this is the first attempt to systematically examine these design features and how they may differ for older or younger mobile users. To develop some understanding in this area, the following exploratory hypotheses are offered.

H3a. *There will be differences between older and younger users in the perception of information design of a mobile device.*

H2b. *There will be differences between older and younger users in the perception of navigation design of a mobile device.*

H3c. *There will be differences between older and younger users in the perception of visual design of a mobile device.*

Design and Satisfaction

For many years customer satisfaction has been studied in physical environments (Balasubramanian et al., 2003; Parasuraman et al., 1988; Oliver, 1980 and 19991999). More recently, research into consumer satisfaction has turned to the Web domain and examines 'stickiness' and 'the sum of all the website qualities that induce visitors to remain at the website rather than move to another site' (Holland & Baker, 2001). According to Anderson and Srinivasan (2002), e-satisfaction is defined as the contentment of the customer with respect to his or her prior purchasing experience with a given electronic firm. In the present research we adapt the definition for online satisfaction presented by Anderson and Srinivasan to suggest mobile satisfaction refers to contentment of the customer with the experience of using the mobile interface. However, there is no requirement for completion of a purchase in the current context. This definition is in alignment with Chae et al. (2002) who examined information quality related to user satisfaction for mobile Internet services.

Online satisfaction motivates online shoppers to stay at the site and return to the site in the future (Flavian et al., 2005; Bhattacherjee, 2001; Doll & Torkzadesh, 1988; McKinney et al., 2002) thus yielding a loyal customer outcome. In research into stationary websites, customer satisfaction is affected by content and context of the website (Flavian et al., 2005; Teo et al., 2003). More specific to information design, Szymanski and Hise (2000) discovered that product information and site design are critical to creating a satisfying customer experience. The experience of online shopping can be affected by the richness of product information presented (Palmer, 2002; McKinney et al., 2002), and is a dominant concern of the user (Kateranttanakul & Siau, 1999; Pitt et al., 1995; Zhang et al., 2000). A positive navigation experience and perception of a well-designed site may likewise result in online consumer satisfaction (Agarwal & Venketesh, 2002; Fogg & Tzeng, 1999; Palmer, 2002; Fogg et al., 2002; Hoffman & Novak, 1996; Koufaris, 2002; Nielsen, 2001), an enjoyable online shopping experience (Childers et al, 2001), and sales (Lohse & Spiller, 1999). Yoon (2002) found navigation functionality resulted in satisfaction, and induced Web visitors to remain at the site. Lohse and Spiller (1998) demonstrated that designing online stores with friendly user interfaces positively influences traffic and sales. Cyr (2008) found information design, navigation design, and visual design to all positively impact satisfaction for users from multiple countries.

While design research is limited within the mobile context, Tarasewich (2003) concluded: '[A]esthetics, along with usability, may also be part of designing an overall enjoyable user experience with mobile devices' (p. 12). Jiang and Benbasat (2003) found that mobile interface features can positively influence users' attitudes regarding product presentations. Graphical user interfaces, information architecture, and content all contributed to differences in perceived satisfaction with mobile data services (Choi et al., 2006). The preceding considerations lead to the final hypotheses:

H4a. *Perception of information design of a mobile device will impact user perceived satisfaction with the mobile device.*

H4b. *Perception of navigation design of a mobile device will impact user perceived satisfaction with the mobile device.*

Figure 1. Research model for mobile services

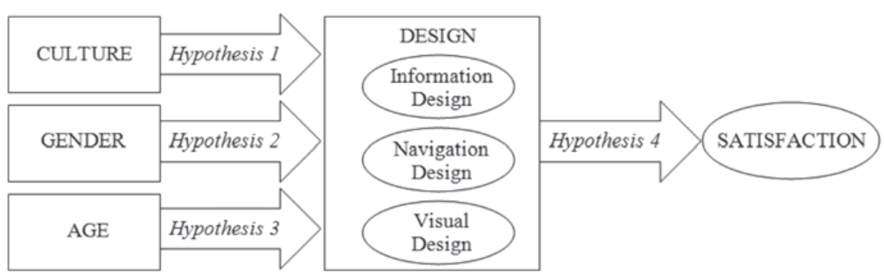


Table 1. Country comparisons (Source: based on Hofstede, 1980)

Country Dimension	Canada	China
Power Distance	Low (39)	High (80)
Uncertainty Avoidance	Low (48)	Medium (60)
Masculine	Medium (52)	Medium (50)
Individualism	High (80)	Very Low (20)
Long-Term Orientation	Very Low (23)	Very High (118)

H4c. *Perception of visual design of a mobile device will impact user perceived satisfaction with the mobile device.*

The above hypotheses as examined in this research are visually depicted in Figure 1.

RESEARCH METHODOLOGY

Participants

Sixty participants were recruited who were either Chinese or Canadian in origin (30 in each category). These countries were chosen to represent diverse cultural characteristics as per Hofstede (1980). Refer to Table 1.

Participants were also categorized by gender and age. Consistent with Aarnio et al. (2002), younger participants were 34 and younger and older participants were 35 or older. A relatively small sample size of 30 was chosen for each group due to the individualized and in-depth requirements of the data collection process. To participate, individuals must have used a cell phone for at least one year. In fact, all participants in the study were experienced users, and had owned a cell phone for approximately 4.5 years. Demographics of the sample appear in Table 2.

Experimental Site and Device

Given the emphasis on aesthetics in usability, an attractive site was required, without interference from a slow or faulty live Internet connection. Most WAP sites in North America and Europe are primarily text based, however the Lonely Planet offered some visual treatment (www.lonelyplanet.com with access from a mobile phone at wap.lonelyplanet.com). Therefore the City Guide version of this site was chosen for the experiment and further enhanced. The site for this study featured a colored background and text cells, as well as photographs, maps and icons, among other features. Each version of the site was created in both English and Chinese. For the Chinese version, site contents were translated to Chinese, 'back-translated' from Chinese to English, and then this version was compared with the original English version to verify content equivalence. Screenshot pages appear in Appendix A. Testing occurred on a Nokia 6600 Internet enabled cell phone. This

Table 2. Participant demographics

	Canadian (n=30)	Chinese (n=30)
Age	< 35: 67% ≥ 35: 33%	< 35: 60% ≥ 35: 40%
Gender	Male: 53% Female: 47%	Male: 43% Female: 57%
Highest level of education	High school: 40% Technical degree: 3% Undergraduate/College: 43% Masters/Doctorate: 13%	High school: 27% Technical degree: 3% Undergraduate/College: 53% Masters/Doctorate: 17%
Time spent online/week	22 hours	17.4 hours
Time owned a cell phone	4.8 years	4.4 years
Mobile Internet browsing experience	20%	17%
Use of the mobile Internet (if available)	Buy movie/concert tickets: 33% Download games/ring tones: 40% Check news: 50% Browse places to eat, shop, etc.: 47% Other: 33%	Buy movie/concert tickets: 30% Download games/ring tones: 60% Check news: 67% Browse places to eat, shop, etc.: 67% Other: 7%

phone was considered very suitable to the aims of the investigation as it has one of the largest screens on the market and has a superior color display (65,536 TFT) capable of laptop quality images. The interface included a joystick option for easy navigation.

To prevent problems with download and browsing delays typical of a real website using WAP technology, website pages for the Lonely Planet were downloaded and saved locally on the cell phone. Based on feedback obtained from a pilot focus group (of 6), participants were not aware of this manipulation and perceived the connection to be '*real*'. Important to this study, saving the site locally allowed the content to be modified and the display speeds to be controlled.

Experimental Tasks

The pilot study with 6 participants was used to pretest potential tasks and the experimental protocol including survey items and interview questions. Participants in the pilot study were initially asked to perform three information retrieval tasks: finding movie listings at a local theatre, choosing a restaurant in a different city, and booking a hotel in a different city. It was decided the restaurant task was most suitable because it was preferred by the pilot subjects and afforded excellent visual design opportunities. For the restaurant task, photos of the venue's interior as well as the map showing the location were included.

In the full study, each participant went through the experiment individually, under the supervision of a research investigator. For Chinese participants, all documents were translated and back translated, and a translator was available as required. The session began with a brief introduction and completion of a background data sheet, followed by familiarization with the Nokia 6600 device including a written summary of key functions. Once it was determined participants were comfortable with the device they were read the following narrative:

Imagine that you have just arrived in San Francisco to meet up with an old friend. Your friend has suggested that you select a restaurant on your cell phone, and call her back with the address. Use the bookmarked CityGuide site to accomplish this task. Spend as much time as you need browsing through the featured listings for San Francisco. There is no need to actually write down any information or make any calls. Just let me know what your selection is when you're finished.

The device was then handed to the participant with the browser opened at the introductory page of the site. The site listing for San Francisco featured four restaurants. Participants first completed the task, and then responded to a survey and were interviewed regarding their experience. At the end of the experiment, participants were debriefed and received a $20 honorarium for their time.

Survey Instrument

Following the completion of the experimental task, subjects were asked to complete a paper-based survey. In this section, we provide an overview of the survey measurement items, focusing on its content and construct validity.

Content validity considers how representative and comprehensive the items are in creating the experimental constructs. Validity is assessed by examining the process by which the construct items were generated (Straub, 1989). Constructs should draw representative questions (items) from a universal pool (Cronbach, 1971; Kerlinger, 1964). In this research, survey items were adapted from previously validated work on Information design (Cyr et al., 2004 and 20052005), Navigation design (Cyr et al., 2004 and 20052005), Visual design (Cyr et al., 2004 and 2005;2005 van

Table 3. Principle components analysis and reliability

		ID	ND	VD	SAT
Cronbach alpha		.61	.84	.89	.84
AVE		.57	.68	.60	.56
Items					
	ID1	**.775**	.221	.094	.101
	ID2	**.730**	.020	.311	.245
	ID3	.598	.271	.559	.176
	ID4	.546	.227	.537	.300
	ND1	.250	**.700**	.302	.175
	ND2	.004	**.891**	.072	.249
	ND3	.222	**.862**	.131	.039
	VD1	.110	.093	**.859**	.146
	VD2	.336	.157	**.833**	.182
	VD3	.101	.122	**.601**	.510
	VD4	.214	.200	**.789**	.291
	SAT1	-.159	.240	.247	**.661**
	SAT2	.326	.060	.223	**.857**
	SAT3	.313	.102	.137	**.810**
	SAT4	.408	.279	.290	**.646**

Notes: ID=Information Design; ND=Navigation Design; VD=Visual Design; SAT=Satisfaction

der Heijden, 2003), and Satisfaction (Cyr et al., 2004 and 20052005). All items were constructed as agree-disagree statements on a seven-point Likert scale. The complete survey appears in Appendix B.

Construct validity assesses the extent to which a construct measures the variable of interest. In other words, there should be high correlations between items of the same construct (convergent validity), and low correlations between items of different constructs (discriminant validity) (Straub, 1989). Results of the principal components analysis with varimax rotation appear in Table 3. The loadings for navigation design, visual design, and satisfaction construct items exceed recommended thresholds (Hair et al., 1998). However information design had two items (ID3 and ID4) that had high cross-loadings with items in the visual design construct. As such, ID3 and ID4 were removed from our analysis in order to maintain discriminant validity.

Discriminant validity can also be assessed by the average variance extracted (AVE) for each construct. As shown in Table 2, the AVEs were all above the recommended 0.50 level (Fornell & Larcker, 1981), which meant that more than one-half of the variances observed in the items were accounted for by their hypothesized factors.

Construct reliability (internal consistency) of the four factors was examined using Cronbach's α-value. As shown in Table 2, α-values ranged from 0.61 (for information design) to 0.89 (for visual design). Rivard and Huff (1988) suggest that this measure for reliability should be higher than 0.5 and ideally higher than 0.7. Navigation design, visual design and satisfaction α-values are well past this recommended threshold, and the α-value for information design (with its two items dropped) is also in an acceptable range. Therefore, our survey instrument encompassed satisfactory content validity (as evidenced from drawing construct items from existing validated literature); satisfactory convergent validity (as evidenced from high item loadings and construct reliability); satisfactory discriminant validity (as evidenced from low cross-loadings of factor items and the AVE for each factor); and satisfactory construct reliability (as evidenced from Cronbach's α-values).

Interviews

Following the completion of the survey, subjects were asked open-ended questions in a tape-recorded interview. The interview questions were meant to solicit additional information about the participants' experiences with the experimental task and interface. The questions probed how participants liked the design of the site, what they would change, and whether or not they found the device useful.

Responses were content analyzed and coded using Atlas.ti. This software provides an effective means to analyze qualitative data such as interview transcripts. The qualitative analysis process consisted of the following steps: (1) data preparation (i.e. interview transcription and formatting); (2) in vivo coding (use of participants' words as code labels) and open coding (use of arbitrary labels for code labels); (3) category and concept building in which semantic relationships between codes are identified to build higher conceptual abstractions; and (4) theory building based on interpretation of the results.

RESULTS

The descriptive statistics and correlations for the perceptual constructs are shown in Table 4. Each of the design variables (ID, ND and VD) was correlated to each other, as well as to overall satisfaction. From the demographic variables, only age is positively correlated with visual design.

Overall, respondents have a rather favorable impression of the mobile interface, with mean

Table 4. Descriptive statistics and correlations

	Mean	Std. Dev.	ID	ND	VD	SAT	GEN	CUL	AGE
ID	5.43	1.04							
ND	5.73	0.90	.37***						
VD	5.28	1.09	.53****	.43***					
SAT	4.98	1.11	.49****	.43***	.59****				
GEN	----	----	.09	.01	.09	.08			
CUL	----	----	-.06	-.09	.02	.17	.10		
AGE	----	----	.04	.07	.26**	.18	-.10	.07	

Notes:
1 ID=Information Design; ND=Navigation Design; VD=Visual Design; SAT=Satisfaction
2. Descriptive statistics not provided for GEN, CUL and AGE, as these are dichotomous variables
3. * p<.10; ** p<.05; *** p<.01; **** p<.001

scores of over 5 out of 7 for information design (ID), navigation design (ND), and visual design (VD). The mean score for satisfaction is 4.98, also indicating overall satisfaction with the site. To probe these results further, a word count was run on transcribed interviews using the atlas.ti software. The word 'easy' appeared 69 times, mostly in response to the question 'Try to describe the navigation experience on the site'. As such, the majority of users find the navigation of the device easy. One respondent elaborates, '*[U]sing the joystick was pretty straightforward, once I remembered to go left and right to go up the links. It was more intuitive to use the scroll button up or down…*'

Although information design receives a relatively high mean score (5.43), more than half the respondents thought there could have been more information about the restaurants on the site. This split into two category codes: one related to not enough information for each restaurant, which typically was lack of a menu (code *menu*) and two, not enough choice of restaurants (under the code *choice*). In terms of visual design of the site, the following quote captures some of the favorable sentiments as expressed by a number of users: '[E]*asy to use, attractively displayed, something that awes people, this display is very graphically appealing. I really want to play with this.'* Another user comments: *'The colors were good. The colors were actually fairly robust. Resolution seemed to be pretty good.'*

Tests of Culture, Gender, and Age as Moderators

T-tests of differences between culture, gender, and age group means for design and satisfaction are shown in Tables 5, 6, and 7 respectively.

Additionally, interview data was coded using two methods: (i) in vivo (using the participant's exact words as the basis for a code), and (ii) open coding (using arbitrary labels to code the data). Interview responses were systematically categorized in each design area (information, navigation and visual design). Altas.ti was used to create a concept map for the design areas, highlighting the actual number of responses in each code based on gender and culture. Refer to Appendix C. Although the numbers are relatively low, they signify response trends between the groups. In addition, representative quotations are included. Further, more theoretical themes were created from the emerging concepts, across the design areas. The main themes from our interview data were:

- **Information breadth:** The number of alternative (restaurant) choices. This is encapsulated by comment such as 'choices are limited' and unmet expectations for 'a lot of restaurants represented'.
- **Information depth:** The amount of detail for each alternative (restaurant) choice. This is encapsulated by participants seeking 'detailed menus', 'prices', 'famous dishes', 'parking', 'exterior shots', 'reviews' and 'hours of operation'.
- **Visual ease:** Utilitarian view on the interface design's capability to facilitate the task. This is captured through comments such as 'adequate resolution', 'easy to read' and 'easy to look at'.
- **Visual beauty:** Hedonic view on the beauty of the interface design. This is captured through comments such as 'cute' design, 'should be more charming', 'put some music', and 'animation would be good'.
- **Navigation layout:** The layout of the information within the site. Captured by comments such as 'laid out in a logical way', and too much 'scrolling'.

Table 5. T-test of differences between culture group means for design and satisfaction

	Group Means	t-value	p-value
ID	Can: 5.48 Ch: 5.37	0.433	.667
ND	Can: 5.81 Ch: 5.66	0.688	.507
VD	Can: 5.27 Ch: 5.32	-0.176	.861
SAT	Can: 4.80 Ch: 5.17	-1.279	.206

Notes: Can= Canadian; Ch=Chinese

Table 6. T-test of differences between gender group means for design and satisfaction

	Group Means	t-value	p-value
ID	M: 5.33 F: 5.52	-0.698	.488
ND	M: 5.72 F: 5.74	-0.076	.939
VD	M: 5.19 F: 5.39	-0.685	.497
SAT	M: 4.87 F: 5.06	-0.576	.567

Notes: M=Males; F=Females

Table 7. T-test of differences between age group means for design and satisfaction

	Group Means	t-value	p-value
ID	<35: 5.39 ≥35: 5.48	-0.277	.784
ND	<35: 5.68 ≥35: 5.82	-0.626	.534
VD	<35: 5.08 ≥35: 5.66	-2.274	.027**
SAT	<35: 4.83 ≥35: 5.25	-1.501	.139

Notes: * $p < .1$; ** $p < .05$; *** $p < .01$

Table 8. Summary of interview analysis across culture, gender and age groups

Design Area	Culture	Gender	Age
Information Design	Canadian: focused on Information Breadth Chinese: focused on Information Depth	Male: no noticeable patterns Female: focused on Information Depth (in particular, Chinese females)	Young: focused on Information Breadth Old: commented on both Information Breadth and Information Depth, with more focus on Depth
Navigation Design	Canadian: commented on both Navigation Layout and Challenges (in particular, Canadian women focused on Navigation Challenges) Chinese: very few comments on navigation	Male: some comments on Navigation Layout. Female: focused on Navigation Challenges (in particular, Canadian women).	Young: very few comments on navigation Old: some comments on Navigation Challenges
Visual Design	Canadian: focused on Visual Ease (in particular, Canadian men) Chinese: focused on Visual Beauty (mostly providing suggestions to augment beauty)	Male: No men commented on Visual Beauty; Canadian men commented on Visual Ease Female: focused on Visual Beauty (in particular, Chinese females)	Young: some comments on Visual Ease Old: commented on both Visual Ease and Visual Beauty (females only for Visual Beauty).

- **Navigation challenges:** Navigation/interaction challenges mostly stemming from inexperience with the new technology. Encapsulated by comments such as ‘counterintuitive joystick’ and not convenient to ‘press the left side to select the options’.

Table 8 summarizes the analysis of the interview transcripts across the above emerging themes and individual differences (culture, gender, age).

Based on the survey data, it is surprising no statistically significant differences are evident for culture or gender. However based on the qualitative analysis of the interview data some differences in these categories are indicated.

These results support H1a that Canadians prefer utility of information design, while Chinese prefer more detail and depth of information. Canadians were much more disappointed with the number of restaurant choices (Information Breadth) than Chinese, who focused on the lack of restaurant details (Information Depth). This Chinese focus on Information Depth was particularly evident among Chinese women. H1b was also generally supported in that both Canadians and Chinese prefer simple and logical navigation formats. In the case of Canadians they thought navigation could in fact be simpler, and the Chinese made few specific comments. Finally, qualitative results also support H1c. It was predicted Chinese would be more concerned with aesthetic beauty of the mobile interface than Canadians. In fact, Chinese (women only) even commented on how to augment the beauty of the interface, and Canadians focused more on visual ease.

It is interesting to note that no male from either culture made comments on the visual beauty or hedonic elements of interface design. However there were some cases where women not only commented on their hedonic preferences, but also made distinctions of what they prefer versus what they think men prefer. Overall, while our quantitative analysis rejects hypothesis 2 concerning differences in gender, our in-depth qualitative analysis suggests there may be some interesting differences in these categories.

Some support is found for hypothesis 3. Using survey data, statistically significant differences exist between older users (35 or older) and younger users (under 35) for visual design, with older respondents indicating the design of the mobile

Table 9. Predicting satisfaction

	R^2	ΔR^2	β	Sig.
GENDER	.063	.013	.174	.551
AGE			.415	.171
CULTURE			.322	.270
ID	.422	.391	.220	.097*
ND			.230	.118
VD			.420	.002***
ID x GENDER	..031	-.021	-.120	.583
ND x GENDER			.192	.432
VD x GENDER			-.025	.924
ID x AGE	.175	.130	.061	.532
ND x AGE			-.111	.343
VD x AGE			.180	.122
ID x CULTURE	.071	.022	.306	.258
ND x CULTURE			-.105	.663
VD x CULTURE			-.104	.738
ID x GENDER x AGE	.018	-.035	-.005	.602
ND x GENER x AGE			.077	.654
VD x GENDER x AGE			.011	.948
ID x GENDER x CULTURE	.012	-.041	.086	.848
ND x GENDER x CULTURE			.181	.546
VD x GENDER x CULTURE			-.261	.573
ID x AGE x CULTURE	.131	.084	.361	.051*
ND x AGE x CULTURE			-.091	.636
VD x AGE x CULTURE			-.190	.448
ID x AGE x GENDER x CULTURE	.013	-.040	.166	.580
ND x AGE x GENDER x CULTURE			.116	.661
VD x AGE x GENDER x CULTURE			-.263	.452

Notes: 1. ID=Information Design; ND=Navigation Design; VD=Visual Design
2. * p < .1; ** p < .05; *** p < .01

interface more appealing. Generally speaking, older respondents seem more impressed with the novelty of the device and its design than the younger group. This is evidenced in the following comment from an older subject: '*The newness of it, it's captivating. Being efficient is fun sometime. Not having to wade through a whole bunch of stuff. Look at this, look at what it can do, you can do it really easily, and you can get the information before everybody else does.*'

Predicting Satisfaction

Regression analysis was performed to assess the determinants of satisfaction based on main effects and interaction effects of the variables in this research. Gender and culture categorization was straightforward, where dummy variables represented males/females and Canadian/Chinese participants. Age was also coded as a dichotomous variable for clarity in presentation (Morris et al., 2005). Participants who are 34 years of age or younger were categorized as 'younger', while those 35 or older were placed in the 'older' category (as per Aarnio et al., 2002). Morris et al. (2005) suggested that gender differences in technology perceptions are more pronounced among older workers, and the interplay between key demographic variables should be examined in addition to investigating isolated demographic characteristics. This notion is supported by others in the information systems field (Butler, 2000; Venkatesh et al., 2000), as well as the field of psychology (Nosek et al., 2002; Kubeck et al., 1996). Therefore, in Table 9 the regression analysis examines various combinations of three-way and four-way interactions between demographic variables (e.g., visual design X gender X age).

In general, results of the regression analysis in Table 9 support hypothesis 4 that design elements do impact satisfaction with mobile services.

The effect size of independent variables on a dependent variable can be determined by comparing the R^2 of the dependent variable with and without the presence of each independent variable (Chin, 1998). The calculation for effect size (f^2) is as follows:

$$f^2 = \frac{R^2_{\text{included}} - R^2_{\text{excluded}}}{1 - R^2_{\text{included}}}$$

The effect size of perceived information design, navigation design and visual design on satisfaction were $f^2=0.09$, $f^2=0.08$ and $f^2=0.23$, respectively. Cohen (1988) provides the following criteria for interpreting effect size: (i) for small effect size, $0.02<f^2\leq 0.15$; (ii) for medium effect size, $0.15<f^2\leq 0.35$; and (iii) for large effect size, $f^2>0.35$. Therefore, both information and navigation design were shown to have a small effect size on satisfaction, while visual design can be classified as having a medium effect size on satisfaction.

The only interaction effect shown to have a statistical impact on satisfaction is information design by age by culture. Not all groups were equally satisfied with information provided at the site. This is supported in our qualitative analysis. Canadians were primarily concerned about the limited number of restaurant options (Information Breadth), whereas Chinese sought more detailed information for each of the restaurant choices (Information Depth). As one Chinese participant noted, '*You need to know price of the restaurant, the surroundings, and what the location and street looks like, but maybe I would have to call them to get information and details*'. In contrast, a Canadian participant commented: '*I was expecting there would be a lot of restaurants represented. For San Francisco I would be disappointed, I wouldn't trust the source'*. This was particularly evident among the older participants.

DISCUSSION AND CONCLUSION

Egan (1998) makes the case for exploring individual differences in interface design, and claims 'differences among people usually account for much more variability in performance than differences in system designs or differences in training procedures' (p. 543). Individual differences may affect what users seek in a system's interface, and how they interpret such interfaces. Elements of a user interface appropriate for one group may not be appropriate for another.

For mobile devices, evidence of the impact of individual differences on design and satisfaction has been preliminary, scattered, and incomplete. In fact, the majority of previous research tends to examine the *adoption* of mobile devices by culture (Harris et al., 2005; Kim et al., 2004; Carlson et al., 1999), gender (Yang, 2005; Brennen, 2000; Joines et al., 2003; Park & Jun, 2003; Rohm & Swaminathan, 2004) or age (Wielenmann & Larsson, 2000) rather than based on design considerations.

In contrast, the current research explores user perceptions of mobile *design* by culture, gender and age. This has included a focus of information design, navigation design and visual design and how each impacts user satisfaction with an Internet enabled cellular phone. Our exploratory analysis using both quantitative and qualitative data reveals some interesting differences for culture and age between user groups related to our three mobile design dimensions and satisfaction. Equally interesting is that there are no gender differences for the mobile device as tested.

In the realm of information design, overall Canadian and Chinese users felt there was adequate information presented on the mobile device, but each group was attentive to different types of information. Canadians focused on 'Information Breadth' (more choice of alternative restaurants) and the utility of information design, while Chinese focused on 'Information Depth' (more detail about the existing choices). This finding is in alignment with the qualitative results of Choi et al. (2006) who discovered that collectivist Koreans and Japanese preferred large amounts of detail on a single screen. With respect to gender and information design, no specific patterns emerged for men however women tended to focus more on 'Information Depth', especially Chinese women. This finding suggests women more than men desire detailed information content, perhaps related to different modes of information processing (Maltby et al., 2003). With respect to age, younger participants focused on 'Information Breadth' while older users desired both breadth and depth of information, with more emphasis on depth. Taken collectively, younger Canadian users were most concerned with 'Information Breadth' while older Chinese females were most focused on 'Information Depth'. These results have implications for mobile interface designers who aim to best connect with users.

On the stationary Internet, there is evidence that preferences for visual design vary by culture (Cyr, 2008; Cyr & Trevor-Smith, 2004). Users from collectivist cultures such as China desire visuals and aesthetic beauty of the interface, with emphasis on "affective" qualities (Cyr et al., 2006). On a mobile interface Choi et al. (2006) similarly found that collectivist Koreans and Japanese preferred colorful screen design, while Finns preferred simpler, less colorful screen design. Results from the current investigation parallel these earlier studies. In the area of visual design, Canadians focused on Visual Ease (utilitarian elements of design) while Chinese commented more on the 'beauty' of the design. Related to gender, men in both cultures and age groups were more concerned with 'Visual Ease' while women were more interested in 'Visual Beauty' and more aesthetic elements of the interface.

Almost all participants found the mobile interface relatively easy to use and navigate. The sites are not deep, are laid out in a logical way, and the joystick is easy to master. Chinese participants had little to say about navigation, while Canadians women and older users generally commented on challenges of navigation the interface. This finding is aligned to other work in which women found stationary websites more difficult to navigate (Cyr & Bonanni, 2005), or where physical limitations may have implications for interface design with older users (Hawthorn, 2000).

Consistent with findings for the stationary Internet, it is not surprising that in this investigation design elements impact satisfaction in a mobile service context. Further, the current work is consistent with Chae et al. (2002) who found that information quality of mobile services contributed to satisfaction. However, Chae et al. examined different elements of the mobile interface than in our investigation. Collectively, these findings support the assertion that regardless of technology or device, the interface is often considered the most important component of the entire system to the end-user (Sarker & Wells, 2003), and plays an important role in user attitudes (Bidgoli, 1990). As such, further research related to effective mobile design has commercial implications for m-commerce and is aligned to investigations of effective website design in e-commerce (Agarwal & Venkatesh, 2002; Fogg & Tzeng, 1999; Hui & Triandis, 1985; Morgan & Hunt, 1994; Schultz, 2003).

This is an initial exploratory study with a limited sample size, representing only two cultures. Although participants are representative of the desired cultural groups, they may not fully represent the socio-economic group within their country. It is recommended that follow-up studies draw samples from larger populations, and from additional culture groups. Also, usability evaluations themselves may be culturally bound. As Yeo (2001) points out, participants from various cultures are prone to provide false statements during usability evaluations to allow the designer to 'save face'. Research is needed to develop usability and research methodologies that accurately reflect personal opinions and preferences across cultures. Further, designers tend to ignore the role culture may play in the design of the interface (Sheridan, 2001), in particular within the Web context (Jagne et al., 2005). While many studies have been inconclusive, cultural factors deserve further investigation (Kwon & Chidambaram, 2000). At the very least, researchers and designers will ideally seek to better understand design elements that promote cultural attractiveness.

It is noteworthy that only one mobile application (a restaurant selection) and only one WAP site (lonelyplanet.com) are used in this experiment. The site offered some visual treatment, but has a narrow structure and simple layout. A more complex site with a deeper hierarchy may reveal more pronounced differences across diverse groups. Future research should consider multiple designs across multiple mobile applications. A positive feature of this investigation is the application of design elements as outlined by the design community (Garrett's 2003 categorizations for information design, visual design, and navigation design) to a mobile treatment. Further, in a nascent area of study as represented here, a strong point of the investigation is the use of both quantitative and qualitative methodologies. In addition to surveys, interview data was evaluated using atlas.ti software to provide a systematic evaluation of words or phrases into categories relevant to this investigation. The use of atlas.ti afforded deeper insights into the user experience than survey data alone can provide.

This research is an important first step in understanding the impact of individual differences on the design and satisfaction of mobile services. This relatively unexplored area is worthy of future attention. The quest for further knowledge as to how culture, gender and age impact technology use and satisfaction implies appreciation that these relationships are dynamic, and subject to continuous transformation (Simon, 2001). Enhanced understanding of subtle dimensions related to specific user groups will eventually enable fulfillment of expectations for optimal anytime and anywhere services.

REFERENCES

Aarnio, E. A., Heikkila, J., & Hirvola, S. (2002). Adoption and use of mobile services empirical evidence from a Finnish survey. In *Proceedings of 35th International Conference of System Sciences, Hawaii, USA.*

Agarwal, R., & Venkatesh, V. (2002). Assessing a firm's web presence: A heuristic evaluation procedure for measurement of usability. *Information Management Research*, *13*, 168–121.

Anckar, B., & D'Incau, D. (2002). Value creation in mobile commerce: Findings from a consumer survey. *Journal of Information Technology Theory & Application*, *4*, 43–64.

Anderson, R. E., & Srinivasan, S. S. (2002). E-satisfaction and e-loyalty: A contingency framework. *Psychology and Marketing*, *20*, 123–138. doi:10.1002/mar.10063

Balasubramanian, S., Konana, P., & Menon, N. (2003). Customer satisfaction in virtual environments: A study of online investing. *Management Science*, *49*, 871–889. doi:10.1287/mnsc.49.7.871.16385

Barber, W., & Badre, A. N. (2001). Culturability: The merging of culture and usability. In *Proceedings of 4th Conference on Human Factors and the Web.* New Jersey: Basking Ridge.

Becker, S. A. (2002). An exploratory study on web usability and the internationalization of US e-businesses. *Journal of Electronic Commerce Research*, *3*, 265–278.

Bernard, M. (2002). *Criteria for optimal web design (Designing for usability).*

Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. *MIS Quarterly*, *2*, 351–369. doi:10.2307/3250921

Bidgoli, H. (1990). Designing a user-friendly interface for a decision support system. *Information Technology*, *12*, 148–154.

Bigne, E., Ruiz, C., & Sanz, S. (2005). The impact of internet user shopping patterns and demographics on consumer mobile buying behavior. *Journal of Electronic Commerce Research*, *6*, 193–209.

Brennan, M. (2000). *Dot-com flavor of the week or hand-held revolution?*

Brown, K. (2001). Using computers to deliver training: Which employees learn and why? *Personnel Psychology*, *54*, 271–296. doi:10.1111/j.1744-6570.2001.tb00093.x

Butler, D. (2000). Gender, girls, and computer technology: What's the status now? *Clearing House (Menasha, Wis.)*, *73*, 225–229.

Carlson, P. J., Kahn, B. K., & Rowe, F. (1999). Organisational impacts of new communication technology: A comparison of cellular phone adoption in France and the United States. *Journal of Global Information Management*, (July): 19–29.

Chae, M., & Kim, J. (2004). Do size and structure matter to mobile users? An empirical study of the effects of screen size, information structure, and task complexity on user activities with standard web phones. *Behaviour & Information Technology*, *23*, 165–181. doi:10.1080/01449290410001669923

Chae, M., Kim, J., & Ryu, H. (2002). Information quality for mobile internet services: A theoretical model with empirical validation. *Electronic Markets*, *12*, 38–46. doi:10.1080/101967802753433254

Chau, P. K., Au, G., & Tam, K. Y. (2000). Impact of information presentation modes on online shopping: An empirical evaluation of a broadband interactive shopping service. *Journal of Organizational Computing and Electronic Commerce*, *10*, 1–22. doi:10.1207/S15327744JOCE100101

Chen, S. C., & Dhillon, G. S. (2003). Interpreting dimensions of consumer trust in e-commerce. *Information Technology and Management*, *4*, 303–318. doi:10.1023/A:1022962631249

Childers, T., Carr, C., Peck, J., & Carson, S. (2001). Hedonic and utilitarian motivations for online retail shopping behavior. *Journal of Retailing*, *77*, 511–535. doi:10.1016/S0022-4359(01)00056-2

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. In G.A. Marcoulides (Ed.), Modern methods for business research (pp. 295–336). Mahwah: Lawrence Erlbaum Associates.

Choi, B., Lee, I., & Kim, J. (2006). Culturability in mobile data services: A qualitative study of the relationship between cultural characteristics and user-experience attributes. *International Journal of Human-Computer Interaction*, *20*, 171–206. doi:10.1207/s15327590ijhc2003_2

Choi, B., Lee, I., Kim, J., & Jeon, Y. (2005). A qualitative cross-national study of cultural influences on mobile data service design. In *Proceedings of Computer-Human Interaction, Portland, Oregon, USA, 2-7 April.*

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* 2nd ed. New York: Academic Press.

Coursaris, C., Hassanein, K., & Head, M. (2003). Understanding the mobile consumer. In S. Nansi (Ed.), *Wireless communications and mobile commerce* (pp. 132–165). Idea Group Inc.

Cronbach, L. J. (1971). Test validation in educational measurement. In R. L. Thorndike, (Ed.), 2nd ed. (pp. 443–507). Washington, D.C: American Council on Education.

Crosby, L. B., DeVito, R., & Pearson, J. M. (2003). Manage your customers' perception of quality. *Review of Business*, *24.*

Cyr, D. (2008). Modelling website design across cultures: Relationships to trust, satisfaction and e-loyalty. *Journal of Management Information Systems*, *24*(4), 47–72. doi:10.2753/MIS0742-1222240402

Cyr, D., & Bonanni, C. (2005). Gender and website design in e-business. *International Journal of Electronic Business*, *6*, 565–582. doi:10.1504/IJEB.2005.008536

Cyr, D., Bonanni, C., Bowes, J., & Ilsever, J. (2005). Beyond trust: Website design preferences across cultures. *Journal of Global Information Management*, *13*, 24–52.

Cyr, D., Bonanni, C., & Ilsever, J. (2004). Design and e-loyalty across cultures in electronic commerce. In *Proceedings of 6th International Conference on Electronic Commerce.*

Cyr, D., Head, J., Larios, H., & Pan, B. (2006). Exploring human images in website design across cultures: A multi-method approach. In *Proceedings for the Fifth Pre-ICIS HCI Research in MIS Workshop 2006, Milwaukee, Wisconsin*. Nominated for Best Paper Award.

Cyr, D., & Trevor-Smith, H. (2004). Localization of web design: An empirical comparison of German, Japanese, and U.S. website characteristics. *Journal of the American Society for Information Science and Technology*, *55*, 1–10. doi:10.1002/asi.20075

Deaux, K., & Kite, M. E. (1987). Thinking about gender. In B. B. Hess & M. M. Ferree (Eds.), *Analyzing gender: A handbook of social science research* (pp. 92–117). Newbury Park, CA: Sage Publications.

Dillon, R. J., & McKnight, C. (1990). The effect of display size and text splitting on reading lengthy text from the screen. *Behaviour & Information Technology*, *9*, 215–227. doi:10.1080/01449299008924238

Doll, W. J., & Torkzadesh, G. (1988). The measurement of end-user computing satisfaction. *MIS Quarterly*, *12*, 259–274. doi:10.2307/248851

Doney, P. M., Cannon, J. P., & Mullen, M. R. (1998). Understanding the influence of national culture on the development of trust. *Academy of Management Review*, *23*, 601–620. doi:10.2307/259297

Duchnicky, R. L., & Kolers, P. A. (1983). Readability of text scrolled on visual display terminals as a function of window size. *Human Factors*, *25*, 683–692.

Egan, D. E. (1998). Individual differences in human-computer interaction. In M. Helander (Ed.), *Handbook of human-computer Interaction* (pp. 543–568). Amsterdam: Elsevier Science Publishers.

Egger, F. N. (2001). Affective design of e-commerce user interfaces: How to maximize perceived trustworthiness. In *Proceedings of International Conference of Affective Human Factors Design*. London Press.

Ellis, R. D., & Kurniawan, S. H. (2000). Increasing the usability of online information for older users: A case study in participatory design. *International Journal of Human-Computer Interaction*, *12*, 263–276. doi:10.1207/S15327590IJHC1202_6

Evers, V., & Day, D. (1997). The role of culture in interface acceptance. In S. Howard, J. Hammond, & G. Lindegaard (Eds.), *Proceedings of Human Computer Interaction, INTERACT '97*. London: Chapman and Hall.

Farkas, K., & Farkas, J. B. (2000). Guidelines for designing web navigation. *Technical Communication, Third Quarter,* (pp. 341–358).

Flavián, C., Guinalíu, M., & Gurrea, R. (2005). The role played by perceived usability, satisfaction and consumer trust on website loyalty. *Information & Management*, *42*, 719–729. doi:10.1016/j.im.2004.07.001

Fogg, B. J., Soohoo, C., & Danielson, D. (2002). *How people evaluate a web site's credibility*? Results from a large study. Persuasive Technology Lab, Stanford University.

Fogg, J., & Tseng, S. (1999). Credibility and computing technology. *Communications of the ACM*, *14*, 39–87.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobserved variables and measurement error. *JMR, Journal of Marketing Research*, *18*, 39–50. doi:10.2307/3151312

Garrett, J. J. (2003). *The Elements of user experience: User-centered design for the web.* Indiana, USA: New Riders Publications.

Gefen, D., & Straub, D. W. (1997). Gender differences in the perception and use of email: An extension to the technology acceptance model. *MIS Quarterly*, *21*, 389–400. doi:10.2307/249720

Gommans, M., Krishan, K. S., & Scheddold, K. B. (2001). From brand loyalty to e-loyalty: A conceptual framework. *Journal of Economic and Social Research*, *3*, 43–58.

Goodman, J., Brewser, S. A., & Gray, P. (2005). How can we best use landmarks to support older people in navigation? *Behaviour & Information Technology*, *24*, 3–20. doi:10.1080/01449290512331319021

Goodman, J., & Lundell, J. (2005). HCI and the older population [Editorial]. *Interacting with Computers*, *17*, 613–620. doi:10.1016/j.intcom.2005.09.001

Hair, F., Anderson, R. E., Tathan, R. L., & Black, W. C. (1998). *Multivariate data analysis.* Englewood Cliffs, N.J.: Prentice Hall.

Harris, P., Rettie, R., & Kwan, C. C. (2005). Adoption and usage of m-commerce: A cross-cultural comparison of Hong Kong and the United Kingdom. *Journal of Electronic Commerce Research*, *6*, 210–224.

Hawthorn, D. (2000). Possible implications of aging for interface design. *Interacting with Computers*, *12*, 507–528. doi:10.1016/S0953-5438(99)00021-1

Hoffman, D. L., & Novak, T. P. (1996). Marketing in hypermedia computer-mediated environments: Conceptual foundations. *Journal of Marketing*, *60*, 50–68. doi:10.2307/1251841

Hofstede, G. (1980). *Culture's consequences: Comparing values, behaviors, institutions and organisations across nations*, 2nd ed. Thousand Oaks, CA: Sage Publication.

Holland, J., & Baker, S. M. (2001). Customer participation in creating site brand loyalty. *Journal of Interactive Marketing*, *15*, 34–45. doi:10.1002/dir.1021

Hu, J., Shima, K., Oehlmann, R., Zhao, J., Takemura, Y., & Matsumoto, K. (2004). An empirical study of audience impressions of B2C web pages in Japan, China and the U.K. *Electronic Commerce Research and Applications*, *3*(2), 176–189. doi:10.1016/j.elerap.2003.09.004

Hui, C. H., & Triandis, H. C. (1985). Measurement in the cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology*, *16*, 131–152. doi:10.1177/0022002185016002001

Jagne, J., Smith-Atatkan, S., Duncker, E., & Curzon, P. (2005). Cross-cultural factors of physical and e-shopping. In *Proceedings of the Eleventh International Conference on Human Computer Interaction.*

Jiang, Z., & Benbasat, I. (2003). The effects of interactivity and vividness of functional control in changing web consumers' attitudes. In *Proceedings 24th International Conference of Information Systems, Seattle, USA.*

Joines, J. L., Scherer, C., & Scheufele, D. (2003). Exploring motivations for consumer web use and their implications for e-commerce. *Journal of Consumer Marketing*, 20.

Jones, M. Y., Stanaland, A. J., & Gelb, B. D. (1998). Beefcake and cheesecake: Insights for advertisers. *Journal of Advertising, Summer.*

Kateranttanakul, P., & Siau, K. (1999). Measuring information quality of web sites: Development of an instrument. In *Proceedings of International Conference of Information Systems, Charlotte, North Carolina* (pp. 279–285).

Kelley, C. L., & Charness, N. (1995). Issues in training older adults to use computers. *Behaviour & Information Technology*, *14*, 107–120. doi:10.1080/01449299508914630

Kerlinger, F. N. (1964). *Foundations in behavioral research*. New York: Holt, Rinehart, and Winston.

Kim, K., Kim, J., Lee, Y., Chae, M., & Choi, Y. (2002). An empirical study of the use contexts and usability problems in mobile Internet. In *Proceedings of 35th Annual Hawaii International Conference on System Sciences, Los Alamitos, CA.*

Kim, M. K., Park, C., & Jeong, D. H. (2004). The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. *Telecommunications Policy*, *28*, 145–159. doi:10.1016/j.telpol.2003.12.003

Knight, M., & Pearson, J. (2005). The changing demographics: The diminishing role of age and gender in computer usage. *Journal of Organizational and End User Computing*, *17*, 49–65.

Koufaris, M. (2002). Applying the technology acceptance model and flow theory to online consumer behaviour. *Information Systems Research*, *13*, 205–222. doi:10.1287/isre.13.2.205.83

Kubeck, E., Delp, N. D., Haslett, T. K., & McDaniel, M. A. (1996). Does job-related training performance decline with age? *Psychology and Aging*, *11*, 92–107. doi:10.1037/0882-7974.11.1.92

Kumar, S., & Zahn, C. (2003). Mobile communications: Evolution and impact on business operations. *Technovation*, *23*, 515–520. doi:10.1016/S0166-4972(02)00120-7

Kwon, H. S., & Chidambaram, L. (2000). A test of the technology acceptance model: The case of cellular phone adoption. In *Proceedings of 33rd Hawaii International Conference on System Sciences.*

Lee, Y., Kim, J., Lee, I., & Kim, H. (2002). A cross-cultural study on the value structure of mobile internet usage: Comparison between Korea and Japan. *Journal of Electronic Commerce Research, 3*, 227–239.

Lohse, G. L., & Spiller, P. (1998). Electronic shopping. *Communications of the ACM*, *41*, 81–88. doi:10.1145/278476.278491

Lohse, G. L., & Spiller, P. (1999). Internet retail store design: How the user interface influences traffic and sales. *Journal for Computed-Mediated Communication, 5.*

MACRO. (2004). Mobile phone usage among the teenagers and youth in Mumbai. Report.

Maltby, Chudry, F., & Wedande, G. (2003). Cyber dudes and cyber babes: Gender differences and internet financial services. *Journal of Financial Services Marketing*, *8*, 152–165. doi:10.1057/palgrave.fsm.4770115

Marcus, A., & Gould, E. W. (2000). Cultural dimensions and global web user interface design. *Interaction*, (July/August): 33–46.

McKinney, V., Yoon, K., & Zahedi, F. M. (2002). The measurement of web-customer satisfaction: An expectation and disconfirmation approach. *Information Systems Research*, *13*, 296–315. doi:10.1287/isre.13.3.296.76

Morgan, R., & Hunt, S. D. (1994). The Commitment-trust theory of relationship marketing. *Journal of Marketing*, *58*, 20–38. doi:10.2307/1252308

Morris, G. M., Venkatesh, V., & Ackerman, P. L. (2005). Gender and age differences in employee decisions about new technology: An extension to the theory of planned behavior. *IEEE Transactions on Engineering Management*, *52*, 69–84. doi:10.1109/TEM.2004.839967

Nielsen, J. (2001). *Designing for web usability.* Indianapolis: New Riders Publications.

Nielsen, J., & Del Galdo, E. M. (1996). *International User Interfaces*. Wiley Computer Publishing, John Wiley & Sons, New York.

Nosek, B. A., Banaji, M. R., & Greenwalk, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics*, *6*, 101–115. doi:10.1037/1089-2699.6.1.101

Nysveen, H., Pedersen, P. E., & Thorbjørnsen, H. (2005). Intentions to use mobile services: Antecedents and cross-service comparisons. *Journal of the Academy of Marketing Science*, *33*, 330–346. doi:10.1177/0092070305276149

Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *JMR, Journal of Marketing Research*, *17*, 460–469. doi:10.2307/3150499

Oliver, R. L. (1999). Whence consumer loyalty? *Journal of Marketing*, *63*, 33–44. doi:10.2307/1252099

Palmer, J. W. (2002). Website usability, design, and performance metrics. *Information Systems Research*, *13*, 151–167. doi:10.1287/isre.13.2.151.88

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring customer perceptions of service quality. *Journal of Retailing*, *64*, 12–40.

Park, C., & Jun, J. K. (2003). A cross-cultural comparison of internet buying behavior: Effects of internet usage, perceived risks, and innovativeness. *International Marketing Review*, *20*, 534–553. doi:10.1108/02651330310498771

Pearson, J. M., Crosby, L., Bahmanziari, T., & Conrad, E. (2003). An empirical investigation into the relationship between organisational culture and computer efficacy as moderated by age and gender. *Journal of Computer Information Systems*, *43*, 58–70.

Perry, M., O'Hara, K., Sellen, A., Harper, R., & Brown, B. A. T. (2001). Dealing with mobility: Understanding access anytime, anywhere. *Transactions on Computer-Human Interaction*, *8*, 323–347. doi:10.1145/504704.504707

Pitt, L. F., Watson, R. T., & Kavan, C. B. (1995). Service quality: A measure of information systems effectiveness. *MIS Quarterly*, *19*, 173–187. doi:10.2307/249687

Putrevu, S. (2001). Exploring the origins and information processing differences between men and women: Implications for advertisers. *Academy of Marketing Science Review, 10.*

Rainer, R. K., Laosethakul, K., & Astone, M. (2003). Are gender perceptions of computing changing over time? *Journal of Computer Information Systems*, *43*, 108–114.

Rau, P. P., Chen, J., & Chen, D. (2006). A study of presentations of mobile web banners for location-based information and entertainment information websites. *Behaviour & Information Technology*, *25*, 253–261. doi:10.1080/01449290500222009

Resiel, J. F., & Shneiderman, B. (1987). Is bigger better? The effects of display size on program reading. In G. Salvendy (Ed.), *Social, ergonomic and stress aspects of work with computers* (pp. 113–122). Elsevier Science Publishers.

Rivard, S., & Huff, S. (1988). Factors of success for end user computing. *Communications of the ACM*, *31*, 552–561. doi:10.1145/42411.42418

Rodgers, S., & Harris, M. A. (2003). Gender and e-commerce: An exploratory study. *Journal of Advertising Research*, *43*, 322–329.

Rohm, J., & Swaminathan, V. (2004). A typology of online shoppers based on shopping motivations. *Journal of Business Research*, *57*, 748–757. doi:10.1016/S0148-2963(02)00351-X

Rosen, D. E., & Purinton, E. (2004). Website design: Viewing the web as a cognitive landscape. *Journal of Business Research*, *57*, 787–794. doi:10.1016/S0148-2963(02)00353-3

Salthouse, T. A. (1992). Reasoning and spatial abilities. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition,* 3rd ed. (pp. 167–212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sanchez-Franco, M. J. (2006). Exploring the influence of gender on the web usage via partial least-squares. *Behaviour & Information Technology*, *25*, 19–36. doi:10.1080/01449290500124536

Sarker, S., & Wells, J. D. (2003). Understanding mobile handheld device use and adoption. *Communications of the ACM*, *46*, 35–40. doi:10.1145/953460.953484

Schmidt, S. H., & Frick, O. (2000). WAP – designing for small user interfaces. In *Proceedings of Conference on Human Factors in Computing Systems.*

Schrott, G., & Gluckler, J. (2004). What makes mobile computer supported cooperative work mobile? Towards a better understanding of cooperative mobile interactions. *International Journal of Human-Computer Studies*, *60*, 737–752. doi:10.1016/j.ijhcs.2003.11.006

Schultz, L. (2003). Effects of graphical elements on perceived usefulness of a library. Masters thesis.

Sheridan, E. F. (2001). Cross-cultural web site design: Considerations for developing and strategies or validating local appropriate online content. *MultiLingual Computing*, *12*.

Simmers, S., & Anandarajan, M. (2001). User satisfaction in the internet-anchored workplace: An exploratory study. *Journal of Information Technology Theory and Application*, *3*, 39–61.

Simon, S. J. (2001). The impact of culture and gender on web sites: An empirical study. *The Data Base for Advances in Information Systems*, *32*, 18–37.

Singh, N., Xhao, H., & Hu, X. (2003). Cultural adaptation on the web: A study of American companies' domestic and Chinese websites. *Journal of Global Information Management*, *11*, 63–80.

Straub, D. W. (1989). Validating instruments in MIS research. *MIS Quarterly*, *12*, 147–170. doi:10.2307/248922

Straub, D. W., Loch, W., Aristo, R., Karahanna, E., & Strite, M. (2002). Toward a theory-based measurement of culture. *Journal of Global Information Management*, *10*, 13–23.

Stroetmann, V. N., Husing, T., Kubitschke, L., & Stroetmann, K. A. (2002). The attitudes, expectations and needs of elderly people in relation to e-health applications: Results from a European survey. *Journal of Telemedicine and Telecare*, *8*, 82–84. doi:10.1258/135763302320302154

Sun, H. (2001). Building a culturally-competent corporate web site: An explanatory study of cultural markers in multilingual web design. In *Proceedings of SIGDOC'01*, 21-24, October (pp. 95–102).

Szymanski, D. A., & Hise, R. T. (2000). E-satisfaction: An initial examination. *Journal of Retailing*, *76*, 309–322. doi:10.1016/S0022-4359(00)00035-X

Tarasewich, P. (2003). Designing mobile commerce applications. *Communications of the ACM*, *46*, 57–60. doi:10.1145/953460.953489

Teo, H., Oh, L., Liu, C., & Wei, K. (2003). An empirical study of the effects of interactivity on web user attitude. *International Journal of Human-Computer Studies*, *58*, 281–305. doi:10.1016/S1071-5819(03)00008-9

Teo, T., Lim, V., & Lai, R. (1999). Intrinsic and extrinsic motivation in internet usage. *Omega International Journal of Management Science*, *27*, 25–37. doi:10.1016/S0305-0483(98)00028-0

Tractinsky, N. (1997). Aesthetics and apparent usability: Empirically assessing cultural and methodological issues. In *CHI Proceedings.*

van der Heijden. (2003). Factors influencing the usage of websites: The case of a generic portal in the Netherlands. *Information & Management*, *40*, 541–549. doi:10.1016/S0378-7206(02)00079-4

Venkatesh, V., Morris, G. M., & Ackerman, P. L. (2000). A longitudinal field investigation of gender differences in individual technology adoption decision-making processes. *Organizational Behavior and Human Decision Processes*, *83*, 33–60. doi:10.1006/obhd.2000.2896

Venkatesh, V., Ramesh, V., & Massey, A. P. (2003). Understanding usability in mobile commerce. *Communications of the ACM*, *46*, 53–56. doi:10.1145/953460.953488

Weilenmann, A., & Larsson, C. (2000). Collaborative use of mobile telephones: A field study of Swedish teenagers. In *Proceedings of NordiCHI 2000, Stockholm, Sweden, 23–25 October.*

Yang, K. (2005). Exploring factors affecting the adoption of mobile commerce in Singapore. *Telematics and Informatics*, *22*, 257–277. doi:10.1016/j.tele.2004.11.003

Yeo, A. (2001). Global software development lifecycle: An exploratory study. In *CHI 2001 Conference on Human Factors in Computing Systems* (pp. 104–111). ACM Press.

Yoon, S. (2002). The antecedents and consequences of trust in online-purchase decisions. *Journal of Interactive Marketing*, *16*, 47–63. doi:10.1002/dir.10008

York, J., & Pendharkar, P. (2004). Human-computer interaction issues for mobile computing in a variable work context. *International Journal of Human-Computer Studies*, *60*, 771–797. doi:10.1016/j.ijhcs.2003.07.004

Zhang, X., Keeling, K. K., & Pavur, R. J. (2000). Information quality of commercial web site homepages: An explorative analysis. In *Proceedings of International Conference on Information Systems, Brisbane, Australia* (pp. 164–175).

ENDNOTES

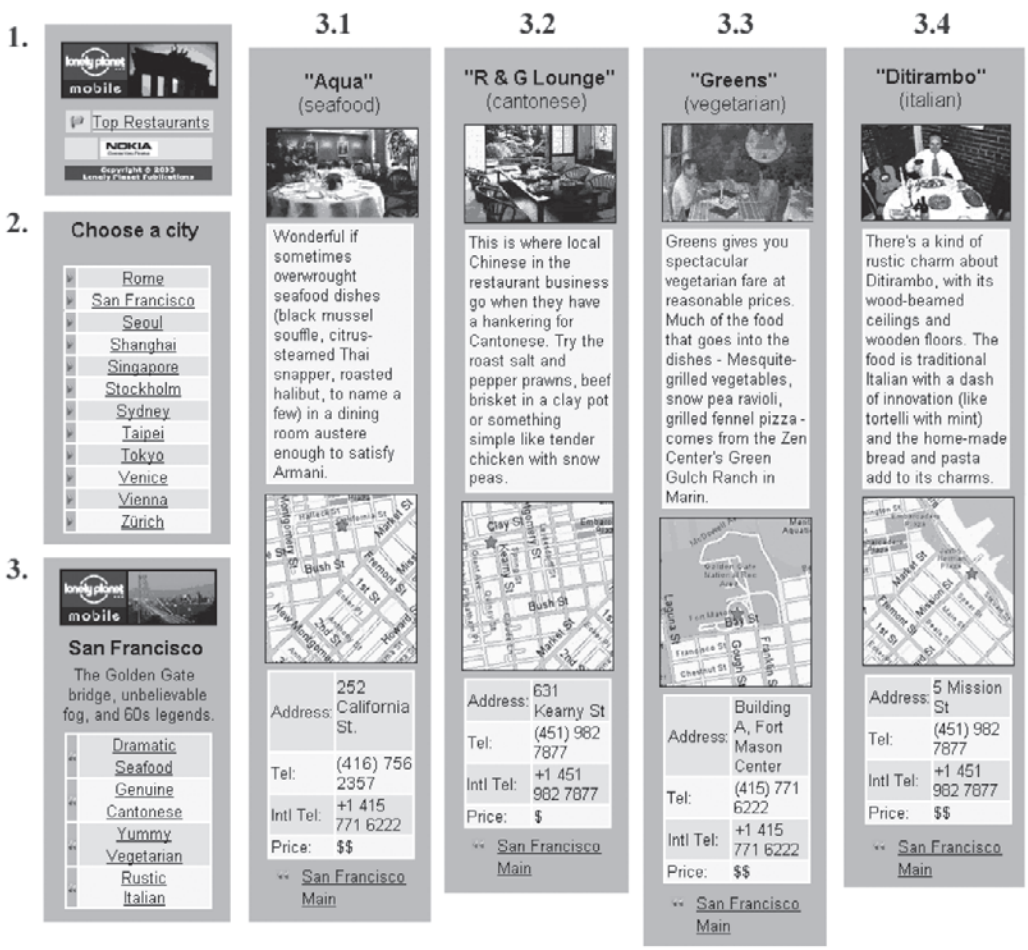
1 Throughout this paper the terms ‘portable’, ‘handheld’, ‘mobile’ and ‘small-screen’ refer to essentially the same type of device, typically called a ‘smart phone’. Also, the term ‘mobile Internet’ refers to accessing the Web directly on a small-screen, handheld, or mobile device that may be connected to the Internet.

2 The work by Geert Hofstede (1980) outlines 5 cultural dimensions: (1) Power distance - extent to which a society accepts unequal distributions of power in organizations and institutions. (2) Uncertainty avoidance - how societies accommodate high levels of uncertainty and ambiguity in the environment. (3) Masculinity-Femininity - in feminine societies there is an emphasis on quality of life and relationships; cultures that focus on material success and assertiveness are considered more masculine in orientation. (4) Individualism-Collectivism - in an individualist society individuals are expected to consider personal interests over interests of the group and individual decision-making is valued; in a collectivist culture the good of the group is more likely to be considered.

(5) Time Orientation - whether the focus in on short-term vs. long-term considerations. For a further elaboration of Hofstede's cultural dimensions, refer to Hofstede (1980) or Simon (2001).

APPENDIX A. SCREEN SHOTS

Figure 2.



English Version: Numbers indicate different pages, or decks

Figure 3.

Chinese version: Numbers indicate different pages, or decks

APPENDIX B. DESIGN SURVEY

Following are the statement used in the survey. Each was answered on a 7-point Likert scale from strongly disagree to strongly agree.*Information Design* [Sources: Cyr et al. 2005; 2004]

ID-1: I find the information logically presented.

ID-2: All service options, service attributes and restaurant information are well presented.

ID-3: I find the information to be well organized.

ID-4: The presentation of information is effective.*Navigation Design* [Sources: Cyr et al. 2005; 2004]

ND-1: This browser provides good navigation facilities to information content.

ND-2: I can easily navigate the CityGuide site.

ND-3: I find the CityGuide site easy to use.*Visual Design* [Sources: Cyr et al. 2005; 2004; van der Heijden 2003]

VD-1: The screen design (i.e. colors, boxes, menus, etc) is attractive.

VD-2: This site looks professionally designed.

VD-3: The graphics are meaningful.

VD-4: The overall look and feel of the site is visually appealing.*Satisfaction* [Sources: Cyr et al. 2005; 2004]

S-1: This site appeals to me emotionally.

S-2: This service completely fulfills my expectations.

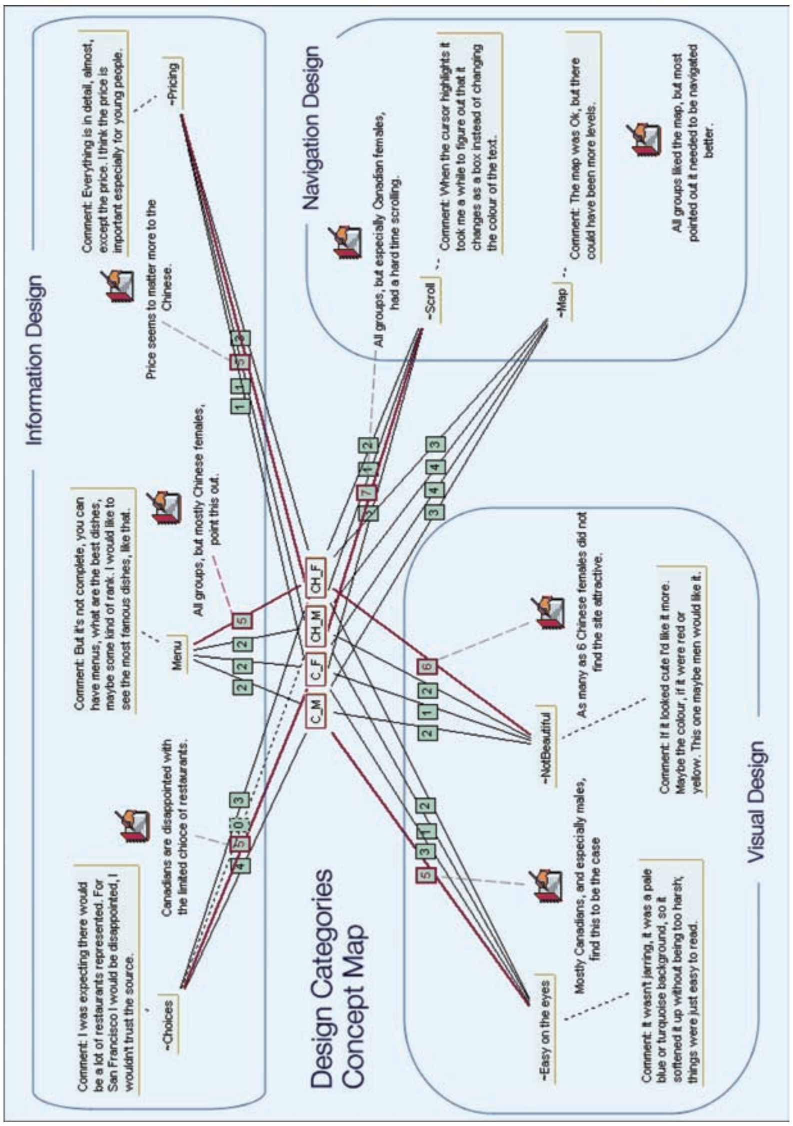
S-3: This service satisfies my needs well.

S-4: Using this service is satisfactory overall.

APPENDIX C: ATLAS.TI CONCEPT MAP

Note: C_M = Canadian male, C_F = Canadian female, CH_M = Chinese male, CH_F = Chinese female.

Figure 4.



This work was previously published in Mobile and Ubiquitous Commerce: Advanced E-Business Methods, edited by M. Head; E. Li, pp. 173-200, copyright 2009 by Information Science Reference (an imprint of IGI Global)

Chapter 2.2
Paralingual Web Design and Trust in E-Government

Roy H. Segovia
San Diego State University, USA

Murray E. Jennex
San Diego State University, USA

James Beatty
San Diego State University, USA

ABSTRACT

Can Web design improve the way governments serve their constituencies through the use of information technology [i.e., e-government]? This article proposes that the use of paralingual Web design can overcome possible trust issues in e-government with bilingual populations. An experiment was conducted where active e-government Web pages were converted to paralingual format and then site visitors were surveyed regarding their trust in the content and readability. The results of the experiment show that trust was improved for the minority language speakers, while the majority language speakers remained neutral with neither group indicating significant decrease in readability. These findings have important implications for societies with large bilingual or multilingual populations, where issues of trust among minority speakers and majority speakers may exist, as they indicate that paralingual Web design can help reduce these trust issues.

INTRODUCTION

Electronic government (e-government) is the use of Information and Communication Technologies (ICT), including the Internet, by government organizations to facilitate providing information and services to their constituents. E-government Web sites provide everything from basic information about governmental bodies and issues, to online services such as registering vehicles and applying for employment and for permits. More recent e-government services include econsultation, which is citizen participation and response to forthcoming consultations and decisions on matters of public interest (Jadu, 2005). The impetus to implement e-government can be attributed to cost control and improved service to citizens. Another driver is government's growing awareness of the need to attain more democratic governance (Coleman and Gotze, 2001; OECD, 2001), coupled with a widespread public interest in the potential of ICT to empower citizens and to increase government accountability and transparency (Hart-Teeter, 2003). An example is the United States E-Government initiative targets use of improved Internet-based technology to make it easier for citizens and businesses to interact with the government, save taxpayer dollars, and streamline citizen-to-government communications (USOMB, 2005). These many drivers make it likely that e-government will be a lasting ICT application leading e-government system designers to look for tools and methodologies that will ensure their acceptance and use by the intended users.

This article introduces one such potential design methodology, paralingual Web design, and uses an experiment to test this design methodology to see if it has potential for improving system acceptance and success. Paralingual is a Web design methodology for presenting information in more than one language. Paralingual Web design involves placing content in the desired languages but instead of having separate pages for each language as is common in a bi or multilingual Web design, the bi or multilingual content is placed side by side on the same page. The inspiration for this article is the trend towards localization in ecommerce systems and a concern that there may be a localization issue for e-government when the target population is bi or multilingual. Localization is defined by the Localization Industry Standards Association (LISA, 2008) as the process of modifying specific products or services for specific markets. In the case of e-government this involves tailoring e-government Web sites to fit the constituent market and in the case of a constituent market that speaks more than one language, allowing for these multiple languages. The concern driving this experiment is that there may be a trust issue affecting the success/adoption of e-government should these systems fail to take into account the bi or multi lingual aspects of their constituents.

A premise of information systems, IS, is that for an IS to be successful the intended system users must "use" the system where Rai et al. (2001) consider "use" to be the consumption of the outputs of the IS by the users as measured in terms such as frequency of use, amount of time of use, numbers of access to the IS, usage pattern, etc. General thinking is that the more an IS is used, the more successful the IS. Two of the more widely accepted IS success/acceptance models, the DeLone and McLean (1992 and 2003) IS Success Model and the Davis (1989) Technology Acceptance Model, TAM, incorporate "use" as a measure of success (DeLone and McLean) or successful adoption (TAM) through constructs such as Intent to Use, Perceived Usefulness, and Perceived Ease of Use.

Several authors (Gefen, et al., 2002; Tan, et al., 2005, Tan, et al., 2008, Warkentin, et al., 2002) suggest that use of e-government is influenced by the trust that potential users have with e-government. This article hypothesizes that this trust in e-government, and thus subsequent use, can be increased in bi and multi lingual societies by using paralingual Web design. This allows readers who are bilingual to easily see both versions and readily determine if the same information is being said in each version. It is expected that trust will be increased through this citizen validation process.

The contribution of this research is showing designers of e-government Web pages how the Web pages can be designed to improve trust in a bi-lingual constituency. While this research did not test this design approach in a multi-lingual environment it is expected this design can also be applied to e-government Web pages for these constituencies.

LITERATURE REVIEW

This article draws from three main bodies of literature, the trust, paralingual, and IS acceptance/success literatures. These literatures are summarized below and provide the theoretical foundation for the article. The trust literature is presented first as it provides the issue of concern for the article. The paralingual literature is second to provide the background for why the proposed design methodology is a good solution for the trust issue. The IS acceptance/success literature is provided third as it helps provide the framing for the experiment.

Trust can be defined as "the subjective assessment of one party [trustor] that another party [trustee] will perform a particular transaction according to his or her confident expectations, in an environment characterized by uncertainty" (Ba and Pavlou, 2002, p. 245.) For e-government this means users trusting that the e-government service is providing correct information, that data will be protected, and that transactions will be conducted in a secure manner and recorded appropriately.

Trust in government has historically been problematic in the United States of America as constituent citizens are known to have a high level of distrust in their governing bodies. Trust in government has been declining for more than three decades and has been the topic of a substantial amount of research in political science (Levi and Stoker, 2000; Hibbing and Theiss-Morse, 2002). In the state of California, a recent study exposed an unexpectedly high level of distrust in government by California citizens. During 2004, a series of dialog-oriented seminars were held by Viewpoint Learning in various locations in California. One of the seven major findings of the study was that an underlying issue was profound mistrust of government and elected officials. Furthermore, this mistrust was more intense and persistent than expected, outstripping the levels that have been measured by polls and focus groups (Rosell, Gantwerk, and Furth, 2005).

In addition to the trust issues above, there are known issues with trust in e-government Web sites. This is clearly the effect of the general mistrust by citizens in their government bodies, as mentioned previously. The principal reason given for mistrust of the Web is an artifact of the internet itself. Namely, the internet is now perceived to be beyond the control of the hosts and providers in terms of security and trust. Despite the use of lock icons, digital signatures, passwords, privacy policy statements, and other security techniques, internet users feel that hosts and providers have lost control of the digital data transport medium as well as the software infrastructure that supports it, impeding the growth of e-government (Mercuri, 2005.) To counter this, the International Telecommunication Union (ITU) is providing support for national e-government projects including enhancing security and trust in the use of public networks (Khalil-babnet, 2005).

Improving trust in government and e-government is a critical issue. In the study by Viewpoint Learning, citizens voiced a strong desire to find constructive solutions to problems facing the state (Rosell, Gantwerk, and Furth, 2005). In a geographical area with a high proportion of bilingual speakers, usage of e-government Web sites may be improved in the same way as has been shown effective in electronic commerce (ecommerce). That is, with regard to language issues, researchers have found that customers are far more likely to buy products and services from Web sites in their own language, even if they can read English well. Furthermore, attention to site visitors' needs should be an important consideration in Web design because such attention can help a site build trust with customers (Schneider, 2003). Gassert (2004) suggests building trust through knowledge by using ICT for better education and information. Additionally, LaVoy (2001) supports the use of e-government as a way of improving trust by improving accountability. Finally, Gefen, et al. (2002) view trust in government as the main driver for e-government adoption. Their analysis show data privacy concerns create the biggest barrier to adoption of e-government. While this form of e-government service, online tax service, consists of real transactions, the trust issue dealt with in this research comes even before citizens attempt such transactions. Namely, the citizens must be given a reason to simply trust in the information that is on a government Web site.

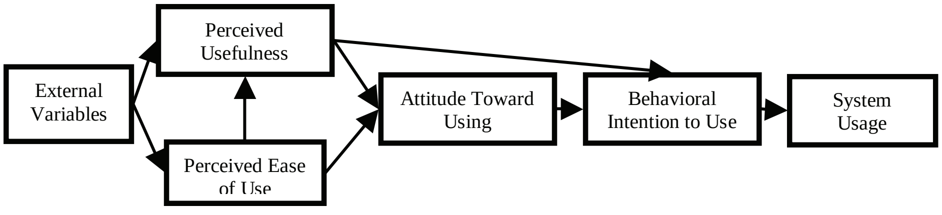
One of the earliest known examples of written multi-language information is the Rosetta Stone. According to Wikipedia (2006a), this archaeological artifact is a granite stone with writing in three different written scripts dated to about 200 BCE (Before Common Era, essentially the same as BC). It contains Greek, Demotic Egyptian, and Egyptian hieroglyphics. The message in the three scripts is the same and is a decree by the Egyptian ruler Ptolemy V regarding taxes and temple construction. The purpose of having the message written in three languages adjacent to each other was to solve a difficult linguistic problem..

A contemporary example of multi-language dissemination of important information is a requirement in the California Labor Code about employers posting worker information in English and Spanish. According to the Labor Code, several important documents, such as the Notice of Workers Compensation Coverage, must be posted in Spanish and English whenever there are employees of Spanish descent (California Labor Code, 2006.) There are similar requirements regarding Minimum Wage, Pregnancy Disability Leave and the California Family Rights Act (OSHA4LESS.COM, 2006)

The term “paralingual” is used to define the layout of information using two sets of text in different languages on the same page, such as in a Web page. The term was coined as an extension of the word “bilingual.” Para is a Greek prefix that means beside, near, or alongside (Wikipedia, 2006). Therefore, paralingual refers to two languages adjacent to each other on the same page. Paralingual Web pages are almost non-existent on the Web. Although many Web sites are now multi-language Web sites, the common layout for these Web sites is to separate the languages to separate pages. This commonality is reflected in the standards on localization (LISA, 2008.). This localization can be found in e-government. Cunliffe et al (2002) reports a case study of developing a bilingual Web site in English and Welsh for users in Wales. This study focused on Web site design for just two languages and recognized that there are many bilingual areas in the world (Cunliffe et al, 2002). One of the most important aspects of designing for bilingual Web site content is to provide rich interconnectivity between materials in the two languages (Cunliffe et al, 2002). While not the main focus of the study, Cunliffe, et al. (2002) does discuss the options for placement of two languages on the same page.

Paralingual Web design is expected to affect e-government use two ways: increased use due to increased trust and possible decrease in use due to impacts on ease of use and usefulness. As

Figure 1. Technology acceptance model (Davis, 1989)



discussed above it is expected to help improve trust. The two models mentioned in the Introduction help to predict the probable impact of paralingual Web design. TAM was developed by Davis (1989) as an explanation of the general case determinants of computer acceptance that are capable of explaining user behavior across a broad range of systems, technologies, and user populations. The model includes use as a determinant but indicates that use is determined by ease of use and perceived usefulness, attitude, and intention to use (see Figure 1). TAM is a derivative of Fishbein and Ajzen's (1975) Theory of Reasoned Action (TRA) model. TRA focuses on situation specific personal beliefs and attitudes, and the effects of the beliefs of others who can influence the individual. The fundamental premise of TRA is that individuals will adopt a specific behavior if they perceive it will lead to positive outcomes (Compeau and Higgins, 2001). However, adoption is also influenced by two factors, Perceived Usefulness and Perceived Ease of Use. Perceived Usefulness reflects that an individual's perception of usefulness influences their intention to use the technology primarily through the creation of a positive attitude. This is consistent with the TRA, which holds that attitude (an individual's positive or negative feelings about performing a behavior) influence behavioral intention. Geffen, et al. (2002) found that trust impacts Perceived Usefulness with increased trust improving Perceived Usefulness. Perceived Ease of Use reflects the user's assessment of how easy a system is to learn and use. TAM includes ease of use as a separate belief construct based on the concept of self-efficacy (an individual's judgment of his/her ability to organize and execute tasks necessary to perform a behavior). Ultimately, TAM predicts that if paralingual Web design can improve trust with the users while not reducing perceptions of ease of use or usefulness then it can be expected that paralingual Web design will be accepted by users.

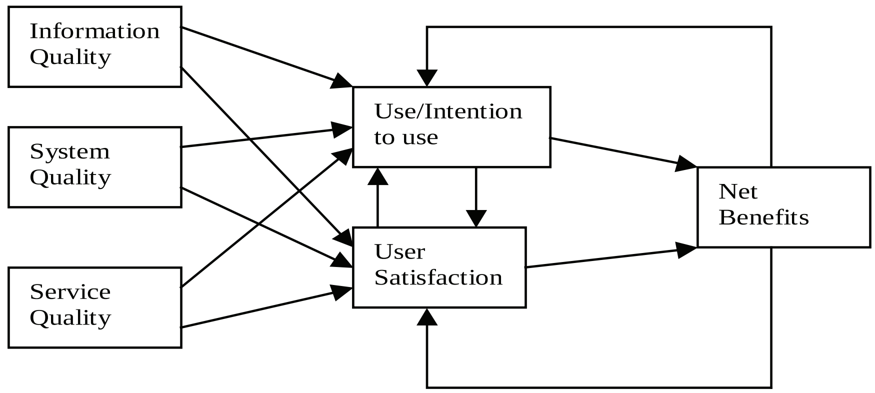
DeLone and McLean (2003) revision of the IS Success Model also helps predict the impact of paralingual Web design. This is a causation model that implies that system quality, information quality, and service quality will lead to increased use or increased intent to use which will lead to benefits and success of the system (see Figure 2.) The intent to use construct is important for this article as it is similar to the ease of use and usefulness constructs from TAM, especially when intent to use is operationalized using the Perceived Benefit Model (Thompson, Higgins, and Howell, 1991). Additionally, trust is reflected in the information quality dimension as users must be able to trust the information in the system for there to be quality. Paralingual Web design provides the trust in the information quality and aids in perceived usefulness in intent to use. Ultimately, the IS Success Model predicts that improved trust will help improve information quality which will increase intent to use/actual use leading to benefits and system success.

The conclusion from the literature review is that Paralingual Web design may be a design tool that will support building trust in content and process by e-government users. As mentioned earlier, Geffen, et al. (2002) consider this essential for e-government adoption as it support perceived usefulness. The only concern is an impact to perceived ease of use resulting in the experiment design discussed in the next section.

RESEARCH METHODOLOGY

This study sought to determine whether paralingual Web site design can improve trust in e-government Web sites while maintaining ease of use and usefulness for both the reader and the provider government. In a bi or multilingual environment, the ability to communicate concepts across diverse cultures and languages has become increasingly important, especially when issues of trust are involved. Further, e-government has

Figure 2. DeLone and McLean's (2003) Revisited IS success model



rapidly increased in usage, making it even more essential for Web designers to conscientiously strive to ensure that concepts have the same meaning across cultures. This study focuses on paralingaul issues in a highly bilingual populated location in the United States, that of San Diego and Tijuana, with a total combined population of over 5 million as of 2004 and over 64 million annual border crossings. The subject municipality is National City, a city of 54,260 (2000 census) located approximately 10 miles from the United States border with Mexico. The municipality population is 60% Hispanic/Latino (all or part), 39% white (all or part), and 30% other (all or part where other includes Asian, Hawaiian/Pacific Islander, Black, and American Indian) (Note that the sum is greater than 100% due to respondents reporting belonging to more than one race) (National City, 2008). Actual percentage of English-Spanish bilingual residents is not reported but is understood to be large. The subject municipality was selected because of its expected bilingual population and its willingness to participate in the experiment. It should be noted that simply because the municipality population is heavily Hispanic or Latino does not mean that the local language is predominately Spanish. This is a US city and the citizens are US citizens. English is the majority language spoken in National City as evidenced by the National City Web site being in English only. However, due to its proximity to the border and its large Hispanic/Latino population it is reasonable to assume there is a substantial bilingual population.

To create the experiment, three informational Web pages were converted to paralingual format consisting of English and Spanish text placed horizontally adjacent to each other. Informational pages were chosen as the National City Web site is primarily informational and does not perform any financial transactions. This is considered acceptable as the majority of e-government Web sites in 2004 were informational rather than financial transaction focused (ICMA, 2004). Translation of the English content in the original selected Web pages was performed taking into account the following:

- Variations in the style and vocabulary of Spanish. A style to reflect the local style was chosen. This style is mostly Mexican in its structure and vocabulary, so these were kept in perspective at all times.
- The level of writing was kept at approximately a high school level of comprehension (the same as the English version).
- The Spanish translation was written to conform strictly to correct language structure, syntax, and spelling. This is demonstrated by the correct application of diacritical marks, such as accents, tildes, and umlauts.

The translation task was performed by a native Spanish speaker with professional training and experience as a translator. The translated content was then evaluated and modified by a Spanish language professor with a specialization in translation studies.

A survey was generated using survey monkey to gauge the opinions of visitors to the experimental Web pages. The survey consisted of eight items. Items 1-4 dealt with demographics of the respondents and were asked at the request of National City:

1. What is your age range? (18-24, 25-34, 35-44, 45-54, 55-64, 65 or over)
2. Are you a resident of National City? (Yes, No)
3. What language or languages do you use for communication (speaking, reading, writing)? (English only, Spanish only, mostly English with some Spanish, about half English and half Spanish, mostly Spanish with some English, some other combination of languages) (Note that this question was used to group responses)
4. Have you visited the National City official Web site before now? (Yes, No)

Items 5 and 6 operationalized trust. Item 5 queries improved trust due to the information on the Web page while item 6 queries improved trust due to the paralingual format. Two items were used to ensure that improvement in trust was due to paralingual format and not to just reading the information (both use a 7 point Likert scale with 1 being Strongly Disagree and 7 being Strongly Agree):

5. Please respond to this statement: I have a greater trust now than before in my understanding of the National City Web site.
6. Please respond to this statement: I have a greater trust now than before in the information on the National City Web site because it is in English and Spanish side by side.

The seventh item was also asked at the request of National City and queried how aware the respondent was of the multilingual nature of National City:

7. Are you aware that other residents in National City speak more than one language? (Yes, No, Not sure)

The final item operationalized ease of use and usefulness by querying perceptions of site readability. Only readability is measured as no transactions are performed with the paralingual Web sites:

8. Please respond to this statement: It was easy for me to read the pages with English and Spanish next to each other: (No, it was very difficult, No, it was somewhat difficult, It was neither difficult nor easy, Yes, it was somewhat easy, Yes, it was very easy)

Municipality officials encouraged participation by constituents and residents in the vicinity through a series of bilingual public announcements encouraging individuals to visit the modified Web pages and to complete a brief survey documenting their opinions on the Web site. The respondents to the survey had the choice of filling out the survey in English or Spanish, presumably selecting their primary language of communication. The respondents who answered the online surveys represent a portion of Web users who visited the National City Web site and chose to view at least one of the paralingual pages. This constitutes a self-selected sample. The visitors to the Web site may have been responding to one of the various methods of advertisement of the research, or may have been incidental visitors to the Web site who then decided to follow a link from the home page to one of the paralingual pages. While there are problems with self selected samples, it is acceptable for this experiment as the goal was to get real users to respond to the Web pages and the inherent difficulties in recruiting a random sample on a public Web page. Data was collected over a three month period with 133 responses being collected. Data were grouped based on the respondent's answer to the question asked about the respondent's primary language for communication. The range of response choices was from "English only" to "Spanish only"; an additional choice was "some other combination of languages." In order to generate analytical results that are more representative of the language component of the sample data, the English sample data included only those who indicated "English only," "Mostly English," or "some other combination of languages." The Spanish sample data included those that indicated "Spanish only," "mostly Spanish," or "half English and half Spanish." This resulted in 97 English responses and 36 Spanish responses.

Statistical data analysis was used to analyze the collected data. Beatty (2000), Siegel (1956), McClave, Benson, and Sincich (2008), Jaccard and Becker (1990), and others have stressed the importance of selecting appropriate statistical procedures that correspond to the methods used to assign numerical values, as well as the type and level of data being analyzed. Since the questionnaires employed in the current study are based on a traditional Likert format, with anchor points ranging from "strongly disagree" to "strongly agree," nonparametric statistics were deemed appropriate. The nonparametric tools used include the Mann-Whitney U test, the Wilcoxon T test, and Spearman's rho. The Mann-Whitney U test is an appropriate tool for comparing central tendencies between *independent* responses to Question 5 and 6 *across* the two language groups, while the Wilcoxon matched pairs T test is an appropriate tool for comparing central tendencies across the *matched* pairs of data. With regard to tests of association, Spearman's rho is an appropriate measure for examining the correlation between matched responses to Questions 5 and 6 *within* each language group. When deemed appropriate and meaningful, means and standard deviations have been calculated for various data sets as well. All statistical tests were performed via SPSS and validated using other software. All results are reported in the format recommended by Jaccard and Becker (1990).

All collected surveys were used for the analysis. There was no exclusion or limitation of surveys based on any criteria. However, it should be noted that some respondents did not fill out all survey questions. For example, a small portion of both English and Spanish respondents did not fill out the second half of the surveys. Thus, the data used in different components of the analysis consist of different sample sizes, depending on the questions being examined. The software used for these analytical tests accounted for the missing data and thus calculated test values based only on the number of questions actually answered.

RESULTS AND DISCUSSION

For all tests, the alpha value was chosen to be $\alpha = 0.05$. For ease and uniformity of comparison between all tests, the z value will be used. For the chosen $\alpha = 0.05$ and 2-tailed tests, the critical z value is $z = \pm 1.96$.

Mann-Whitney U on Grouping by Language Choice

The Mann-Whitney U test was used to compare the central tendencies of responses to item 5: I have a greater trust now than before in my understanding of the National City Web site, based on the 97 English-based responses and the 36 Spanish-based responses. This item was designed to measure differences between respondents' improvement of trust in their *understanding* of the information on the paralingual page only. The Mann-Whitney

U and its corresponding *z*-value (-1.907) resulted in no statistical difference in perceptions between the English and Spanish respondents.

A second Mann-Whitney U test was used to compare the central tendencies of responses to item 6: I have a greater trust now than before in the information on the National City Web site because it is in English and Spanish side by side, again based on the 97 English-based responses and the 36 Spanish-based responses. This item was designed to measure differences between the respondents' improvement of trust in regard to the *information* on the paralingual page, based on having the information in the two languages on the same page. This test yielded a *z*-value of -4.406 ($p < .01$), indicating that there were significant differences between the responses of the English and Spanish respondents with regard to this question.

These results partially support the hypothesis that use of paralingual Web pages will increase trust in the page content and government sponsor. Since trust levels are different only when the paralingual information is considered, and the Spanish respondents showed higher medians than the English respondents in both items 5 and 6, the Spanish respondents show an increased level of trust based on the paralingual content.

Wilcoxon T Test on Grouping by Language Choice

The Wilcoxon matched-pairs T test was used to compare the central tendencies on data sets similar to those tested with the Mann-Whitney U test, items 5 and 6. The calculated z value for English items 5 and 6 is $z = -3.188$. This indicates that there is a statistical difference in the means of the answers to each item. For Spanish items 5 and 6, the calculated z value is $z = -1.034$ and we fail to reject the null hypothesis, therefore we conclude that there is no statistical difference in the means of the answers to each item.

This result seems to indicate that Spanish speakers improved their trust because the text was in Spanish, this does not necessarily support that a paralingual Web design is necessary to improve trust and that simply being bi-lingual in some format will increase trust of the minority speakers.

Spearman's rho on Grouping by Language Choice

Spearman's rho was used to examine the relationship between responses on English item 5 (n=97) and 6 (n=97) to determine how well the English respondents' answers correlated between the two items regarding trust. The test measures the level of correlation between these two items. The two-tailed test yielded $\rho = 0.504$ ($p < .01$), which was statistically significant.

Spearman's rho was also calculated for Spanish items 5 (n=36) and 6 (n=36). The calculated correlation, $\rho = 0.563$ ($p<.01$), is significant for a 2-tailed test as calculated by SPSS.

Spearman's rho was performed to see the degree of correlation between items 5 and 6 in each group. The rho value for English items 5 and 6 is 0.504, so the correlation is "significant". For Spanish items 5 and 6 the rho value is 0.563, also significant. Thus each of the two groups answered consistently in the trust questions.

Readability Reflecting Ease of Use and Usefulness of the Paralingual Web Pages

Item 8 of the survey is a measure of the readability as reflecting ease of use and usefulness of the paralingual pages. The choices for responses had a range of five from "it was very difficult" to "it was very easy." The percentages of the English respondents (n=97) who answered "it was somewhat easy" or "it was very easy" is 61.3 % and the percentage of Spanish respondents (n=36) answering similarly is 85.7%. This is interpreted to imply that having the page in paralingual format did not diminish readability and thus ease of use and usefulness, important to predicting acceptance of paralingual Web pages by users. However, it is also implied that English respondents were more likely to find an impact to ease of use and usefulness. This may suggest that these respondents will be less accepting of a change to paralingual format as they perceive paralingual format to be less useful.

Alternative Analytical Calculations

Performing nonparametric tests is the appropriate method for analysis of ordinal data. These results have been shown in the previous section. However, means and standard deviations are more commonly understood and therefore used more commonly to describe data. Table 1 is a summary of the means and standard deviations for items 5, 6, and 8.

CONCLUSION

This article is primarily intended to provide evidence to support government decision makers, e-government researchers, and e-government Web designers in applying paralingual Web page design for improving trust in government in regions where there is a high proportion of bilingual residents. An experiment was performed to test the hypothesis that paralingual Web design will improve trust in the content of the e-government Web page without significantly affecting ease of use and usefulness. This hypothesis was confirmed, but not quite as expected. It was found that the paralingual format improved trust for the minority speaker but not the majority speaker. Upon reflection this is an expected finding.

An additional finding with respect to ease of use and usefulness was noted. It was found that in general respondents did not find the paralingual format hard to read, however, it was noted that the majority speakers (English) were less enthusiastic than the minority (Spanish) speakers about the paralingual Web design. This finding has implications for the adoption of paralingual Web design in that it may show that there will be resistance to adoption of a paralingual Web approach by the majority speakers. The implication to policy makers is that there needs to be additional research done with respect of citizen attitudes towards bilingual government prior to implementing a paralingual Web strategy. This is particularly important in different regions of the United States where it could be expected that the generally favorable bilingual acceptance attitudes found in California may not exist.

The conclusion of this article is that paralingual Web design is useful for e-government in areas with significant bilingual populations. However, there are limitations to this approach as it appears that there is a risk of backlash and rejection from the majority speaking population. The implication for policy makers is that paralingual Web design should be used when there are known trust issues between majority and minority speakers that translate into trust issues with government and e-government initiatives.

Table 1. Means and standard deviations of survey data

Item	Value Range	English: Mean/StdDev	Spanish: Mean/StdDev
5	1 - 7	4.2209/1.785	4.8182/1.286
6	1 - 7	3.5930/1.811	5.0606/1.540
8	1 - 5	3.88/1.259	4.366/0.994

Limitations

This experiment has a small sample with limited items testing improved trust and ease of use and usefulness. The conclusion that paralingual Web design improved trust for the minority Spanish speakers is supported by the statistical analysis but it cannot be totally discounted that trust may have been increased simply because the content was in Spanish. Additionally, this experiment only looked at informational Web pages and the conclusions may not apply to financial or other transactional Web pages. Finally, only one city was looked at, one minority language used (Spanish,) and the sample population was self selected meaning that the results may not be reflective of all populations, cultures, and languages.

Areas for Future Research

There are several areas for future research of which the first are those areas that address the limitations to this research. This includes further studies using other languages and locations; obtaining a large sample size; and using transaction based Web pages in addition to informational pages. It is expected that there may be ease of use and usefulness issues associated with paralingual Web design used for transactional Web pages which could affect adoption of the pages and requiring that the trust improvement from paralingual Web design be balanced against ease of use and usefulness.

An additional area for future research is in using paralingual Web design in a multilingual environment. In this case multilingual implies more than two languages and is a topic of relevance in many countries in Europe and Asia (Finland and Switzerland are two examples). While there are historical examples of multilingual designs, none are in Web page layouts (the Rosetta Stone is an example of a three language representation). Given the limitations on screen areas, especially for mobile and/or handheld devices, a paralingual Web design involving three or more languages may not be practical or may induce substantial ease of use and usefulness issues. Research needs to be done to see if this is a practical approach or if more traditional approaches of having multiple versions of the same Web page each in a different language is a preferable approach.

A final area for future research is in e-government policy. The conclusion that majority speakers may resist paralingual implementations is very important to policy makers. Research into what factors may influence majority speakers to accept and adopt paralingual implementations or which factors may influence majority speakers to reject paralingual implementations is critical to policy makers for crafting appropriate e-government strategies and policies. It is expected that paralingual Web design will only be appropriate in regions where there is significant lack of trust by the minority speakers but this needs to be confirmed.

REFERENCES

2005 The Year of Languages, (2005). Special project of the American Council on the Teaching of Foreign Languages (ACTFL). Available from http://www.yearoflanguages.org/i4a/pages/index.cfm?pageid=3419 accessed March 21, 2006.

Ba, S. and Pavlou, P.A., (2002). Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior, MIS Quarterly, (26:3), 2002, pp. 243-268.

Beatty, J. R. (2000). Statistical Methods, Sixth Edition. McGraw-Hill, New York.

Becerra, H., (2006). Welcome to Maywood, Where Roads Open Up for Immigrants. Los Angeles Times, March 21, 2006, p. A-1.

Bharat, K. and Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. In 7th WWW,.

California Labor Code, Section 3550(d). Accessed on March 24, 2006 from http://www.leginfo.ca.gov/cgi-bin/displaycode?section=lab&group=03001-04000&file=3550-3553.

Compeau, D. and Higgins, C., (1995). Computer Self-Efficacy: Development Of A Measure And Initial Test. MIS Quarterly, 19(2), pp. 189-211.

Coleman, S., and Gøtze, J, (2001). Bowling Together: Online Public Engagement in Policy Deliberation. Hansard Society, London. Retrieved October 4, 2005 from http://bowlingtogether.net.

Castelli, W.A., Huelke, D.F., and Celis, A., (1969). Some basic anatomic features in paralingual space. Oral Surgery, Oral Medicine, And Oral Pathology, 27(5), pp. 613-621.

Cunliffe, D., Jones, H., Jarvis, M., Egan, K., Huws, R., and Munro, S., (2002). Information Architecture for Bilingual Web Sites. Journal of the American Society for Information Science and Technology, 53(10), p. 866.

Davis, F., (1989). Perceived Usefulness, Perceived Ease Of Use, And User Acceptance Of Information Technology. MIS Quarterly, 13, pp. 319-339.

DeLone, W.H. and McLean, E.R., (1992). Information Systems Success: The Quest for the Dependent Variable. Information Systems Research, 3, pp. 60-95.

DeLone, W.H. and McLean, E.R., (2003). The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. Journal of Management Information Systems, 19(4), pp. 9-30.

Fishbein, M., and Ajzen, I., (1975). Belief, Attitude, Intention And Behavior: An Introduction To Theory And Research. Addison-Wesley, Reading, MA.

Gassert, H. (2004). "How to Make Citizenz Trust E-Government," University of Fribourg, E-Government Seminar, Information Systems Research Group. Available from http://edu.mediagonal.ch/unifr/egov-trust/slides/html/title.html accessed November 29, 2005.

Gefen, D., Warkentin, M., Pavlou, P.A., and Rose, G.M., (2002). EGovernment Adoption. Eighth Americas Conference on Information Systems, Association for Information Systems.

Girion, L., (2006). Language Becoming an Issue for Health Insurers. Los Angeles Times, March 20, 2006, p. C-1.

Grefenstette, G and Nioche, J., (2000). Estimation of English and non-English Language Use on the WWW. Xerox Research Centre Europe. Available from http://arxiv.org/pdf/cs.CL/0006032 accessed March 19, 2006.

Grönlund, A. and Horan, T. A. (2004). Introducing E-Gov: History, Definitions, and Issues. Communications of the Association for Information Systems, 15: 713-729.

Gulli, A. and Signorini, A., (2005). The Indexable Web is More than 11.5 Billion Pages. WWW 2005, May 10–14, 2005, Chiba, Japan.

Hart-Teeter, (2003). The New E-government Equation: Ease, Engagement, Privacy and Protection. A report prepared for the Council for Excellence in Government, Retrieved November 27, 2005 from http://www.excelgov.org/usermedia/images/uploads/PDFs/egovpoll2003.pdf.

Hibbing, J.R. and Theiss-Morse, E. (2002). Stealth Democracy: Americans' Beliefs About How Government Should Work. Cambridge University Press, Cambridge.

International City/County Management Association (ICMA) (2004). Electronic Government Survey 2004. accessed February 22, 2008 from: http://icma.org/upload/bc/attach/%7B9BA2A963-DDCC-40B7-836D-F1CFC17DCD98%7Degov2004Web.pdf,

Jaccard, J. and Becker, M.A., (1990). Statistics for the Behavioral Sciences. Wadsworth Publishing

Company, Belmont, CA, USA.

Jadu, (2005). Enterprise Content Management for Public and Private Sector. Available from http://www.jadu.co.uk/egov/jadu_egov_econsultation.php accessed November 26, 2005.

Khalil-babnet, M. WSIS Prepcom-2: Cybersecurity an Issue for All. Available from http://www.babnet.net/en_detail.asp?id=935 accessed November 29, 2005.

La Prensa-San Diego. Available from http://laprensa-sandiego.org/rates/rates.html accessed November 27, 2005.

LaVoy, D.J., (2001) Trust and Reliability. Public Manager, Fall 2001, 30(3), p 8.

Levi, M. and Stoker, L., (2000). Political trust and trustworthiness. Annual Review of Political Science 3: 475-507.

Localization Industry Standards Association, LISA, (2008). Localization. Accessed February 23, 2008 from http://www.lisa.org/Localization.61.0.html.

Macias, E. and Temkin, E., (2005). Trends And Impact Of Broadband In The Latino Community. Tomás Rivera Policy Institute, 2005. Available from http://www.trpi.org/PDFs/broadband.pdf accessed March 21, 2006.

Mercuri, R.T. (2005). Trusting in Transparency. Association for Computing Machinery. Communications of the ACM, 48(5), p. 15.

McClave, J.T., Benson, P.G., and Sincich, T., (2008). Statistics for Business & Economics, 10/E. Prentice Hall, Upper Saddle River, New Jersey, USA.

Moller, R.M., (2000). Profile of California Computer and Internet Users. California Research Bureau. California State Library. Available from http://www.library.ca.gov/crb/00/01/00-002.pdf accessed April 9, 2006.

National City, (2008). National City Web Site. Available from http://www.ci.national-city.ca.us/main.asp accessed February 22, 2008.

National Performance Review, (1993). From Red Tape To Results: Creating A Government That Works Better And Costs Less. Washington, D.C.: Government Printing Office.

Organization for Economic Co-operation and Development (OECD), (2001). Citizens as Partners, Information, Consultation and Public Participation in Policy-Making.

Osborne, D. and Gaebler, T., (1992). Reinventing Government: How The Entrepreneurial Spirit Is Transforming The Public Sector. Reading, MA: Addison-Wesley.

OSHA4LESS.COM, "Sales Book - lew 2.pdf", 2006. Received as attachment to electronic mail message, March 24, 2006.

Paralinguism in the Theatres and the International Theatre Festivals. Sasho Ognenovski. Intercultural Communication, ISSN 1404-1634, 1999, August, issue 1. Available from http://www.immi.se/intercultural/ accessed November 19, 2005.

Rai, A., Lang, S., and Welker, R., (2002). Assessing the Validity of IS Success Models: An Empirical Test and Theoretical Analysis. Information Systems Research, 13(1), pp. 50-69.

Rosell, S., Gantwerk, H., and Furth, I., (2005). Listening To Californians: Bridging The Disconnect. Viewpoint Learning, Inc. Retrieved January 15, 2006 from http://www.viewpointlearning.com/pdf/HI_Report_FINAL.pdf.

Sandoval, V.A. and Adams, S.H., (2001). Subtle Skills for Building Rapport: Using Neuro-Linguistic Programming in the Interview Room. Available from http://www.fbi.gov/publications/

leb/2001/august2001/aug01p1.htm accessed November 19, 2005.

Schneider, G.P., (2003). Electronic Commerce, Fourth Annual Edition. Thomson Course Technology, Boston.

Secretariat, Treasury Board of Canada, (2006). Communications Policy of the Government of Canada. Available from http://www.tbs-sct.gc.ca/pubs_pol/sipubs/comm/comm1_e.asp#04 accessed March 21, 2006.

Siegel, S., (1956). Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill Book Co., New York.

Tan, C.W., Pan, S.L., and Lim, E.T.K., (2005). Towards the Restoration of Public Trust in Electronic Governments: A Case Study of the E-Filing System in Singapore", Proceedings of the 38th Annual Hawaii International Conference on System Sciences, IEEE Computer Society Press,

Tan, C.-W., Benbasat, I., and Cenfetelli, R.T., (2008). Building Citizen Trust towards e-Government Services: Do High Quality Web sites Matter? Proceedings of the 41st Hawaii International Conference on System Sciences, IEEE Computer Society Press.

Thompson, R.L., Higgins, C.A., and Howell, J.M., (1991). Personal Computing: Toward a Conceptual Model of Utilization. MIS Quarterly, March, pp. 125-143.

United States Census Bureau, Census 2000. "QT-PL. Race, Hispanic or Latino, and Age: 2000. Data Set: Census 2000 Redistricting Data (Public Law 94-171) Summary File Geographic Area: National City city, California." Available from http://factfinder.census.gov/servlet/QTTable?_bm=y&-geo_id=16000US0650398&-qr_name=DEC_2000_PL_U_QTPL&-ds_name=D&-_lang=en&-redoLog=false, accessed March 2, 2006.

United States Department of Labor, The Family and Medical Leave Act of 1993, 29 CFR 825.300, 2006. Available from http://www.dol.gov/dol/allcfr/ESA/Title_29/Part_825/29CFR825.300.htm accessed March 24, 2006.

United States Office of Management and Budget (USOMB), (2005), E-Gov: Powering America's Future With Technology. Retrieved October 5, 2005 from http://www.whitehouse.gov/omb/egov/index.html.

Warkentin, M., Gefen, D., Pavlou, P.A., and Rose, M., (2002). Encouraging Citizen Adoption of e-Government by Building Trust. Electronic Markets, 12(3), 157-162.

Wikipedia, (2006). Paralanguage. Available from http://en.wikipedia.org/wiki/Paralanguage#Linguistics accessed September 19, 2006.

Wikipedia, (2006a). Rosetta Stone. Available from http://en.wikipedia.org/wiki/Rosetta_stone

This work was previously published in International Journal of Electronic Government Research, Vol. 5, Issue 1, edited by V. Weerakkody, pp. 36-49, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 2.3
Designing Medical Research Web Sites

Jonathan P. Grady
University of Pittsburgh, USA

Michael B. Spring
University of Pittsburgh, USA

Armando J. Rotondi
University of Pittsburgh, USA

ABSTRACT

This chapter discusses the design of Web sites to be used as the basis of medical research. It is broken down into three sections: Part 1 discusses the various issues that have to be addressed in the design of a Web site that will be used to assess some intervention based on the Web site. Part 2 discusses the design of such a Web site and the development of a tool to facilitate this process. Part 3 presents the results of preliminary usability analysis for the tool to assist medical researchers in constructing Web sites that can meet the needs and requirements of medical intervention studies. The results of the preliminary interviews, prototype walkthroughs, and preliminary usability studies are presented laying the groundwork for future development and more formal usability studies.

DOI: 10.4018/978-1-60566-356-2.ch053

INTRODUCTION

The Internet, specifically the World Wide Web, is being tested by the medical community as a potentially more efficient and effective mechanism to serve the needs of patients and caregivers. The Web can be used to quickly and cost-effectively disseminate information, collect data and target communications to specific people or groups of people. May et al., (2005) describe a range of efforts in the United Kingdom to provide services via telemedicine. In the United States, researchers have begun to study the efficacy of web-based medical interventions. Studies have been performed in recent years examining the effectiveness of internet-based interventions with patients suffering from a variety of medical conditions, including heart transplants (Dew et al., 2004), smoking cessation (Lenert et al., 2003), osteoarthritis (McAlindon, Formica, Kabbara, LaValley, & Lehmer, 2003), clinical depression (Christensen, Griffiths, & Jorm, 2004), *traumatic brain injury (*Rotondi, Sinkule, & Spring, 2005), and schizophrenia (Rotondi et al., 2007). While these studies had different degrees of success achieving their goals, they generally concluded that 1) the Internet is a potentially useful tool for medical interventions, and 2) the success of an Internet-delivered treatment depends largely on the design of the website and other electronic communications used in conjunction with the treatment.

The cost to design, implement and maintain a website to be used as the basis for medical intervention research can be high. Projects at the University of Pittsburgh by Rotondi and Spring budgeted equipment, staff and maintenance costs for a medical research website well in excess of $100,000. Marshall and Haley (2000) report the cost of setting up the site infrastructure alone at $35,000. Mounting a website for medical research takes from three months to a year, depending on the amount of usability testing that needs to be done. At the same time, an under-budgeted or poorly designed website can result in sub-optimal usability for subjects, data that cannot be used to answer research questions, exposure of sensitive patient data, etc. Poorly designed research websites are created because:

- The medical researcher does not understand the nature of website design. (A researcher may not anticipate the need to translate their professional model of the subject matter into a model manageable by the consumer.)
- The individual or team hired to design the website may not have knowledge about the nature of the subjects or the treatment. (The designer may not be aware of physical or cognitive impairments that will require simplified navigational systems or reduced page complexity.)
- Medical researchers may not understand the capabilities or the limitations of data gathering related to the website and designers may not be aware of the data analysis needs of the intervention. For example, how will session information be maintained, how will IDs be assigned, and how will treatment groups be identified?
- The designers may not have sufficient knowledge of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) which sets rigorous standards for protecting patients' personal information (Neale & Schwartz, 2004); data collection, storage, and sharing mechanisms must be carefully designed to adhere to these standards.

Websites used in a variety of interventions reported in the literature had many common elements. This suggested that it might be possible to develop tools to aid the development of reliable and functional medical intervention websites at lower cost. With this in mind, we set out to develop a system that would:

1. Eliminate common pitfalls in design
2. Enhance data collection to better meet research objectives.
3. Reduce cost.
4. Ensure that the website is maximally usable by the target population.
5. Ensures that the website is as secure as possible.

The effectiveness of a system can be assessed based on:

1. Time required to build a website.
2. Monetary cost of building a website.
3. Ability to meet HIPAA requirements for security and privacy of medical data.
4. Usability of the website from both the end-users' and researchers' perspectives.

The rest of the chapter is structured as follows. First, we describe the needs of web-based medical interventions, as well as HIPAA and its implications for medical websites. This section also provides a survey of recent efforts to standardize the collection and presentation of medical information. Second, we present the framework for designing Internet-based interventions based on the critical characteristics identified from recent literature and introduce Webmaster-R (Webmaster for Research), a system built around the framework that is designed to augment a researcher's ability to mount a medical intervention website. Third, we present some formative evaluation data related to the efficacy of the Webmaster-R system using a heuristic evaluation and walkthrough with researchers from the University of Pittsburgh's School of Medicine. The last section discusses directions for future work.

BACKGROUND

Bush et al. (2004) proposed a framework for Internet and Web access that medical researchers should consider when designing an online study. The authors divided these issues into two groups, connectivity and human interface. *Connectivity* encompassed all issues related to the capabilities of the hardware and software used in conjunction with Internet access, as well as the availability of access based on geography, locale, and degree of convenience. *Human interface* involved all issues related to demographics, skill sets, and special needs of potential participants. Bush et al. then applied the highly detailed framework to the recruitment of subjects for a pilot study communicating breast cancer risk. Marshall and Haley (2000) presented a ten-step process for building a secure website for collaborative research from the ground up. The authors focused on the site's database, including 1) developing the database schema, 2) selecting a DBMS package, and 3) creating a separate table of variables for each investigator. Next, the authors 4) designed the screen layouts of each Web page, 5) selected a middleware product for connecting the site's pages to the DBMS, and 6) coded the pages, using JavaScript validation code on each page to verify user input. After completing the database and website, the authors 7) set up a machine to house both the database and web server software, 8) obtained a unique IP address and domain name, 9) applied security measures to the infrastructure, and 10) trained the staff who will enter data. The authors provided a detailed outline of suggested security practices, including security procedures for the website, servers, client machines, and facilities housing the servers.

Spring and Rotondi at the University of Pittsburgh collaborated on the development of several sites to support research on internet-based medical interventions. This chapter is in part a report of a series of interviews conducted in July, 2005 with Spring, Rotondi, and other researchers involved in the early stages of website development to gather a preliminary list of needs. The list of critical needs generated from these interviews included:

- Navigation and design elements appropriate for audience
- Methods to ensure information is presented at the right level of detail for the intended audience.
- Tools to optimize data collection and analysis.
- Naming conventions for various elements of an internet-based intervention, including the usernames of participants and the names of web pages.
- Ability to segment participants and isolate interactions based upon these segments.
- The site developed complies with HIPAA guidelines.

Based upon this input, a prototype system was developed, whose design and preliminary evaluations are explained in greater detail in Sections 3 and 4 respectively.

Figure 1. Comparison of the identified types of medical websites

	Medical Information	**Medical Intervention**	**Medical Intervention Research**
Objective	Provide general information on medical conditions; match potential patients with physicians	Provide treatment and mechanisms to help consumers cope with their condition; improve well-being	Collect data and test hypotheses for research while providing treatment to consumers
Access	Public	Private	Private
Bound by HIPAA?	No	Yes	Yes
Subject to IRB Review?	No	No	Yes
Data Analysis Needs	No	Minimal	Significant

A Website Taxonomy

Websites serve different purposes. We classify sites generally as 1) providing information, 2) supporting interaction, 3) allowing transactions, 4) supporting transformation. Generally, the evolution of a website moves from information only to one that provides the first three functions. The last stage of website evolution is one where it transforms the nature of the service or business. For example, in phase 3, the web provides an alternative mechanism for transactions. In phase 4, new services may be provided which are not possible in non-web formats. Our work here is primarily directed to sites that provide for information and interaction, with some minimal consideration of transactions. The use of websites to transform how medical interventions are delivered is likely to come about only after significant research data has been collected on the efficacy of parallel interventions.

We further classify medical websites in terms or whether they are intended to support: 1) general information dissemination, 2) patient interventions, or 3) research on interventions. Figure 1 summarizes a few of the differences. The nature of the types of sites is described below.

Information Websites

Medical information websites provide information on one or more medical conditions. No relationship with a physician is assumed. Because these websites do not provide any treatment, they are not bound by HIPAA guidelines. If discussion groups and chat rooms are publicly accessible, users are informed that their postings are not private. One of the largest and well-known medical information websites is WebMD's Consumer Portal (http://www.webmd.com), which provides public access to treatment information, discussion boards, and expert chats concerning a wide variety of medical conditions, all targeted to a general audience.

Medical Intervention Websites

Medical intervention websites are designed for physicians to provide information and/or treatment to specific consumers who have specific medical conditions. While the information provided on medical intervention websites may be similar to medical information sites, it is targeted at specific individuals and may contain references to an individual's specific condition. Thus, measures are taken to ensure that interactions among site users remain secure and private. Because client-physician relationships are established, medical intervention websites must adhere to HIPAA guidelines. Therefore, these sites tend to serve a much smaller audience whose members must be granted access to the sites resources explicitly. Interactions on the site (e.g., discussion groups) are monitored and accessible only to group members and authorized individuals. Data collected from the website must be secured, and all personally identifiable information eliminated from general exposure.

Medical Intervention Research Websites

Similar in structure to medical intervention websites, *medical intervention research websites* are specifically designed for collecting data on client interaction as clinical trials are conducted through the website. Medical intervention research websites share the same characteristics of medical intervention websites, but also include features – such as log and data analysis tools - to aid researchers in the analysis of the efficacy of the intervention. The studies cited in the next section all employed secured, private medical intervention research websites available only to specific audiences dealing with specific medical conditions.

Common Characteristics of Medical Intervention Websites

Many of the early Web-based medical interventions use the web to deliver treatment and provide guidance to patients, family members and the caregivers of those suffering from a medical condition. These interventions have become more feasible over the last 10 years as the penetration of Internet connectivity in U.S. households has increased. There is increasing support for the use of the web for a wide range of treatments (Ikemba et al, 2002.) Our work on intervention websites has focused on two common functions:

Information Dissemination

The websites include a repository of condition-specific medical information designed for patients and their families, as well as links to selected external resources. Presenting the information at an appropriate level with the right amount of detail was found to be especially critical. Rotondi et al. (2007) used a concrete, text-rich design with a shallow navigation structure that has been found to be more appropriate for those suffering from cognitive impairments, such as schizophrenia. Christensen et al., (2004) included a depression literacy website designed with an eighth-grade reading level. Lenert et al. (2003) attributed the high drop-out rate in their multi-stage smoking cessation program to relevant medical information being presented at an inappropriate level of detail (namely confusing navigation structure and text-heavy design).

Communication Mechanisms

Communication between individuals and professionals is an important aspect of many research intervention studies. The public intervention websites (e.g., WebMD and Schizophrenia.com) provide public discussion boards for individuals to share experiences related to the medical condition or topic. The research-oriented intervention websites included a mechanism for doctors to send messages to groups of patients, and most included a hyperlink or contact form for individual patients to send messages to doctors. Gray et al. (2000) included monitored discussion boards for interaction among patient families, as well as video conferencing capabilities for real-time communication between families and medical staff members. Communication was not anonymous in the discussion groups, but privacy of communication was paramount in doctor-family interactions. Because the site fostered a great deal of support among families and caregivers, the Internet-based intervention was well-received. Lenert et al.'s (2003) smoking cessation intervention had limited interaction capabilities, other than doctor-patient and site-patient communication via email, which may have also led to the program's high dropout rate.

HIPAA Concerns

The Health Information Portability and Accountability Act (HIPAA), which took full effect in 2003, specifies procedures for handling the personally identifiable information (PII) of patients in a secure, private, and confidential manner. HIPAA places restrictions on the PII that can be collected, stored, and transmitted via websites and email communications. The Act also requires organizations to disclose their procedures for handling PII in their privacy policies. While many organizations have updated their privacy policies to meet new guidelines, a recent study found the new policies have actually become harder to read and understand (Anton et al., 2007.) Given the ramifications of HIPAA's guidelines, as well as the long-existing needs to preserve patient confidentiality and the integrity of research experiments, a number of recent studies have focused on privacy and security implementations within the context of Internet-based interventions.

Schmidt (2003) examined HIPAA's implications on the privacy and security of electronic patient-physician communications. The original text of the HIPAA Security Rule required all email communications between patients and physicians to be encrypted. Because no universal solution for email encryption exists, this rule would have significantly discouraged the use of email by both patients and doctors, impeding communication between the two. The current text of the Rule now states that email encryption should be performed if it is feasible, but allows organizations to address alternative methods. Schmidt advocates examination of security measures (ex. PGP, Web-based Encryption) to avoid accidental disclosure but suggests having patients authorize the use of unencrypted email.

Baer, Sariou, and Koutsky (2002) provided a detailed explanation of their data collection and storage methods of sensitive health information from university students. Men and women were assigned access to different URL's, where all communication (including login) was encrypted using SSL 3.0. All form data provided by subjects were validated using a JavaScript-based validator before being posted. Validated data were encrypted using PGP v.2.6.2 and then written to the server in text files. Only authorized administrators decrypted the data in the text file, using a separate computer disconnected from all networks.

The need for website authentication and access controls was critical to the integrity of experiments, especially when identifying the subjects of experimental groups using the website during trials, and making sure only they were the ones using the website. McAlindon et al. (2003), while using their website to also recruit potential subjects for their osteoarthritis trials, used offline consent forms and medical records to authenticate subjects before accepting them into the study. Gray et al.'s (2000) Internet-based intervention for high-risk infants utilized RSA's SecureID handheld device, allowing users to log in to the website with a one-time hash generated from the device along with their memorized PIN.

FRAMEWORK FOR MEDICAL INTERVENTION RESEARCH WEBSITES

The frameworks suggested by Bush et al. (2004) and Marshall and Haley (2000) provide reasonable starting points for medical intervention researchers. However, these solutions require a significant working relationship between researchers and web designers. As Rotondi and Spring reported in interviews, the cost of this kind of web site development and maintenance exceeded $200,000 in their studies. Examination of the frameworks suggests that the difficult problems exist in three areas: infrastructure, site development, and data analysis. There are interrelations among these three areas, but they can to some extent be treated separately. The area of greatest cost variance, and the area most amenable to automation is the second – site development. The primary focus of our work has been on addressing this second area by developing a tool called Webmaster-R (Webmaster for Research). While the tool is conceptually similar to generalized tools such as Dreamweaver and FrontPage, the tool is also designed to insure that the site meets the requirements for security and privacy of medical data as outlined in the Health Insurance Portability and Accountability Act, meets the needs of the researchers for ease of construction and maintenance, and insures that data is appropriately collected and organized for research study data analysis. This work has been funded in part by the Dean of the School of Information Sciences at the University of Pittsburgh. Below we discuss some of the infrastructure considerations, which are generally outside the scope of the current work. Next, we introduce the website development framework considerations. Finally, we examine some of the data analysis considerations, which are not yet addressed, but which are within the scope of this research.

Infrastructure

Infrastructure includes the hardware running the intervention website, as well as the facilities and network supporting the site. With the risks of privacy, security, and confidentiality breaches of patient data, setting up a properly-designed and robust infrastructure requires careful consideration. Infrastructure concerns include:

- **Physical Facilities:** Equipment should be housed in locked rooms with access limited to a small number of people. Servers should be connected to uninterruptible power supplies with surge protection to prevent loss of data and minimize server outages.
- **Network:** The bandwidth and Quality-of-Service(QoS) needs of a medical intervention website will vary depending upon the web components, the number of users accessing the site, frequency of access, and the amount of data transmitted per visit. Employing a bandwidth-intensive component such as video conferencing, as Gray et al. (2000) did in their high-risk infant study, makes Quality-of-Service a potentially critical aspect of a medical intervention website. Service disruptions may cause undue anxiety to the patients (or caregivers), diminishing the effectiveness of the intervention and suggesting the need for redundant servers, disks, and network access points.
- **General Server Security:** The number of users with access to the data on the server should be kept to a minimum. Default passwords should be changed to ones that cannot be easily guessed or discovered using brute-force attacks. Anti-virus, firewall, and intrusion detection software should be installed and configured to run automatically. All unnecessary network services (ex. Telnet, FTP, IRC) should be disabled.

Network connections to the server for purposes of development should be restricted to authorized domains.

- **Logging, auditing:** Auditing and log analysis is critical in determining not only who is accessing the site, but also in providing statistics on how often subjects are visiting the site and what activities they engage in during their visits. Logging was important in Lenert et al.'s (2003) smoking cessation intervention for determining how far subjects progressed in the multi-stage program before quitting the trial, as those subjects most often did not respond to e-mail follow-ups. When assigning user logins, employ naming conventions that make identifying different categories of authorized users in the log files easier.
- **Backup procedure:** Data critical to the intervention – database tables, web pages, and site structures – should be backed up each night on tape media. RAID redundancy should be employed to prevent loss of data in case of a fatal hard drive failure. All backup procedures should be documented.
- **Web and DBMS servers:** All pages, including the root page of the website, should be accessed only through authenticated login, ideally using SSL encryption (thus requiring a certificate issued from a certificate authority such as VeriSign or Thawte.) The URL of the research website should be advertised only to authorized users. Permissions should be set on directories and/or specific pages to prevent access from unauthorized users within the research study (e.g., to prevent participants from gaining access to administrative pages.) Security procedures should be documented, and staff members should be trained on website security techniques.

While this chapter focuses on the development of the website itself, the following checklist provides an indication of the complex issues that a medical researcher should insure are being addressed by those responsible for installation and management of the infrastructure.

Related to the Operating System (assuming Microsoft Windows):

- Apply all security patches before connection to the Internet, from a CD or other removable media.
- Do NOT set up the server as part of a Windows domain.
- Set the administrator password to at least 8 characters in length.
- Use a password that is difficult to guess, not based on a dictionary word, using a mixture of numerical, upper and lower case characters.
- Create separate logins for each user, and give them access only to the resources they need.
- Install anti-virus software (such as Symantec Anti-Virus).
- Set the Windows Firewall to "on".
- Obtain (for free) Microsoft's Baseline Security Analyzer for routine security checks.
- Check event and system logs on a regular basis (daily or weekly).
- Set Windows Automatic Updates to notify you when patches are available, download them, but NOT to install them. Verify and only then install them.

Related to the Web Server (assuming Microsoft IIS):

- Turn on only those services (e.g.. FTP, mail) that you need.

° If using IIS 5.0, turn off all unnecessary services.

° If using IIS 6.0, check that all other services are turned off (by default, they should be.)

- Ensure logging is enabled

- Ensure that Directory Traversal is turned OFF.
- Disable Internet Printing
- Give write privileges only to administrators or other users authorized to edit the content of the website.

Securing Your Database Server (assuming Microsoft SQL Server):

- Run the database server under an account without administrator privileges
- Assign a password to the "sa" account.
- Do NOT allow direct connections to the database server.
- Devise a backup plan for all databases
- Do NOT grant full access to your databases to all users. Limit users' access to only those resources that they need.
- Ensure that all scripts and code programmed by developers are secure.

° Verify input data types (ISNUMERIC)

° For string data, replace single quotes with two single quotes

° Create reusable input validation modules

° Use stored procedures to abstract data access.

This is only a sample list. More definitive and up-to-date lists can be obtained by your organizational security offices or from organizations such as CERT (http://www.cert.org/).

Website Design Principles

As with websites generally, the components of the website can be classified into four categories: 1) *information*, referring to the medical content and related resources of the site, as well as any means of searching, indexing, or managing the site's content and resources; 2) *interaction*, which describes person-to-person communication facilitated by the website, 3) *transaction*, which refers to any medical treatment performed via electronic communication (ex. medical diagnosis, testing, prescriptions), and 4) *transformation*, which refers to unique services available only through web services. We examine the nature of the first three below.

Information components provide access to the medical content and related web resources located internally or externally, as well as any means of searching, indexing, or managing the site's content and resources. The presentation of information should center on the characteristics and needs of the target audience. Information should be presented at the appropriate level and appropriate detail for the understanding of the intended audience, and mechanisms should be put in place to ensure the appropriateness of the content (ex. checking reading level). Navigation structures should coincide with a user's mental model of the presented information. Mechanisms should be provided for collaborating researchers to insert and review resources. Resources should be indexed with search capabilities, so users can easily locate relevant material. Other Web resources, such as documents on external websites, should be included only if they are relevant to the treatment.

Interaction entails any aspect of person-to-person or person-to-group communication facilitated by the website. In a medical intervention environment, interactions can have many distinct forms: Doctor-site, Patient-doctor, Caregiver-doctor, Caregiver-patient, and Patient-patient. Doctor-site interaction refers to "broadcast" communications that are posted by a medical researcher onto the website, such as responses to questions or event announcements, targeted to the general audience. The remaining interactions, however, are all individual-individual, and require special consideration in terms of HIPAA compliance and confidentiality:

- **Isolation of discussions:** Create separate chats or discussion boards, granting permissions only to members of specific groups to facilitate treatment and preserve patient confidentiality.

- **Privacy of communications:** Interactions should only be disclosed either according to the site's privacy policy, or by explicit consent of the patient or caregiver. Mechanisms should be put in place (such as options on an HTML form) that allow participants to control disclosure of their communications.
- **Anonymity of communication:** According to HIPAA guidelines, virtually all personally identifiable information (PII) must be stripped from communication. When possible, data collection tools should be set up to exclude PII and only use usernames, making identifying individuals difficult should information be disclosed unintentionally.
- **Real-time communication:** IRC and video-conferencing capabilities should be included if synchronous communication is required as part of the intervention.

Other elements of communication that require modifications to infrastructure procedures should also be considered:

- **Archiving of communication:** Communication relevant to the research, particularly any interaction involving patients, should be stored following HIPAA guidelines, eliminating PII. Archives should be organized and stored on a secure server, preferably in a secure directory on the web server where it can be recorded for backup.
- **Flow-through mailing of communication:** Mechanisms should be created that allow for automatic classification and redirection of electronic communication received by the site's mail server to the appropriate researchers or staff members.

Transactions – namely *diagnosis*, *testing*, and *prescriptions* – are the most difficult to implement via internet-based interventions, and are rarely attempted in the cited literature. Mechanisms for supporting these transactions must allow the physician the ability to perform the requested transaction without violating *medical ethics*. Physicians who diagnosis conditions electronically must be able to do so with the relative ease and certainty as through traditional means, especially the ability to authenticate the patient and verify the presence of a condition. Web-based testing should only be performed within acceptable risk levels, and if the outcome measures of the tests can be transmitted electronically. As is the case with diagnosis, authenticating patients and verifying the need for requested medications are paramount when prescribing medications via electronic means. McAlindon et al.'s (2003) osteoarthritis clinical trials tested the effects of glucosamine, a safe nutritional product, on osteoarthritis of the knee; with a testing procedure that had a very low risk of complications, McAlindon et al. felt comfortable conducting the experiment remotely. They also stress that osteoarthritis of the knee is relatively easy to diagnose remotely, so they relied heavily on recruitment and diagnosis through electronic means (although, as mentioned earlier, participation in the study depended on the subjects' medical records and their signing of an offline consent form.) Finally, they also argue that testing through electronic means is feasible if the outcome measures of the trials can be transmitted electronically, such as their verified osteoarthritis test that relies on a computer-based self-evaluated questionnaire.

Automated Website Design

Initial meetings were held with M.B. Spring and A.J. Rotondi to review their experiences in developing the websites for NIH studies. They provided access to the software developed for these sites, as well as information about the design process, mistakes and oversights that were made, and other problems encountered. Combining a general design framework developed by Spring intended for conference and proceedings management and the custom-designed software for websites that served as the basis for previous medical intervention research studies, a system was developed that poses a series of questions to principal investigator in terms that they should understand. Based on those answers a fully functional website is developed that can be manually enhanced if the PI so desires. Users specify the design elements of their site and the components they wish to include. These preferences are then dynamically merged with a series of predefined templates to produce the site's pages. The system design and how it addresses the theoretical framework are discussed below.

Information - A series of components were developed to address the presentation and navigational needs of a website, as well as the appropriateness and management of content. The system's Website Options allow the researcher to create the presentation and navigational elements of a website, including:

- Cascading style sheets
- Global website navigation menu
- Website logo
- Color and font schemes for text, links, and backgrounds

Tools are also available to help researchers determine the appropriateness of content in terms of level and detail. Researchers can compute a simple Flesch-Kincaid reading level score for documents, or they can evaluate a document's readability based on the cognitive abilities of its intended audience.

All content added to Webmaster-R is indexed and linked appropriately according to the researcher's preferences. Search functions are built into information-intensive modules, such as resource libraries, using sophisticated document indexing techniques. Threaded discussion boards use simpler database search techniques.

Interaction - Components were developed to provide functionality commonly found on medical intervention research websites to facilitate communication among participants and staff. Interactive modules available through Webmaster-R include:

- Threaded discussion boards
- Blogs, either as personal journals or announcements from staff
- IRC chat
- Video streaming
- Contact forms

Researchers can create multiple instances of the above modules, and specify which groups of participants have access to specific instances. Additionally, researchers have the ability to create their own online questionnaires and specify how 1) they will be presented to users, and 2) how they will be measured for quantitative analysis. Secure mail facilities are included with appropriate modules (i.e. discussion boards, contact forms) to direct communication from participants or the website to appropriate staff members.

Transaction - Components of Webmaster-R can be used for transactions, but in general, the framework stops short of explicitly advocating transaction-type components because of the complex implications, as well as ethical considerations. Even in cases where the medical researcher is also the primary care physician for the intervention's participants, the risks of misidentifying a patient, misdiagnosis, or providing an erroneous prescription often outweighs the benefits of performing a transaction via electronic means. The obstacles to building a general-purpose set of tools for diagnosing any ailment and prescribing any drug electronically via this type of system are great. With cautions, PI's can use components such as self-report surveys and mail – included under the interaction class to approximate transaction interactions.

Content Management & Data Analysis Tools

Content management facilities through a protected administrative section are automatically included with every website created by Webmaster-R. Items in resource and question-and-answer libraries can be added, removed, or edited either directly online or via web upload. Researchers may organize the content in both libraries by categories that may be added or re-labeled online. Access to real-time chat rooms may be granted to groups or specific users through the created website, and full transcripts of all chat room discussions are available at any time. With the consent of participants, researchers can also add and edit profiles, bios, and photos of participants and staff.

Data and log analysis tools are provided to help researchers audit their site's usage, track usage for individual users, and identify potential problems with the sites usability. Website logs are parsed to remove entries that represent support files (i.e., images, external stylesheets and HTML framesets) and filtered to remove visits by staff members. Tools are available to help researchers identify users' sessions, as well as search queries issued by participants. Researchers can then create customized summaries or detailed reports via a web interface. For example, a researcher can produce a report summarizing the number of postings and read messages within a discussion group over a given timeframe, or how many times a particular user accessed specific threads over the duration of the study. Full automation of the data collection has not yet been implemented, but all the tools and procedures have been documented that allow researchers to collect and analyze both web logs and database tables.

Implementation

Webmaster-R was implemented as a Java-based application, producing standard XHTML or JSP documents as output from pre-defined templates. For output that requires database-driven components, Webmaster-R includes a series of preconfigured Java servlets able to communicate with any Microsoft SQL Server database. Included are servlets that automatically write out the necessary database tables for the website, all of whose schema conform to HIPAA guidelines. At the time of this writing, a typical Webmaster-R session of an hour will result in the creation of more than 250 files, most of which are programs that process data, distributed in 25 folders representing almost a Megabyte of data.

Webmaster-R is available for free to anyone in the medical research community to use.

PROTOTYPE EVALUATION

After development of a prototype version of Webmaster-R, meetings were held with two researchers from the University of Pittsburgh who were in the early stages of programs that would entail research on web based medical interventions. The researchers provided more insights into the budget process, usability issues encountered, technical limitations overcome, and techniques used to comply with HIPAA. This section reports on a series of prototype walkthroughs and an initial usability study with a research team developing a new medical intervention research website. After Figure 2, Figure 3 presents a summary of the feedback evaluating the effectiveness of the framework. Five metrics were used:

- Saves money?
- Saves time?
- HIPAA compliant?
- Acceptable usability - application?
- Acceptable usability - resulting websites?

Figure 2. The functional components of Webmaster-R

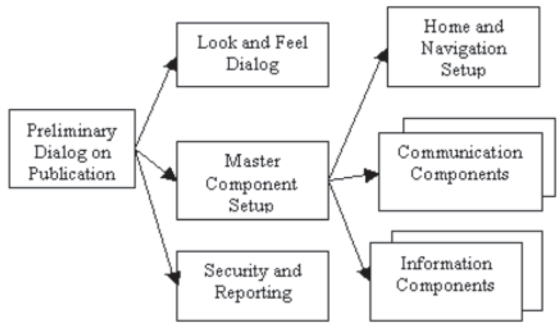
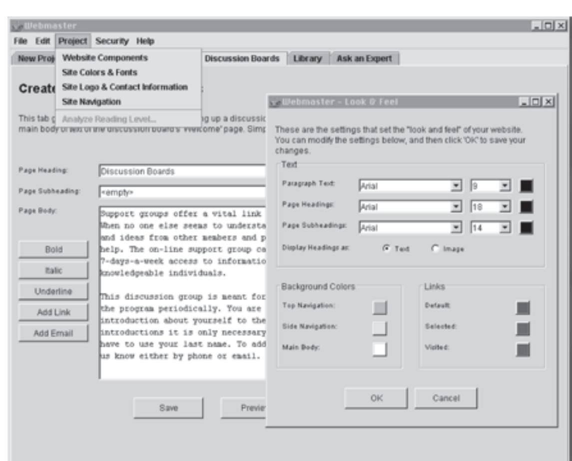


Figure 2 shows the functional components of Webmaster-R. The PI is able to incrementally develop the components, adding and changing the components at any time.

Users can preview their work in HTML as they work, and also have the option of publishing the site to their local computer (for later uploading) or directly to a web server, provided they have a network connection to the server. As the components are added, a system of files is built up recording the decisions. At any time, the basic configuration can be modified or changed. As the PI answers questions that have been designed to make sense to them, complex pre-configured common facilities are added. For example, answering a question such as "Do you want periodic reports of website activity?" will activate code fragments and scripts that will do database queries automatically at specified intervals and mail the results to given email addresses. Answering yes to a simple question will add code and structure that would take at least several hours and, in many cases, several days of programming if done from scratch.

A typical view of Webmaster-R is provided in Figure 3 showing the kinds of information that the PI is expected to provide.

Figure 3. Screenshot of Webmaster-R



Prototype Walkthroughs

The first of three prototype walkthroughs was conducted in July, 2005 with three researchers from the University of Pittsburgh. The walkthrough consisted of a demonstration of an early Webmaster-R prototype, with feedback gathered during and after the demonstration. The Webmaster-R prototype demonstrated at this meeting had fully functional Website Option tools, but only one functional Website component, the Home Page.

Participants were asked the following questions regarding the application:

1. In what areas of research would you utilize a medical intervention website?
2. What special needs of your audience need to be addressed in the design of the site?
3. Are there aspects of Webmaster-R that you feel would be helpful to you?
4. What do you think is missing from Webmaster-R?
5. Does Webmaster-R meet your research needs?
6. Would you consider using Webmaster-R to develop a website?

The initial impressions of Webmaster-R were generally favorable. Two researchers indicated that they would consider using the application to build a medical intervention research website, and that it would reduce the time needed to implement the site. One participant indicated he could not because the websites generated by the program would not meet the usability needs of consumers with cognitive impairments. Participants found the choices of website components suitable for their information and interaction needs, although they could not fully judge the application's ability to reduce development costs or comply with HIPAA without seeing the finished components. They also cited the program's ease of use and ability to generate a website from a small set of preferences as strengths. Participants also indicated, however, that some of the program's user interface elements were confusing, particularly the names of some of the input fields and buttons. They suggested improving the interface by providing more visual cues as to what website content a particular field will generate – for example, pre-filling each component's text fields with descriptive text rather than presenting empty fields.

The suggestions were incorporated into Webmaster-R's user interface, and development continued on the component parts of Webmaster-R. In November, 2005 a second walkthrough was conducted with one of the researchers interviewed in the initial meeting. Website components for discussion boards, resource libraries, and group member profiles had been completed and were demonstrated. The participant indicated that the demonstrated components, as well as the rest of the program would be useful, but that a streaming video component would also be important to her research. The demonstrated components would help her reduce both development time and monetary costs, while still complying with HIPAA regulations. She also indicated that a minimal set of documentation (10-20 pages) should accompany the application as a user's guide. This walkthrough was followed by the initial usability test of the Webmaster-R prototype. Following the meeting, online help documents were created and added to the application.

In the third meeting held in December, 2005, two researchers from the WPIC Behavioral Medicine Program were presented with a complete prototype of Webmaster-R, including design changes incorporated from previous feedback. The researchers were given the option of using the application to prototype the website themselves rather than hire outside consultants to build the site. The researchers indicated the application would be very useful for building a website and would reduce their development costs in terms of both time and money. They found the set website components sufficient for their research. Both researchers still found some of the program's terminology confusing, and felt some of the resulting pages (particularly in the Library component) would be unsuitable for their target audience – i.e. adolescents. Following this meeting, the application's interface and templates were altered to address the researchers' feedback. The changes were implemented prior to the second usability test. A set of documents was also created to help the researchers with issues outside the scope of the program, including 1) instructions for installing the program, 2) generating and saving additional web pages using a word processor, and 3) using document templates to enforce a consistent style across all documents.

Usability Tests

After further development of the prototype based on the feedback from the aforementioned interviews, three usability test sessions were conducted – two with a researcher from the Starzl Transplant Institute (STI) and one with a research assistant from the WPIC Behavioral Medicine Program (WPIC). Both sessions with the STI researcher were conducted as cognitive walkthroughs (Wharton et al., 1994) requiring the researcher to show the steps she would take to 1) begin a website project, 2) add website components to the project, and 3) configure a discussion board for all group members using a prototype of Webmaster-R. The session with the WPIC research assistant was conducted similar to Battleson et al. (2001) using the think-aloud protocol on a more advanced version of the Webmaster-R prototype. The research assistant, who was not familiar with HTML, was asked to start a new website project and complete the following sub-tasks: enter the project's technical settings, design the site's layout, define the website's administrator and contact information, create a navigation menu, add and configure five website components, supply text for the Home Page component, and add at least one hyperlink to the Home Page's text. The goal of each usability test was to assess Webmaster-R's user interface – not necessarily to test if the user could construct a production quality website in one sitting.

Based on the test results and the participants' feedback, the tool shows promise but is not yet a complete solution to researchers' needs. We identified three critical areas where the tool needs to be enhanced:

- **Provide more component choices:** In her second usability session, the STI researcher requested a number of website components that were unavailable in Webmaster-R. Her requested site functionality included components based on "Web 2.0" technologies – technologies that provide a more interactive and seamless experience for visitors. Two such components – a personal blog for patients and a real-time chat requiring only a standard web browser, have since been incorporated into Webmaster-R. Other requested components, such as an online questionnaires and video streaming, are more challenging to include in an automated tool because of the complexity of the final products. For questionnaires, it was relatively easy to construct a tool to build questionnaires based on a series of questions using a Likert scale. The prospect of an instrument with complex logic and varying response types was beyond the scope of this current effort. For video presentations, the tool should not only be able to incorporate items from various multimedia formats, but it should also build presentations that will work for participants with older, slower computers and low-bandwidth Internet connections. Building a streaming media server from scratch was deemed beyond the scope of this effort.
- **Focus on end-user design options:** Another unsolved challenge in Webmaster-R is the

Figure 4. Summary of the evaluations of the effectiveness of the framework

	Walkthrough 1	Walkthrough 2	Walkthrough 3
Saves Time?	Yes	Yes	Yes
Saves Money?	Inconclusive	Yes	Yes
HIPAA Compliant?	Inconclusive	Yes	Yes
Acceptable Usability (Application)?	No	No	Yes
Acceptable Usability (Websites)?	No	Yes	No

lack of mechanisms that ensure a website will have an appropriate design for its intended audience. For example, the WPIC research assistant was able to build a functioning website within 90 minutes, but the text-heavy appearance of the constructed site would have been inappropriate for younger audiences. Other researchers have expressed an interest in building sites that meet the requirements of Section 508 of the Rehabilitation Act to be used by people with cognitive and physical impairments. One potential automated solution to this problem is to ask researchers a series of questions describing the characteristics of their audience, and then using one of a set of pre-defined design templates that is most appropriate for the described audience.

- **Make the tool more usable for researchers:** Our goal is to make Webmaster-R easy for a researcher to use without having to wade through volumes of documentation. Although improvements to the interface made the tool more intuitive for the STI researcher in the second session compared to the first, she found some of the terminology in the directions and labels too technical. In the session with the WPIC research assistant, online help pages were provided to help overcome blocks; however, these pages were never accessed during the session. Visual cues need to be added to the interface that make clear what pieces of information will be viewed "publicly" on the website by consumers and caregivers versus information that it intended for internal purposes only.

CONCLUSION

This chapter introduces a framework for classifying the critical needs of a medical intervention research website. Within this framework four main categories have been identified for describing these critical needs: infrastructure, information, interaction, and transaction. Based upon this underlying framework, the Webmaster-R system was built to test the framework and assess the feasibility of automating the web development process for medical intervention research websites. Preliminary evaluations conducted on the system indicate that researchers find the tool helpful and able to meet most of their research objectives. The formative evaluation indicates that the design tools are not yet flexible enough to meet some specific objectives. Notably, the tools do not yet insure that the resulting material is simple enough for younger audiences, or consumers with cognitive impairments. On the other hand, in 5 different situations, we have been able to reduce the time required to mount a complex automated research website, with most of the work done by the medical researcher his or her self, from three or four months to one to two weeks. Further, the actual time required to obtain a polished final website has been reduced from three to four months of technical work to less than 40 hours.

While these results are promising, much work still needs to be done. The current studies suggest that such a tool can aid in the rapid and inexpensive development of websites that can be used as the basis of clinical trials. At this point in time, there is no guarantee that the resulting sites will be mounted so as to guarantee HIPPA compliance. It is not clear whether it is possible to insure this. In addition, while a number of scripts have been written to aid in the analysis of the data collected via the website, we have treated this process individually to date so as to answer as many of the questions researchers have as possible. We have not found many questions that could not be answered by analysis of the data, but it is not clear that this process could be fully automated at this point. Finally, each researcher we have worked with has asked for additional functionality that we have added to the basic model. This process seems to accelerate as the sophistication of researchers and subjects increases and as the functionality it is possible to realize through browser technology increases. While we have found it easy to extend the framework to accommodate new functionality, each option makes the system more complex from a design perspective. Ultimately, validation of the system ultimately requires the same kind of empirical validation as the studies it endeavors to support - i.e., research studies of interventions based on this tool more successful than research studies using ad hoc site development.

In the introduction we referred to a paper by May et al. (2005), which examines the provision of telehealth services in the United Kingdom. Their work, and this current effort approach the issue of developing supporting evidence of the efficacy of web based interventions from different perspectives. At the same time, the goal is the same – the provision of some degree of medical treatments via new mechanisms. From this perspective, the ultimate goal of this work is the development of tools and systems that make medical services available and accessible to those who need them.

REFERENCES

Anton, A., Earp, J., Vail, M., Jain, N., Gheen, C., & Frink, J. (2007). An analysis of Web site privacy policy evolution in the presence of HIPAA. *IEEE Security and Privacy*, *5*(1), 45–52. doi:10.1109/MSP.2007.7

Baer, A., Sariou, S., & Koutsky, L. (2002). Obtaining sensitive data through the Web: An example of design and methods. *Epidemiology (Cambridge, Mass.)*, *13*, 640–645. doi:10.1097/00001648-200211000-00007

Battleson, B., Booth, A., & Weintrop, J. (2001). Usability testing of an academic library Web site: A case study. *Journal of Academic Librarianship*, *27*(3), 188–198. doi:10.1016/S0099-1333(01)00180-X

Bush, N., Bowen, D., Wooldridge, J., Ludwig, A., Meischke, H., & Robbins, R. (2004). What do we mean by Internet access? A framework for health researchers. *Preventing Chronic Disease*, *1*, 1–17.

Christensen, H., Griffiths, K., & Jorm, A. (2004). Delivering interventions for depression by using the Internet: Randomized controlled trial. *British Medical Journal*, *328*, 265–268. doi:10.1136/bmj.37945.566632.EE

Dew, M., Goycoolea, J., Harris, R., Lee, A., Zomak, R., & Dunbar-Jacob, J. (2004). An Internet-based intervention to improve psychosocial outcomes in heart transplant recipients and family caregivers: Development and evaluation. *The Journal of Heart and Lung Transplantation*, *23*(6), 745–758. doi:10.1016/j.healun.2003.07.002

Gray, J., Safran, C., Davis, R., Pompilio-Weitzner, G., Stewart, J., Zaccagnini, L., & Pursley, D. (2000). Baby CareLink: Using the Internet and telemedicine to improve care for high-risk infants. *Pediatrics*, *6*, 1318–1324. doi:10.1542/peds.106.6.1318

Ikemba, C., Kozinetz, C., Feltes, T., Fraser, C., McKenzie, E., Shah, N., & Mott, A. (2002). Internet use in families with children requiring cardiac surgery for congenital heart disease. *Pediatrics*, *109*, 419–422. doi:10.1542/peds.109.3.419

Lenert, L., Munoz, R., Stoddard, J., Delucchi, K., Bansod, A., Skoczen, S., & Perez-Stable, E. (2003). Design and pilot evaluation of an Internet smoking cessation program. *Journal of the American Medical Informatics Association*, *10*, 16–20. doi:10.1197/jamia.M1128

Marshall, W., & Haley, R. (2000). Use of a secure Internet Web site for collaborative medical research. *Journal of the American Medical Association*, *284*, 1843–1849. doi:10.1001/jama.284.14.1843

May, C., Finch, T., Mair, F., & Mort, M. (2005). Toward a wireless patient: Chronic illness, scarce care, and technological innovation in the United Kingdom. *Social Science & Medicine*, *61*, 1485–2494. doi:10.1016/j.socscimed.2005.03.008

McAlindon, T., Formica, M., Kabbara, K., LaValley, M., & Lehmer, M. (2003). Conducting clinical trials over the Internet: Feasibility study. *British Medical Journal*, *327*, 484–487. doi:10.1136/bmj.327.7413.484

Neale, A., & Schwartz, K. (2004). A primer of the HIPAA privacy rule for practice-based researchers. *The Journal of the American Board of Family Practice*, *17*(6), 461–465. doi:10.3122/jabfm.17.6.461

Rotondi, A. J., Sinkule, J., Haas, G. L., Spring, M., Litschge, C., & Newhill, C. (2007). Designing Web sites for persons with cognitive deficits: Design and usability of a psychoeducational intervention for persons with severe mental illness. *Psychological Services*, *4*(3), 202–224. doi:10.1037/1541-1559.4.3.202

Rotondi, A. J., Sinkule, J., & Spring, M. (2005). An interactive Web-based intervention for persons with TBI and their families: Use and evaluation by female significant others. *The Journal of Head Trauma Rehabilitation*, *20*(2), 173–185. doi:10.1097/00001199-200503000-00005

Website: WebMD Consumer Portal (http://www.webmd.com).

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods*, (pp. 105-140). New York: John Wiley.

KEY TERMS AND DEFINITIONS

Content Management: The processes and tools that are used to introduce and maintain the content used by an organization or a website.

Data Analysis: Related to a website, data analysis involves the interpretation of logged data maintained generally for a website and specifically related to the needs of researchers.

Intervention Research: In the case of medicine, research that is designed to assess the efficacy of a particular intervention.

Navigation System: A series of website features that allow a user to move from one part of a website to another without error and to keep track of where they are.

Security: Concerns about authentication, access control and data privacy. With respect to a website, it relates to who is allowed to do what on a website and to the assurance that data is kept private, accessible, and its integrity is maintained.

Usability: The measure of the ability of users to navigate a website and use it to accomplish some goal with a minimum of mistakes.

Website Design: The process of gathering requirements, analyzing needs and designing a website that meets those needs.

This work was previously published in Handbook of Research on Information Technology Management and Clinical Data Adminstration in Healthcare, edited by A. Dwivedi and J. Grady, pp. 852-869, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 2.4
Designing Web Information Systems for a Framework-Based Construction

Vítor Estêvão Silva Souza
Universidade Federal do Espírito Santo, Brazil

Ricardo de Almeida Falbo
Universidade Federal do Espírito Santo, Brazil

Giancarlo Guizzardi
Universidade Federal do Espírito Santo, Brazil

ABSTRACT

In the Web Engineering area, many methods and frameworks to support Web Information Systems (WISs) development have already been proposed. Particularly, the use of frameworks and container-based architectures is state-of-the-practice. In this chapter, we present a method for designing framework-based WISs called FrameWeb, which defines a standard architecture for framework-based WISs and a modeling language that extends UML to build diagrams that specifically depict framework-related components. Considering that the Semantic Web has been gaining momentum in the last few years, we also propose an extension to FrameWeb, called S-FrameWeb, that aims to support the development of Semantic WISs.

DOI: 10.4018/978-1-60566-278-7.ch011

INTRODUCTION

The World Wide Web (also referred to as WWW or simply Web) was created as a means to publish documents and make them available to people in many different geographical locations. However, the advent of the Common Gateway Interface (CGI), in 1993, allowed for authors to publish software instead of documents and for visitors to execute them, producing dynamic results.

The evolution of Web development technology and the emergence of high-level languages (such as PHP, ASP, JSP, etc.) and platforms (such as Microsoft .NET and Java Enterprise Edition) allowed for more complex applications to be built on the Web. Soon enough, a handful of large B2C (business-to-consumer, such as online stores) and B2B (business-to-business, such as supply chain management systems) applications were being deployed on the Internet.

Thus, the concept of Web Applications (WebApps) was born. WebApps consist of a set of Web pages or components that interact with the visitor, providing, storing and processing information. WebApps can be informational, interactive, transactional, workflow-based, collaborative work environments, online communities, marketplaces or web portals (Ginige & Murugesan, 2001).

In this chapter, however, we focus on a specific class of Web Applications, called Web-based Information Systems (WISs). WISs are just like traditional information systems, although deployed over the Internet or on an Intranet. These systems are usually data-centric and more focused on functionality rather than content and presentation. Examples are online stores, cooperative environments, and enterprise management systems, among many others.

Although many Software Engineering principles have long been established before the creation of the Web, first-generation WebApps were constructed in an ad-hoc manner, with little or no concern for them. However, with the increase of complexity of the WebApps, which is especially true for WISs, the adoption of methodologies and software processes to support the development team becomes crucial.

Thus, a new discipline and research field was born. Web Engineering (or WebE) can be defined as "the establishment and use of engineering principles and disciplined approaches to the development, deployment and maintenance of Web-based Applications" (Murugesan et al., 1999, p. 2). Pressman (2005) complements this definition stating that WebE borrows many conventional Software Engineering fundamental concepts and principles and, in addition, incorporates specialized process models, software engineering methods adapted to the characteristics of this kind of application and a set of enabling technologies.

In this field, a lot of methods and modeling languages have been proposed. Some well known works are WebML (Ceri et al., 2000), WAE (Conallen, 2002), OOWS (Fons et al., 2003), UWE (Koch et al., 2000), and OOHDM (Schwabe & Rossi, 1998), among others.

Parallel to the academic research, the industry and the developer community have also proposed new technologies to provide a solid Web infrastructure for applications to be built upon, such as frameworks and container-based architectures. Using them we can improve productivity at the coding phase by reusing software that has already been coded, tested and documented by third parties. As their use becomes state-of-the-practice, methods that focus on them during software design could provide a smoother transition from models to source code.

This has motivated us to develop a WebE design method that focuses on frameworks. The Framework-based Design Method for Web Engineering (FrameWeb) (Souza & Falbo, 2007) proposes a basic architecture for developing WebApps and a UML profile for a set of design models that brings concepts used by some categories of frameworks, which are applied in container-based architectures as well.

Meanwhile, many researches have been directed to the construction of what is being considered the future of the WWW: the Semantic Web. Coined by Berners-Lee et al. (2001), the term represents an evolution of the current WWW, referred by some as the "Syntactic Web". In the latter, information is presented in a way that is accessible only to human beings, whereas in the former data is presented both in human-readable and machine-processable formats, in order to promote the development of software agents that would help users carry out their tasks on the Web.

However, for Berners-Lee's vision to become a reality, Web authors and developers must add semantic annotations to their Web Applications. This is neither an easy nor a small task and support from tools and methods is needed. Thus, an extension of FrameWeb was proposed. The Semantic FrameWeb (S-FrameWeb) (Souza et al.,

2007) incorporates into the method activities and guidelines that drive the developer in the definition of the semantics of the WISs, resulting in a "Semantic Web-enabled" application.

The objective of this chapter is to discuss the current research state regarding Web Engineering (methods and modeling languages), frameworks and the Semantic Web, and to present FrameWeb, and its extension S-FrameWeb, as a new method based on best practices for the development of Web-based Information Systems. We close the chapter by presenting future trends and opportunities for this research.

BACKGROUND

Web Engineering was born from the need to apply Software Engineering principles to the construction of WebApps, adapting them to the application's size, complexity and non-functional requirements. A lot of research on methods and modeling languages has already been conducted, providing an extensive background for new research.

Meanwhile, companies and independent developers create frameworks and propose container-based architectures to promote reuse and improve productivity while maintaining good design principles. Furthermore, research on the Semantic Web has been pointing out some directions on what the Web may become in the future.

This section discusses the current state-of-the-art and state-of-the-practice on Web Engineering, modeling languages for WebE, frameworks for development of WebApps and the Semantic Web.

Web Engineering

Web Engineering (or WebE) uses scientific, engineering, and management principles and systematic approaches to successfully develop, deploy, and maintain high-quality WebApps (Murugesan et al., 1999).

As with conventional software engineering, a WebE process starts with the identification of the business needs, followed by project planning. Next, requirements are detailed and modeled, taking into account the analysis and design perspective. Then the application is coded using tools specialized for the Web. Finally, the system is tested and delivered to end-users (Pressman, 2005).

Considering that, in general, the platform in which the system will run is not taken into account before the design phase of the software process, developing a WebApp would be just like developing any other application up to that phase. However, many differences between Web Engineering and Conventional Software Engineering have been identified by researchers and practitioners (Ahmad et al., 2005), such as sensitivity to content, short time frames for delivery, continuous evolution, focus on aesthetics, etc (Pressman, 2005).

This has motivated researchers to propose different methods, modeling languages and frameworks for Web Engineering. The amount of propositions is quite vast, demonstrating that academics and practitioners have not yet elected a standard concerning Web development. In this subsection we briefly present some methods, while the following subsections focus on modeling languages and frameworks.

Web Application Extension (WAE) (Conallen, 2002) defines an iterative and incremental software process, centered on use cases and based on the Rational Unified Process (Krutchen, 2000) and the ICONIX Unified Process (Rosenberg & Scott, 1999). It proposes activities such as business analysis, project planning, configuration management and an iterative process that includes the usual software development cycle from requirement gathering to deployment.

OOWS (Object Oriented Web Solution) (Fons et al., 2003) is an extension of the OO-Method (Pastor et al., 2001) for WebApp specification and development. It divides the software process in two main steps: conceptual modeling and solution development. In the conceptual modeling step, the system specification is obtained by using conceptual models. For that, OOWS introduces new models for representing navigational and presentational characteristics of web applications. In the solution development step, the target platform is determined, and a specific architectural style is chosen. Then, a set of correspondences between abstraction primitives and the elements that implement each tier of the architectural style are applied in order to automatically obtain the final system (Pastor et al., 2003).

The UML-based Web Engineering (UWE) (Koch et al., 2000) is a development process for Web applications with focus on systematic design, personalization and semi-automatic generation. It is an object-oriented, iterative and incremental approach based on the Unified Modeling Language (UML) and the Unified Software Development Process (Jacobson et al., 1999). The notation used for design is a "lightweight" UML profile. The process is composed by requirement analysis, conceptual navigation and presentation design, supplemented with task and deployment modeling and visualization of Web scenarios (Koch & Kraus, 2002).

Lee & Shirani (2004) propose a component-based methodology for WebApp development, which is divided in two major parts: component requirements analysis and component specifications. Analysis begins identifying the required component functions and is followed by a comparison with the functions available in existing components. The component specification phase has three activities: rendering specification, integration specification and interface specification.

The Ariadne Development Method (Díaz et al., 2004) proposes a systematic, flexible, integrative and platform-independent process for specification and evaluation of WebApps and hypermedia systems. This process is composed of three phases: conceptual design, detailed design and evaluation. Each phase is further subdivided into activities, which in turn defines sets of work products to be built.

Díaz et al. (2004, p. 650) also define the hypermedia paradigm as one that "relies on the idea of organizing information as a net of interrelated nodes that can be freely browsed by users selecting links and making use of other advanced navigation tools, such as indexes or maps". We consider hypermedia methods quite different than methods for the development of WISs, as they focus on content and navigational structures instead of functionality and seem to be better suitable for information-driven WebApps.

Although hypermedia development methods are not on our focus, it is worthwhile to cite OOHDM (Object Oriented Hypermedia Design Method) (Schwabe & Rossi, 1998), a well-known method that is representative of hypermedia methods. It was born from the need to represent hypermedia structures such as links, text-based interfaces and navigation, and more recently has also been applied to Web development. For instance, an extension of this method, called OOHDM-Java2 (Jacyntho et al., 2002), was proposed, which consists of a component-based architecture and an implementation framework for the construction of complex WebApps based on modular architectures (e.g. Java EE). The OOHDM process is divided into five steps: requirements gathering, conceptual design, navigational design, abstract interface design and implementation.

During our research we have also found several other methodological approaches that target specific contexts or scenarios, such as:

- The Business Process-Based Methodology (BPBM) (Arch-int & Batanov, 2003), which blends advantages of the structured and object-oriented paradigms for identifying and designing business components. The central idea of business component modeling is reusability of elementary units, which are business activities. An

elementary unit that represents an atomic changeable business process can be implemented with a portable set of Web-based software components;
- The Internet Commerce Development Methodology (ICDM) (Standing, 2002), which is focused on the development of B2C e-commerce applications, emphasizing not only technical aspects, but also strategic, business and managerial aspects.

Some of the methods presented above also propose a modeling language that better suits its purposes, such as WAE and UWE . In the next subsection, some of them are briefly presented.

Modeling Languages for Web Engineering

Modeling languages define notations to be used on the creation of abstract models to solve problems. The Unified Modeling Language (UML) (Booch et al., 2005), for instance, is a modeling language that defines on its metamodel standardized notations for different kinds of models, such as class diagrams and use case diagrams. However, UML does not define when and to which purpose each model should be used.

Hence, methodologies usually present their own modeling language or, as is most commonly seen, use and extend UML, defining a UML Profile. For this purpose, UML includes extension mechanisms, such as stereotypes (definition of a new model element based on an existing one), tagged values (attachment of arbitrary textual information to elements using label/value pairs) and constraints (semantic specification for an existing element, sometimes using a formal language).

Based on these extension mechanisms, Conallen (2002) proposed the Web Application Extensions (WAE), which extends UML to provide Web-specific constructs for modeling WebApps. WAE also advocates the construction of a new model, the User Experience (UX) Model, which defines guidelines for layout and navigation modeling from requirements specification through design. Models, like the navigation diagram, the class diagram and the component diagram (the last two specific for the web tier), use WAE to represent Web components such as screens, server pages, client pages, forms, links and many more.

The UML-based Web Engineering (UWE) (Koch et al., 2000) also defines a UML profile. Based on class and association elements, it defines new elements to describe Web concepts, such as navigation, indexes, guided tours, queries, menus and many others.

Another method that defines a modeling language based on UML is OOWS (Fons et al., 2003). For the construction of the navigational model, UML packages represent navigational contexts and form a directed graph where the arcs denote pre-defined valid navigational paths. Each context is further modeled using a class diagram to show the navigational classes that form them.

Not all modeling languages are UML-based. WebML (Ceri et al., 2000) is an example. It allows developers to model WebApp's functionalities in a high level of abstraction, without committing to any architecture in particular. WebML is based on XML, but uses intuitive graphical representations that can easily be supported by a CASE tool. Its XML form is ideal for automatic generation of source code, producing Web applications automatically from the models.

Methods and modeling languages aid developers mostly during analysis and design of information systems. However, one can also find tools that focus on the implementation phase. In the next subsection, we discuss frameworks that have been extensively used for the development of WISs.

Frameworks for Web Development

WISs have very similar architectural infrastructure. Consequently, after the first systems started to be built, several frameworks that generalize this infrastructure were developed to be reused in future projects. In this context, a framework is viewed as a code artifact that provides ready-to-use components that can be reused via configuration, composition or inheritance. When combined, these frameworks allow large-scale n-tier WISs to be constructed with less coding effort.

Putting together several of these frameworks can produce what we call a container-based architecture. A container is a system that manages objects that have a well-defined life cycle. A container for distributed applications, such as the applications servers for the Java Enterprise Edition (Shannon, 2003), manage objects and offer services such as persistence, transaction management, remoting, directory services, etc.

The use of these frameworks or container-based architectures has a considerable impact in the development of a WIS. Since it is possible to find many frameworks for the exact same task, we categorized them according their objectives into the following classes:

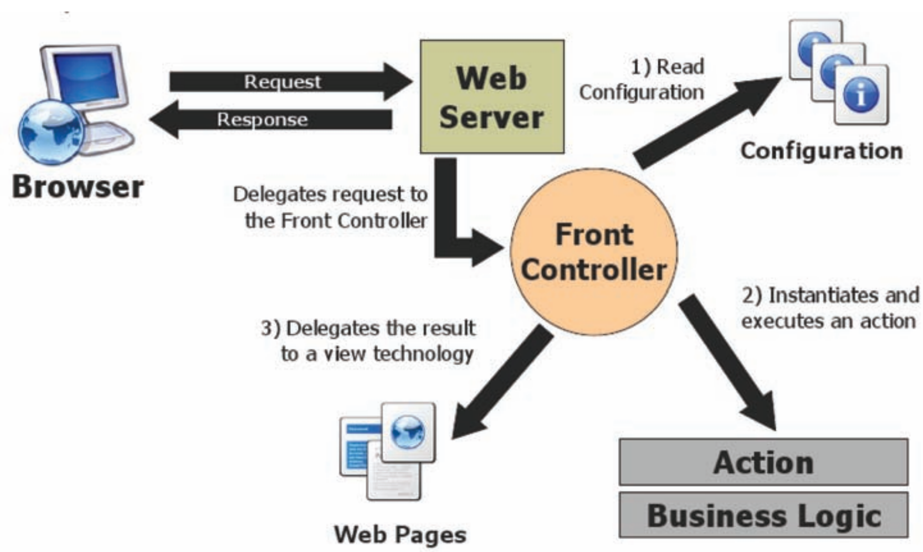
- Front Controller (or MVC) frameworks;
- Decorator frameworks;
- Object/Relational Mapping frameworks;
- Dependency Injection frameworks;
- Aspect-Oriented Programming frameworks;
- Authentication & Authorization frameworks.

Front Controller Frameworks

MVC stands for Model-View-Controller (Gamma et al., 1994). It is a software architecture that was developed by the Xerox PARC for the Smalltalk language in 1979 (Reenskaug, 1979) and has found great acceptance by Web developers. When applied to the Web, the MVC architecture is adapted and receives the name "Front Controller" (Alur et al., 2003, p.166). Both terms are used indistinguishably by Web developers.

The Front Controller architecture is depicted in Figure 1. When structured in this architecture, a WebApp manages all requests from clients using an object known as Front Controller. Based on a customizable configuration, this object decides which class will respond to the current request (the action class). Then, following the Command design pattern (Gamma et al., 1994), it instantiates an object of that class and delegates the control

Figure 1. General architecture of a Front Contoller framework



to it, expecting some kind of response after its execution. Based on that response, the controller decides the appropriate view to present as result, such as a web page, a report, a file download, among other possibilities.

One of these possibilities is using a template engine that defines a template language that is usually more suitable for the view layer than the usual dynamic web technology (such as JSP, ASP or PHP). The template language is usually simpler, making it possible for Web Designers without specific programming skills to build them. Also, they tend to help developers not to break the MVC architecture by restricting what can be done in the template language (e.g. can not directly connect to a database from a template).

MVC Frameworks usually provide the front controller, a super-class or interface for action classes, several result types and a well defined syntax for the configuration file. The template engine is a separate tool, but the framework usually provides integration to it. Note that on n-tier applications, this framework belongs to the Web tier and should delegate business and persistence tasks to components on appropriate tiers.

Only for the Java platform, for instance, there are more than 50 MVC frameworks. Some of the most popular are Struts[1], Spring MVC[2] and Tapestry[3].

Decorator Frameworks

Decorator frameworks automate the otherwise tedious task of making every web page of the site have the same layout, meaning: header, footer, navigational bar, color schemes and other graphical layout elements produced by a Web design team. Figure 2 shows how a decorator framework works.

They work like the Decorator design pattern (Gamma et al., 1994), providing a class that intercepts requests and wraps their responses with an appropriate layout before it is returned to the client. It also provides dynamic selection of decorators, making it easy to create alternate layouts, such as a "print version" of the page. Examples of this kind of framework are SiteMesh[4] and Tiles[5].

Figure 2. The process of decoration of websites

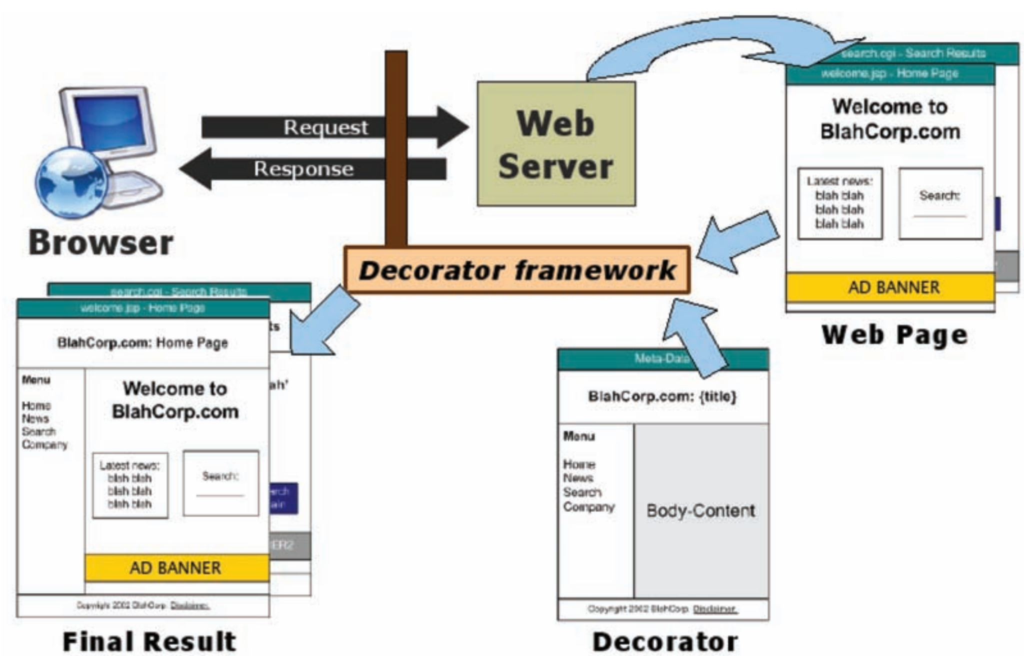
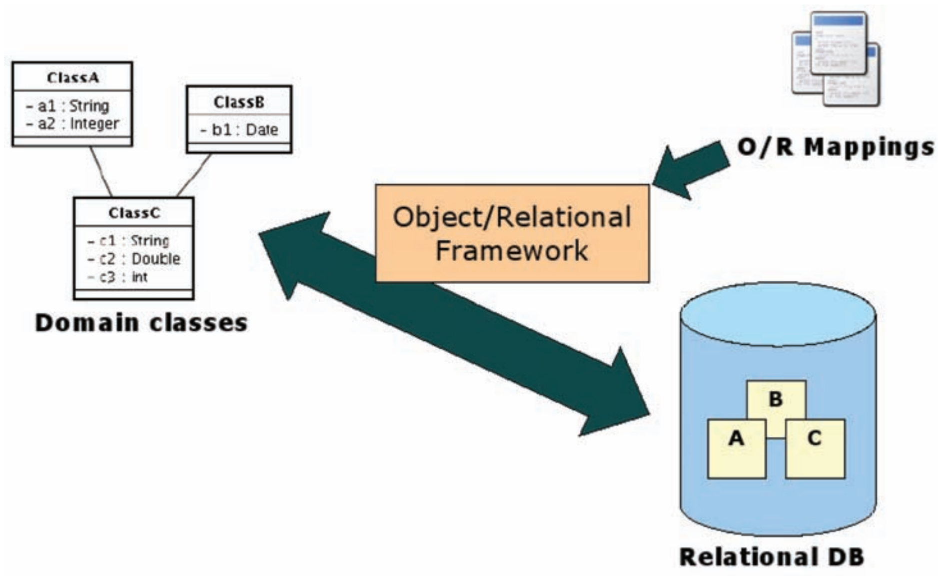


Figure 3. Persistence of objects using an ORM framework



Object/Relational Mapping Frameworks

Relational Database Management Systems (RDBMS) have long been the *de facto* standard for data storage. Because of its theoretical foundations (relational algebra) and strong industry, even object oriented applications use it for object persistence, giving rise to a "paradigm mismatch" (Bauer & King, 2004): tables, rows, projection and other relational concepts are quite different from a graph of interconnected objects and the messages they exchange.

Among the many options to deal with this problem, there is the Object/Relational Mapping (ORM) approach, shown in Figure 3, which is the automatic and transparent persistence of objects to tables of a RDBMS using meta-data that describe the mapping between both worlds (Bauer & King, 2004). Instead of assembling a string with the SQL command, the developer provides mapping meta-data for the classes and call simpler commands, such as save(), delete() or retrieveById(). An object-oriented query language can also be used for more complex retrievals.

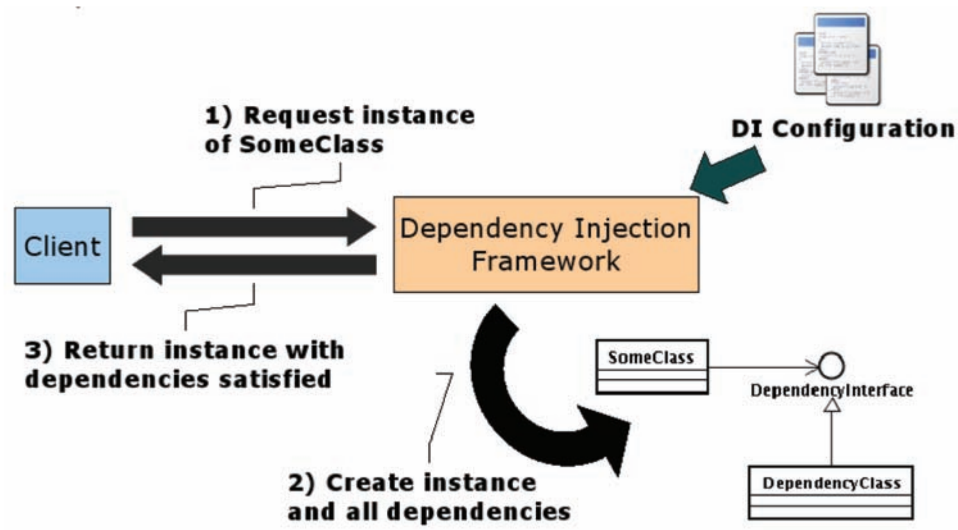
The use of ORM frameworks is not restricted to Web applications and has been in use for quite some time now in all kinds of software. The most popular Java ORM framework is Hibernate[6]. Other well-known frameworks are Java Data Objects[7], Apache Object Relational Bridge[8] and Oracle Toplink[9].

Dependency Injection Frameworks

Object-oriented applications are usually built in tiers, each of which having a separate responsibility. According to Fowler (2007), when we create classes that depend on objects of other classes to perform a certain task, it is preferred that the dependent class is related only to the interface of its dependencies, and not to a specific implementation of that service.

Creational design patterns, such as Factory Method, Abstract Factory and Builder (Gamma et al., 1994), help implementing this good practice in programming, known today as "programming to interfaces, not implementations" (Schmidt, 2007). For instance, if a service class depends on a data access class, it does not need to know

Figure 4. Dependency injection using a framework



how the data access class will perform its duty, but only *what* it will do and what method should be called for the job to be done.

Dependency Injection (DI) frameworks allows the developer to program to interfaces and specify the concrete dependencies as meta-data in a configuration file. When a certain object is obtained from the DI framework, all of its dependencies are automatically injected and satisfied. An abstract example is shown in Figure 4: when the client asks for an instance of SomeClass, the DI framework first satisfies SomeClass' dependencies and delivers the object with all dependencies fulfilled – in the example, an instance of DependencyClass.

These frameworks are also known as Inversion of Control (IoC) frameworks, since the control (who creates the objects) is removed from the dependent classes and given to the framework. As well as ORM frameworks, DI frameworks are not used exclusively for WebApps, although they tend to integrate more seamlessly with applications that run inside containers, just like a WebApp runs inside a Web server. Lots of frameworks provide this service, including Spring Framework, PicoContainer[10], Apache Hivemind[11], etc.

Aspect-Oriented Programming Frameworks

The Aspect-Oriented paradigm is based on the concept of separation of concerns: the idea is to separate different concerns of a system to be treated separately, thus reducing the complexity of development, evolution and integration of software (Resende & Silva, 2005). Although it concerns the whole development process, its biggest influence is at the coding phase, with Aspect Oriented Programming (AOP).

Once a *cross-cutting concern* is identified (e.g.: logging, transaction management), instead of repeating similar code in different points, the functionality can be implemented in a single place, becoming an *aspect*. Then, the different places where that aspect should be applied are identified (these are called *pointcuts*) and, before the code is executed, a process called *weaving* is conducted to automatically spread the aspect all over the code.

The weaving can be conducted by an AOP framework during runtime or by an AOP compiler during compilation time. Many infrastructure concerns that are usual in Web applications are good candidates for this separation, making AOP frameworks very popular. One example, depicted in Figure 5, is that of transaction management. An AOP framework can make all business methods transactional with few configuration steps, avoiding the effort of repeatedly implementing the same logic in all of them.

Some well-known AOP frameworks for the Java platform are AspectJ[12], Spring Framework and JBoss AOP[13].

Authentication & Authorization Frameworks

Another common concern of Web information systems is that of guaranteeing the security of the information. This is usually done by two different procedures: authentication (verifying if an access key is valid to access the application) and authorization (verifying the level of access of the authenticated user and what she is allowed to do).

Being such an important task, frameworks were created to guarantee its proper execution. They can be configured to support many different "auth" methods, using, as usual, meta-data and configuration files. Some well-known auth frameworks for the Java platform are Acegi Security for Spring[14], Apache Cocoon Authentication[15] and the Java Authentication and Authorization Services[16].

In spite of frameworks being much used, there is no Web Engineering method that explores their use in the design phase of the software process. To fill this gap, we proposed FrameWeb, a Framework-based Design Method for Web Engineering (Souza & Falbo, 2007), which is presented later in this chapter.

The Semantic Web

The Semantic Web is being proposed as an evolution of the current WWW, in which information is provided both in human-readable and computer-processable formats, in order to allow for the semi-automation of many tasks that are conducted on the Web.

In order for the software agents to reason with the information on the Web (reasoning meaning that the agents are able to understand it and take sensible actions according to a predefined goal), web pages have to be presented also in a machine-readable form. The most usual way for this is annotating the pages using formal knowledge representation structures, such as ontologies.

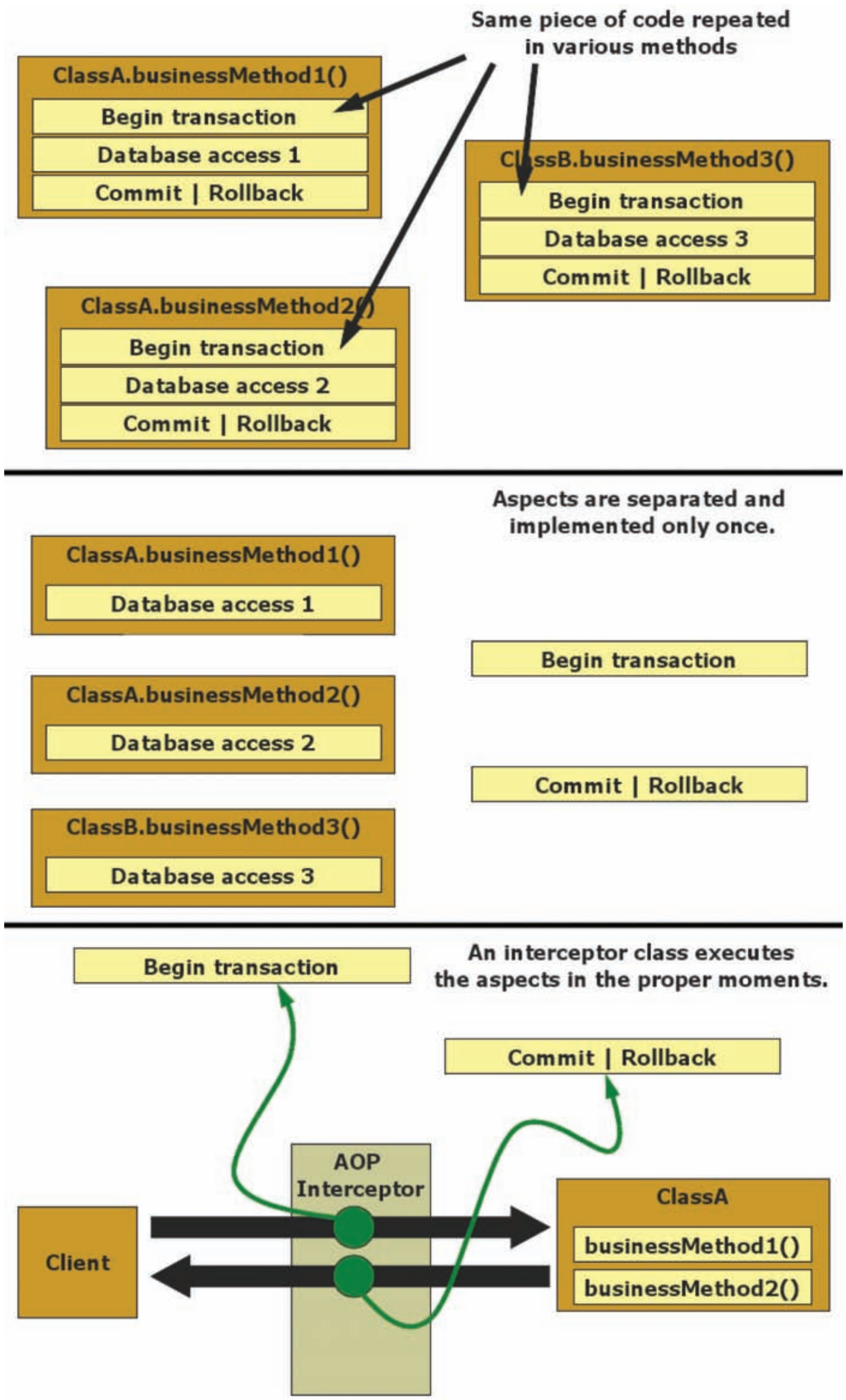
An ontology is an engineering artifact used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of its vocabulary words (Guarino, 1998). Along with ontology representation languages such as OWL (W3C, 2007a), they are able to describe information from a website in formal structures with well-defined inference procedures that allow software agents to perform tasks such as consistency checking, to establish relationships between terms and to systematically classify and infer information from explicitly defined data in this structure.

Designing an ontology is not a straightforward task. There are many methodologies for their construction (Gomez-Perez et al., 2005) and attention has to be given to the selection of concepts, their properties, relationships and constraints. However, after the ontology is built, the annotation of static Web pages with languages such as OWL becomes a simple task, especially with the aid of tools, such as OILEd[17] and Protégé[18].

However, several websites have their Web pages dynamically generated by software retrieving information from data repositories (such as relational databases) during runtime. Since these pages cannot be manually annotated prior to their presentation to the visitor, another approach has to be taken. Two approaches that have been proposed are dynamic annotation and semantic Web services.

The former works by recognizing whether the request belongs to a human or a software agent, generating the proper response depending on the

Figure 5. example of application of AOP using an AOP runtime framework



client: in the first case, a HTML human-readable Web page; in the second, a document written in an ontology specification language containing meta-data about the information that would be displayed in the HTML version. Although the solution seems appropriate, many aspects still need to be addressed, such as: how are the agents supposed to find the Web page? How will they know the correct way to interact with it? For instance, how will they know how to fill in an input form to submit to a specific request?

The latter approach is based on Web Services, which are software systems designed to support interoperable machine-to-machine interaction over a network (W3C, 2007b). Web Services provide a nice way for software agents to interact with other systems, requesting services and processing their results. If semantic information is added to the services, they could become interpretable by software agents. Meta-data about the service are written in a markup language, describing its properties and capacities, the interface for its execution, its requirements and the consequences of its use (McIlraith et al., 2001). Many tasks are expected to be automated with this, including service discovery, invocation, interoperation, selection, composition and monitoring (Narayanan & McIlraith, 2002).

As the research on the Semantic Web progresses, methods are proposed to guide developers on building "Semantic Web-enabled" applications. An example of this is the Semantic Hypermedia Design Method (SHDM) (Lima & Schwabe, 2003). Based on OOHDM (Schwabe & Rossi, 1998), SHDM is a comprehensive model-driven approach for the design of Semantic WebApps.

SHDM's process is divided in 5 activities. In the first step, Requirements Gathering, requirements are gathered in the form of scenarios, user interaction diagrams and design patterns. The next phase, Conceptual Design, produces a UML-based conceptual model, which is enriched with navigational constructs in the Navigational Design phase. The last two activities are Abstract Interface Design and Implementation.

Some Considerations

In our research, we haven't found a method focused on the use of frameworks for the construction of WISs nor for the development of Semantic Web applications. In the next sections, we present FrameWeb, our proposal for the design of framework-based WISs, and its extension, S-FrameWeb, which incorporate into the method activities and guidelines that drive the developer in the definition of the semantics of the WISs, resulting in a "Semantic Web-enabled" application.

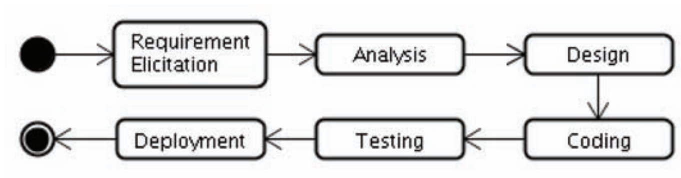
FRAMEWEB

FrameWeb is a design method for the construction of Web-based Information Systems (WISs) based on frameworks. The main motivations for the creation of this method were:

1. The use of frameworks or similar container-based architectures has become the *de facto* standard for the development of distributed applications, especially those based on the Web;
2. There are many propositions in the area of Web Engineering, including methodologies, design methods, modeling languages, frameworks, etc. However, we haven't found one that deals directly with the particularities that are characteristic of the use of frameworks;
3. Using a method that fits directly into the architecture chosen for the implementation promotes a greater agility to the software process, which is something that is desired in most Web projects.

In general, FrameWeb assumes that certain types of frameworks will be used during the

Figure 6. A simple software process suggested by FrameWeb



implementation, defines a basic architecture for WISs and proposes design models that are closer to the implementation of the system using these frameworks.

Being a design method, it doesn't prescribe a complete software process. However, it suggests the use of a development process that includes the following activities, as presented in Figure 6: requirements elicitation, analysis, design, coding, testing and deployment. For a more systematic usage of the method, it also suggests that, during Requirement Elicitation and Analysis, use case diagrams are used to model requirements and class diagrams are used to represent the conceptual model.

Also, as mentioned earlier, one of the motivations for the creation of FrameWeb is the demand for agility that surrounds Web projects. Thus, although the method brings more agility especially to the design and coding phases, developers are advised to follow principles of agility during requirements analysis, as the ones proposed by Agile Modeling (Ambler & Jeffries, 2002).

The main contributions of the method are for the Design phase: (i) the definition of a basic architecture that divides the system in layers with the purpose of integrating better with the frameworks; (ii) a UML profile for the construction of four different design models that bring the concepts used by the frameworks to the design stage of the software process.

Figure 7. A simplified use case diagram for LabES Portal

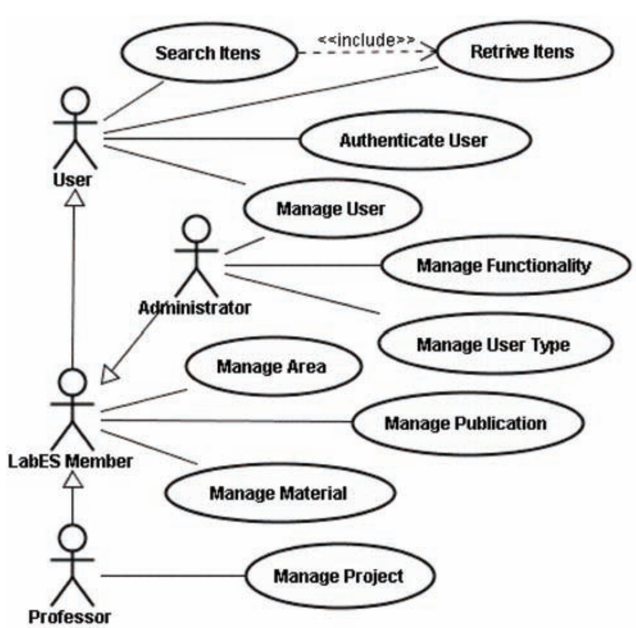
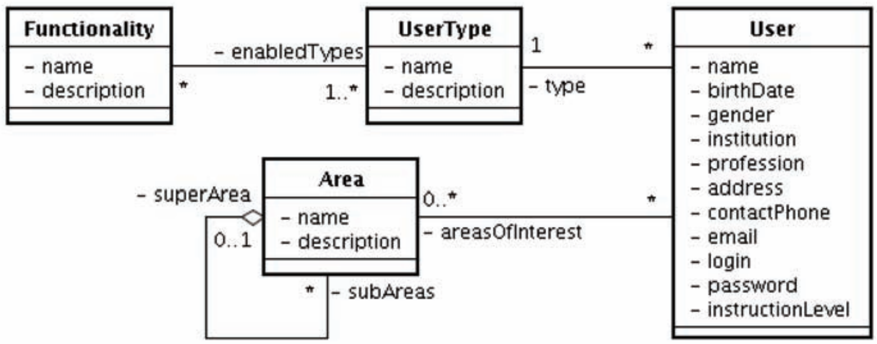


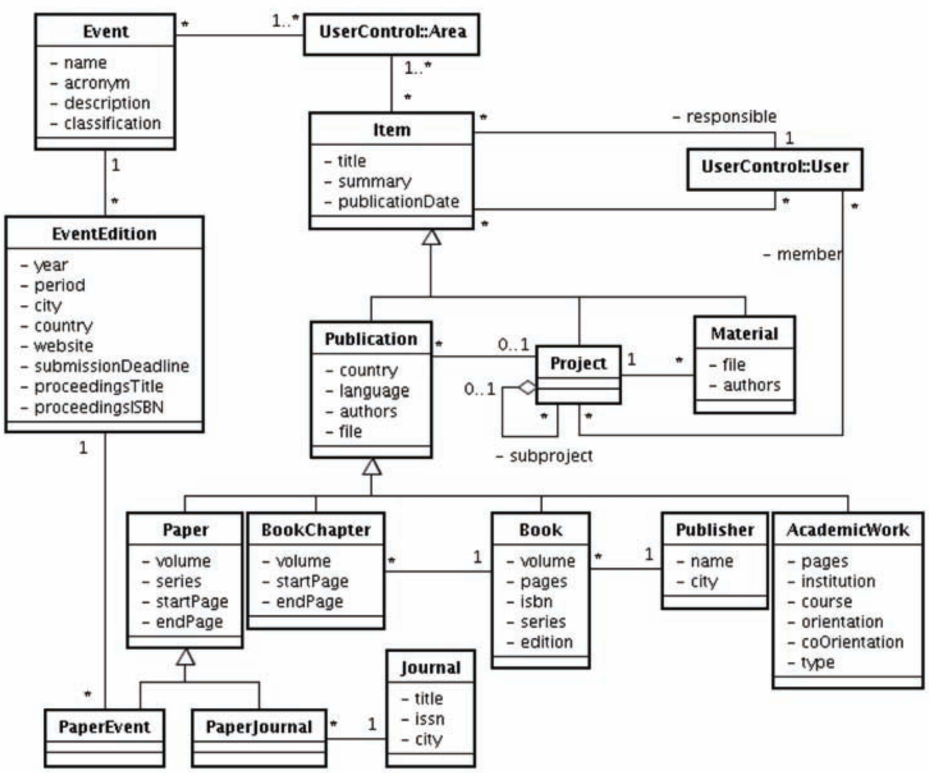
Figure 8. Conceptual model for the User Control module of the LabES Portal



The Coding phase is facilitated by the use of frameworks, especially because design models show components that can be directly related to them. The use of frameworks can also have impacts on Testing and Deployment, but these are yet subject to study and research.

Throughout the next subsections we detail FrameWeb's basic architecture and its UML pro-

Figure 9. Conceptual model of the Item Control module of the LabES Portal



file. Examples diagrams were taken from the development of a portal for the Software Engineering Lab (LabES) of the Federal University of Espírito Santo State using FrameWeb. Figure 7 shows its use case diagram, simplified for brevity.

The "LabES Portal" was proposed to provide a better interaction with the Software Engineering community. This WIS has a basic set of services providing information about current LabES projects, areas of interest, publications and other material available for download. Figures 8 and 9 show the conceptual models produced during Analysis.

Basically, the portal makes a collection of items available. These items can be organized in projects and subprojects or belong to the lab in general. Publications (papers, books, book chapters and academic works) and generic materials can be published in the portal. Items are also related to users (responsible user, editing users) and areas of interest.

Framework-Based WebApp Architecture

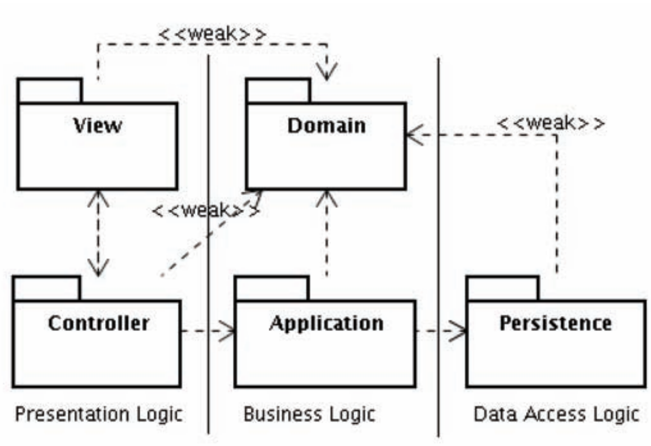
The Design activity, traditionally executed after requirement elicitation and analysis, has as purpose the description of the logical and physical architectures of the system as well as the development of structural and behavioral models built based on the models developed in the previous phases, but that now consider the specific characteristics of the chosen implementation platform.

FrameWeb defines a logical architecture for WISs based on the architectural pattern Service Layer (Fowler, 2002, p. 133). As depicted in Figure 10, the system is divided in three main layers: presentation logic, business logic and data access logic.

The first layer concerns the graphical user interfaces. The View package contains the Web pages, style sheets, images, client-side scripts, templates and everything else related to the exhibition of information to the user. The Controller package encompasses action classes and other files related to the Front Controller framework. These two packages are mutually dependent, since View elements send user *stimuli* to Controller classes while these process the response using pages, models and other View components.

The business logic is implemented in the second layer, divided in two packages: Domain and Application. The former contains classes that represent concepts of the problem domain identified and modeled by the class diagrams during analysis and refined during design. The latter

Figure 10. FrameWeb's basic architecture for WISs



has the responsibility of implementing the use cases defined in the requirements specification, providing a service layer independent of the user interface. The Application classes deal directly with Domain objects to implement system functionality and, thus, this dependency is represented in the diagram.

The Controller package, on the presentation layer, depends on the Application package since it mediates the user access to the system functionalities. User *stimuli* coming from View are transformed by the Controller's classes in method calls to classes in the Application package. Controller and View have also dependency relationships with Domain, but this is tagged as weak to denote low coupling: Domain objects are used only for exhibition of data or as parameters on method invocations between one package and another, i.e., the presentation layer does not have the right to alter domain entities.

The third and last layer regards data access and has only the Persistence package. This package is responsible for the storage and retrieval of persistent objects in long-term duration media, such as databases, file systems, naming services, etc. In the case of FrameWeb, it expects the use of an ORM framework through the Data Access Object (DAO) pattern (ALUR et al., 2003, p. 462). The DAO pattern adds an extra abstraction layer, separating the data access logic of the chosen persistence technology in a way that the Application classes do not know which ORM framework is being used, allowing for its replacement, if necessary. It also facilitates unit testing, as one can provide mock DAOs for the Application classes to be tested alone.

As we can see in Figure 10, the Application package depends on the Persistence package to retrieve, store and delete domain objects as the result of use case execution. Since the Persistence package works with Domain objects, a weak dependency is also portrayed in the figure.

This architecture provides a solid base for the construction of WISs based on the types of frameworks presented earlier in this chapter. Each package contains classes or other elements that integrate with these frameworks and, to model all these elements, FrameWeb proposes a modeling language based on the UML, which is presented next.

Modeling Language

During design, besides specifying the system architecture, the artifacts that will be implemented by the programmers on the coding phase should be modeled. Since FrameWeb is based on the frameworks presented earlier, we felt the need for a modeling language that would represent the concepts that are present in these frameworks.

Following the same approach as other modeling languages such as WAE and UWE, FrameWeb uses UML's lightweight extensions to represent typical Web and framework components, creating a UML profile that is used for the construction of four kinds of diagrams, which are presented in the following subsections: domain model, persistence model, navigation model and application model.

Domain Model

The domain model is a UML class diagram that represents domain objects and their persistence mapping to a relational database. This model is used by the programmers to implement the classes of the Domain package. FrameWeb suggests its construction in two steps:

1. Adapt the conceptual model produced during the Requirement Analysis phase to FrameWeb's architecture and to the chosen platform of implementation. This requires choosing data types for attributes, defining navigabilities of the associations, promoting attributes to classes (if necessary), etc.;
2. Add persistence mappings.

Persistence mappings are meta-data that allow ORM frameworks to convert objects in memory to tuples in Relational Data Base Management Systems and vice-versa. Mappings are added to the domain model using stereotypes and constraints that guide developers in the configuration of the ORM framework during implementation. Despite the fact that these mappings are more related to persistence than domain, they are shown in this model because the classes that are mapped and their attributes are shown here.

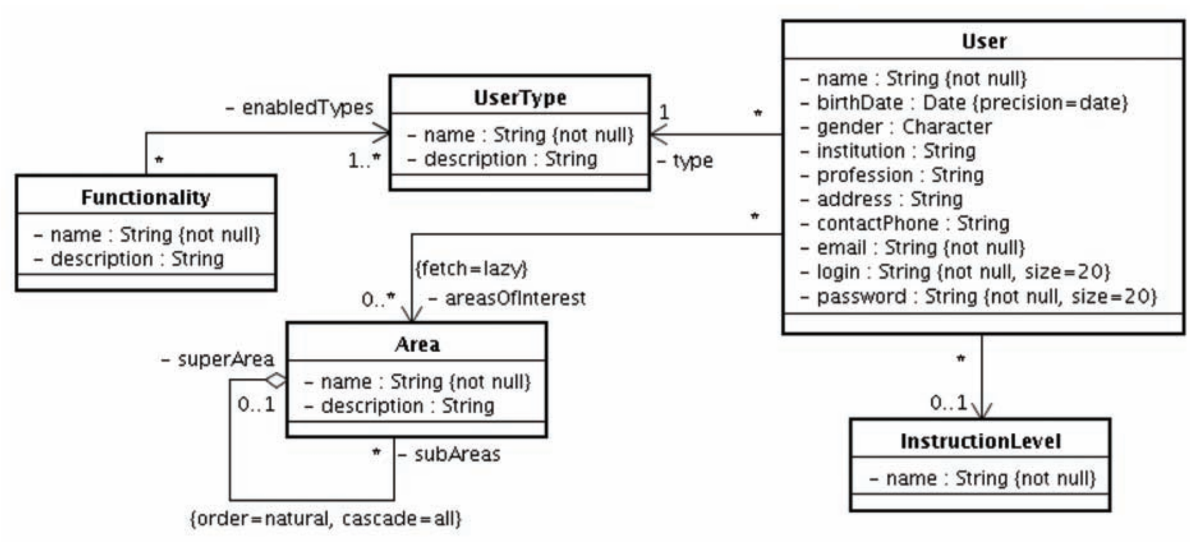
Table 1 describes the possible O/R mappings for the domain model. For each mapping, the table presents the extension mechanism used and what are its possible values or syntax. None of the mappings is mandatory and most of them have sensible defaults, reducing the amount of elements that have to be modeled. The default values are shown in the third column in boldface.

The Domain Model for the User Control module of sLabES Portal is shown in Figure 11. According to the default values, all classes are

Table 1. Possible OR mappings for the Domain Model

Mapping	Extension	Possible Values
If the class is persistent, transient or mapped (not persistent itself, but its properties are persistent if another class inherits them)	Class stereotype	**<<persistent>>** <<transient>> <<mapped>>
Name of the table in which objects of a class will be persisted	Class constraint	table=*name* **(default: class' name)**
If an attribute is persistent or transient	Attribute stereotype	**<<persistent>>** <<transient>>
If an attribute can be null when the object is persisted	Attribute constraint	**null** not null
Date/time precision: store only the date, only the time or both (time-stamp)	Attribute constraint	precision = (date | time | **time-stamp**)
If the attribute is the primary-key of the table	Attribute stereotype	<<id>>
How the ID attribute should be generated: automatically, obtained in a table, use of IDENTITY column, use of SEQUENCE column or none	Attribute constraint	generation = (**auto** | table | identity | sequence | none)
If the attribute represents the versioning column.	Attribute stereotype	<<version>>
If an attribute should be stored in a large object field (e.g.: CLOB, BLOB)	Attribute stereotype	<<lob>>
Name of the column in which an attribute will be persisted	Attribute constraint	column=*name* **(defaults to the attribute's name)**
Size of the column in which an attribute will be persisted	Attribute constraint	size=*value*
If the association should be embedded (instead of having its own table, the associated child class' attributes are placed in the parent's table)	Attribute stereotype	<<embedded>>
Inheritance mapping strategy: one table for each class using UNION, one table for each class using JOIN or single table for the entire hierarchy	Inheritance stereotype	<<union>> <<join>> **<<single-table>>**
Type of collection which implements the association: bag, list, set or map	Association constraint	collection = (bag | list | **set** | map)
Order of an association's collection: natural ordering (implemented in code) or order by columns (ascending or descending)	Association constraint	order = (natural | *column names* [asc | desc])
Cascading of operations through the association: nothing, persists, merges, deletions, refreshs or all	Association constraint	cascade = (**none** | persist | merge | remove | refresh | all)
Association fetching strategy: lazy or eager.	Association constraint	fetch = (lazy | **eager**)

Figure 11. Domain Model for the User Control module of LabES Portal



persistent and class and attribute names are used as table and column names respectively.

As we can see in the diagram, attributes have received mappings such as nullability and size. The birthDate attribute was mapped as date-only precision. The recursive association in Area was configured to be sorted naturally (will be implemented in the programming language) and to cascade all operations (e.g. if an area is deleted, all of its subareas are automatically deleted).

None of the classes have ID or version attributes because they are inherited from a utility package, as shown in Figure 12. The mapped stereotype indicates that DomainObjectSupport and HibernatePersistentObject are not persistent entities, but their subclasses, which are entities, inherit not only their attributes but also their O/R mappings. All domain classes in LabES Portal are said to extend HibernatePersistentObject, inheriting, thus, the UUID[19], the persistence ID and the version attribute.

The parameters I and V are generic, allowing for the user to choose the type of ID and version attributes. HibernateBaseDAO is a base class for data access objects, described in the persistence model, discussed in the next subsection.

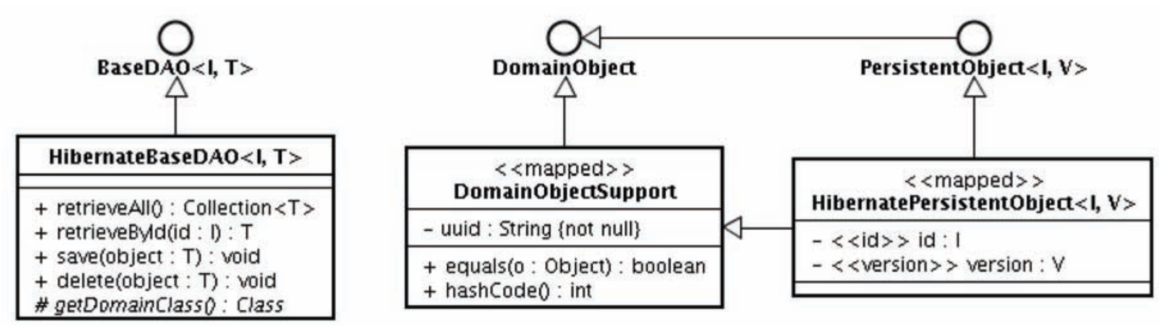
Persistence Model

As mentioned before, FrameWeb indicates the use of the DAO design pattern (ALUR et al., 2003, p. 462) to the construction of the data access layer. Thus, the persistence model is a UML class diagram that represents DAO classes responsible for the persistence of the domain classes. Therefore, it guides the implementation of the classes from the Persistence package. FrameWeb suggests three steps for its construction:

1. Model the interface and concrete implementation of the base DAO (an example is shown in Figure 12);
2. Define which domain classes need basic persistence logic and create a DAO interface and implementation for each one;
3. For each DAO, evaluate the need of specific database queries, adding them as operations in their respective DAOs.

The persistence model presents, for each domain class that needs data access logic, an interface and a concrete DAO that implements the interface. The interface has to be unique and

Figure 12. Utility classes for persistence



defines the persistence methods for a specific domain class. One concrete class is modeled for each persistence technology used.

To avoid repeating in each DAO operations that are common in all of them (e.g.: save, delete, retrieve by ID, etc.), a Base DAO (interface and implementation class) is modeled in a utility package. Automatically all DAO interfaces inherit from the BaseDAO interface and the same happens with concrete implementations, without the need to explicitly state that in the diagram. Also, to avoid repeating methods in the interface and implementations, the designer can choose to display them in one of the two only and it is inferred that all public methods are defined in the interface and implemented in the concrete class.

Figure 12 shows the interface and implementation using Hibernate ORM framework, designed for the LabES Portal project. Both interface and class and declared using generic types, leaving to their subclasses to specify which class is being persisted and what is the type for its ID attribute. The Base DAO defines methods to retrieve all persistent entities of a given class, retrieve an entity given its ID, save and delete an entity. As stated before, all public methods modeled in HibernateBaseDAO are inferred to be defined in the BaseDAO interface.

Figure 13. Persistence model of the User Control module of LabES Portal

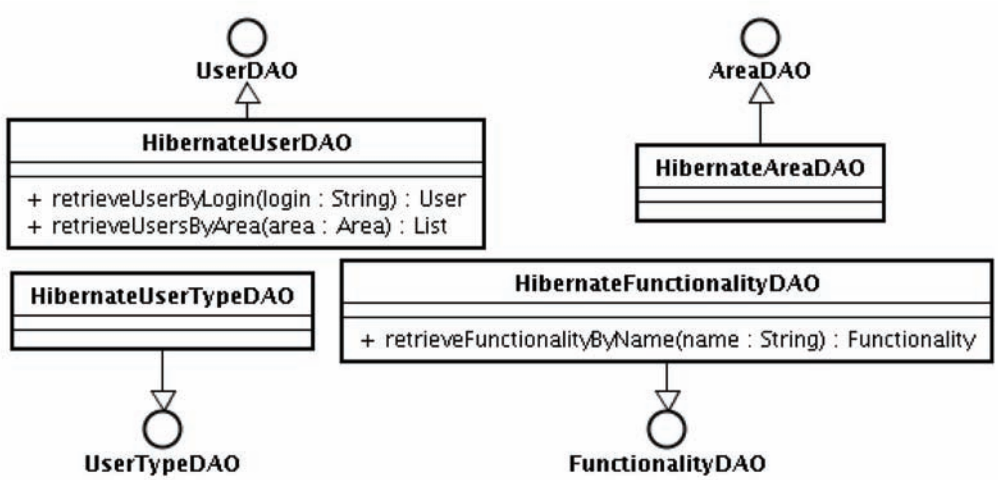


Figure 13 shows the modeling of four DAOs from the LabES Portal project, for the persistence of the classes in the User Control module. AreaDAO and UserTypeDAO are simple, as they inherit all basic operations from the Base DAO and don't need to define any extra ones. The other two define extra operations. For example, UserDAO defines an operation to retrieve all users that have a given area of interest. This is necessary because there is no navigability from Area to User (see Figure 11) and the "Manage Area" use case needs to prevent an area from being deleted if it is associated with any user.

As we can see, the persistence model does not define any UML extensions to represent the concepts that are needed to implement the data access layer, but only some rules that make this modeling simpler and faster.

Navigation Model

The navigation model is a UML class diagram that represents different components that form the presentation layer, such as Web pages, HTML forms and action classes from the Front Controller framework. Table 2 shows the UML stereotypes used by the different elements that can be represented in a navigation model. This model is used by developers to build classes and components of the View and Controller packages.

For Web pages and templates, the attributes of the classes represent information from the domain that is supposed to be displayed in the page. Dependency relationships between them indicate hyperlinks while composition associations between pages and forms denote the presence of the form in that page.

In HTML forms, attributes represent the form fields and their types follow the HTML standard for types of fields (e.g.: input, checkbox, etc.) or the names of the JSP tags used by the framework (e.g., for Struts2, textfield, checkbox, checkboxlist, etc.).

The action class is the main component of the model. Its dependency associations show the control flow when an action is executed. Table 3 lists the different meanings of this kind of association, depending on the components that are connected by it. Dependencies that are navigable towards an action class represent method calls, while the others represent results from the action execution.

The attributes of the action class represent input and output parameters relevant to that action. If there is a homonymous attribute in an HTML form being submitted to the action, it means that the data is injected by the framework in the action class (input parameter). Likewise, when one of the result pages/templates show an attribute with the same name of an attribute of the action class, this indicates that the framework makes this information available for the output.

When an action is executed, the framework will execute a default action method or allow/request the explicit definition of which method to execute. In the latter case, the designer must specify which method is being executed using the constraint {method=method-name} in the

Table 2. UML stereotypes used in the navigation model

Stereotype	What it represents
(none)	An action class, to which the Front Controller framework delegates the execution of the action.
<<page>>	A static or dynamic Web page.
<<template>>	A template that is processed by a template engine and is transformed into a Web page.
<<form>>	A HTML form.
<<binary>>	Any binary file that can be retrieved and displayed by the browser (e.g.: images, reports, documents, etc.).

Table 3. Dependency associations between an action class and other elements

From	To	What it represents
Page / template	Action class	A link in the page/template that triggers the execution of the action.
Form	Action class	Form data are sent to the action class when the form is submitted.
Action class	Page / template	The page/template is shown as one of the results of the action class.
Action class	Binary file	A binary file is shown as one of the results of the action class.
Action class	Action class	An action class is executed as result of another. This process is known as “action chaining”.

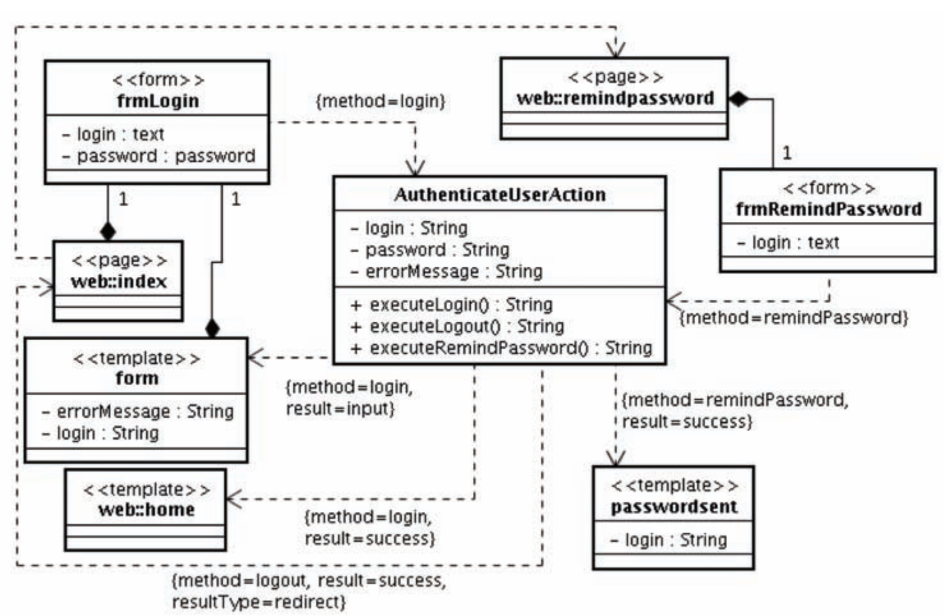
dependency association. The same is true for associations that represent results. Naturally, these methods should be modeled in the diagram.

When modeling action chaining, it’s sometimes necessary to indicate the method that was executed in the first action and the one that will be executed in the following. These can be specified with the constraints outMethod and inMethod.

For dependency associations that represent results there are two other constraints that can be used:

- {result=result name} specifies a keyword that represents this control flow, i.e., when the action class returns this keyword as result of the action execution, the framework will follow this flow and show the appropriate result page/template/binary file;
- {resultType=type name} determines the type of result, among those supported by the framework. Usually, at least the following types of result are available: binary (display a binary file), chain (action chaining), dispatch (dispatches the request), redirect (redirects the request) and template (processes a template using a template engine).

Figure 14. Navigation Model for the use case “Autheticate User”



The difference between a dispatch and a redirection is that the first makes the action's output parameters available to the view, while the second does not. When a dependency association doesn't specify a type, it means it is a dispatch. The default result is defined by the framework.

The designer is free to choose the granularity of the action classes, building one for each use case scenario, one for each use case (encompassing many scenarios), one for multiple use cases, and so forth. Moreover, he/she should decide if it's best to represent many actions in a single diagram or have a separate diagram for each action. Figure 14 is the navigation model for a use case of LabES Portal.

The figure shows that in the initial page of the portal (represented by web::index), there is a form where login and password can be filled. When submitted, this data goes to the action class for the execution of the executeLogin() method, which would access the business logic layer to perform the use case. If the information filled is correct (result = success), the user is taken to web::home, which represents the starting page for authenticated users. Otherwise, the user will be taken back to web::index (result = input), showing once again the login form and an error message.

If the user forgot his/her password, he/she can click on a link in the initial page to go to the web::remindpassword page, where his/her login would be informed and sent to the action class. The executeRemindPassword() method requests the business logic layer to send the password to the user's email address and informs the user that the message has been sent. To log out, the user clicks on the appropriate link and is redirected back to the initial page.

During the conception of FrameWeb, there has been a discussion on whether the navigation model would be better represented by a sequence diagram, as it could represent better the control flow. Two main reasons led to the choice of the class diagram: (a) it provides a better visualization of the inner elements of action classes, pages and forms; and (b) it models composition between pages and forms with a more appropriate notation. Nonetheless, designers are advised to build sequence diagrams to represent complex flows when they see fit.

Last but not least, FrameWeb suggests four steps for the construction of a navigation model:

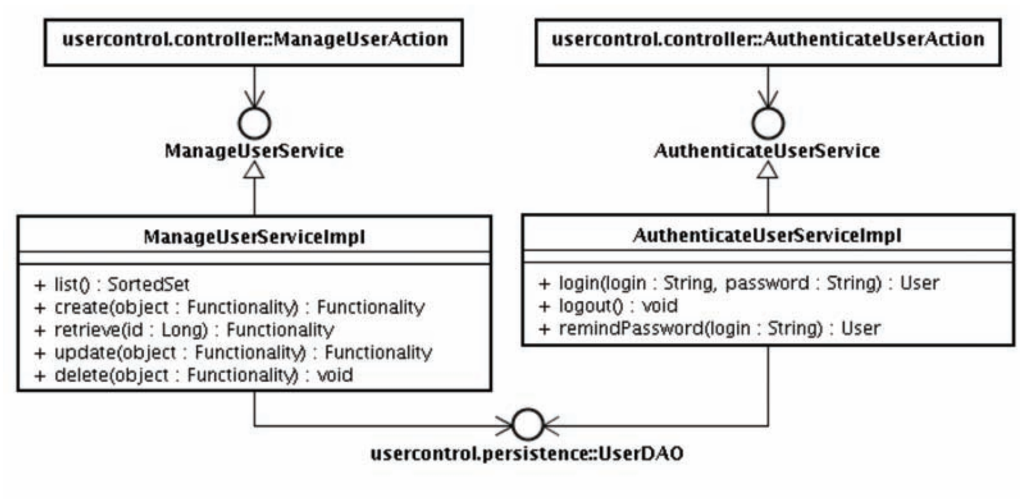
1. Study the use cases modeled during requirements analysis to define the granularity of the action classes (using, preferably, names that can relate the actions to the use cases/scenarios they refer to);
2. Identify how the data gets to the action class, modeling input pages and forms and the appropriate attributes on them and in the action class;
3. Identify what are the possible results and model the output pages/templates/binary files, also adding attributes when appropriate. We suggest that results that come from exceptions should not be modeled to avoid polluting the diagram;
4. Periodically check if the model is getting too complex and consider dividing it into two or more navigation models.

Application Model

The application model is a UML class diagram that represents classes from the Application package and their relationship with the Controller and Persistence packages. Besides guiding the implementation of application classes, this diagram also instructs developers on the configuration of the Dependency Injection framework, which is responsible for managing the dependencies among these three packages.

The granularity of the application classes can be chosen by the developer in the same way as the granularity of the action classes. The application model also shares similarities with the persistence model, as it does not define any UML extension and uses the "programming to interfaces" prin-

Figure 15. Part of an Application Model of the User Control module of LabES Portal



ciple, indicating the modeling of an interface for each application class.

When an application class is modeled, all action classes that depend on it should be displayed in the diagram, with the appropriate namespaces and relationships depicted. Analogously, all DAOs required by the application class to execute the use case should have their interfaces shown in the model, along with the relationship with the application class. Both relationships are represented by directed associations and the multiplicity is not required, as it is always 1.

Figure 15 shows part of an application model of LabES Portal, depicting the classes that implements the "Manage User" and "Authenticate User" use cases and its relationships with controller and persistence components. The methods of the classes represent each scenario of each use case and define the parameters that should be given for them.

Application classes manipulate domain objects and, thus, depend on them. These relationships, however, are not shown in the diagram to avoid increasing the complexity of the model. One can know about these relationships by reading the description of each use case.

FrameWeb suggests four steps for the construction of an application model:

1. Study the use cases modeled during analysis to define the granularity of the application classes (using, preferably, names that can relate the classes to the use cases/scenarios they implement);
2. Add to the interfaces/classes the methods that implement the business logic, giving special attention to the name of the method (as before, with the name of the class), its parameter, the parameters types and its return type;
3. By reading the use case descriptions, identify which DAOs are necessary for each application class and model the associations;
4. Go back to the navigation model (if already built) and identify which action classes depend on which application class and model their associations.

By defining the standard architecture and a UML profile for the construction of these four diagrams, FrameWeb provides the appropriate tools for the design of framework-based WISs.

To promote the construction of "Semantic Web-enabled" WISs, an extension called S-FrameWeb was proposed and it is presented in the next section.

S-FRAMEWEB

The main goal of S-FrameWeb is to make WISs "Semantic Web-enabled". Being a framework-centered method, the chosen approach is to have the Front Controller framework produce dynamic annotations by identifying if requests come from human or software agents. In the former case, the usual Web page is presented, while in the latter, an OWL document is returned.

To accomplish this, S-FrameWeb extends FrameWeb in the following manners:

- The activity of Domain Analysis should be conducted in the beginning of the project to build an ontology for the domain in which the software is based. If it already exists, it should be reused (and eventually modified);
- Requirement Specification and Analysis go as usual, except for the fact that conceptual models build during Analysis can now be based on the domain ontology built in the previous activity;
- During design, FrameWeb's Domain Model (FDM) receives semantic annotations based on the domain ontology;
- During implementation, the MVC framework has to be extended in order to perform dynamic annotation.

Figure 16 shows the software process suggested by S-FrameWeb while Table 4 summarizes the evolution of the models throughout that software process.

The following subsections go through the suggested software process discussing it in more detail.

Domain Analysis

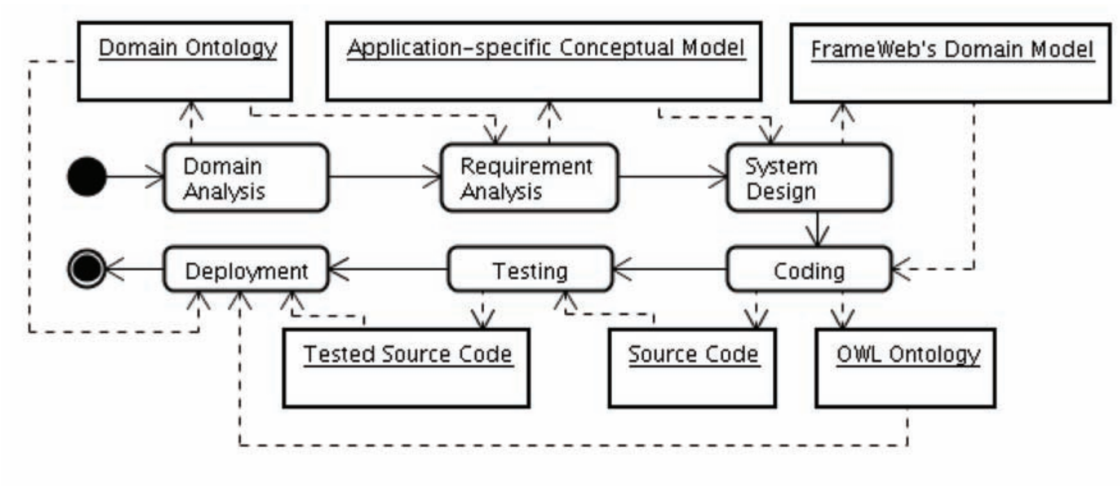
The first step for bringing a WIS to the Semantic Web is formally describing its domain. As discussed previously in this chapter, this can be achieved by the construction of an ontology. S-FrameWeb indicates the inclusion of a Domain Analysis activity in the software process for the development of a domain ontology (we don't use the term "domain model" to avoid confusion with FrameWeb's Domain Model – FDM –, which is a design model).

Domain Analysis is "the activity of identifying the objects and operations of a class of similar systems in a particular problem domain" (Neighbors, 1981; Falbo et al., 2002). When a software is built, the purpose is to solve a problem from a given domain of expertise, such as medicine, sales or car manufacturing. If the domain is analyzed prior to the analysis of the problem, the knowledge that is formalized about the domain can be reused when another problem from the same domain needs a software solution (Falbo et al., 2002).

S-FrameWeb does not impose any specific method for the construction of ontologies. It also doesn't require a specific representation language, but suggests the use of OMG's[20] Ontology Definition Metamodel (ODM) (OMG, 2007), "a language for modeling Semantic Web ontologies in the context of MDA" (Đurić, 2004). ODM defines an ontology UML profile that allows developers to represent ontologies in UML class diagrams.

In the development of the LabES Portal, the SABiO method (Falbo, 2004) was followed, resulting in the construction of an ontology for educational portals that deals with competency questions such as: what are the roles of the people in the educational institution?, what are the areas of interest of these people and the institution?, how is the institution organized?, etc. The ontology was divided into two separate diagrams: one for the general structure of educational protals and another specific for publications. Figure 17 shows the first one.

Figure 16. The software process suggested by S-FrameWeb (SOUZA et al., 2007)



The domain ontology serves as a basis for the construction of the application's conceptual model (during Requirement Analysis), which should derive some classes and associations from the ontology, adding and modifying elements as needed, concerning the specific problem being solved.

Requirement Specification and Analysis

The activities of Requirement Specification and Analysis should be conducted by the development team using its methodology of preference. S-FrameWeb, like FrameWeb, does not prescribe any methods or languages to this phase of the software process. However, as during Domain Analysis, it suggests the use of ODM for the graphical representation of the conceptual model, as it eases its conversion to FDM and, later on, to code (using OWL).

Figure 18 shows the conceptual model for the User Control module of the LabES Portal.

The stereotype <<OntClass>> indicates domain classes, <<ObjectProperty>> models associations between domain classes, <<DataType>> represents XML data types and <<DatatypeProperty>> models associations between classes and data types.

The reader accustomed with UML conceptual models may notice that associations are represented as classes in ODM. This is because in OWL associations are independent from classes and, for instance, can form their own subsumption

Table 4. Models produced by the software process suggested by S-FrameWeb (SOUZA et al., 2007)

Activity	Artifact	What the model represents
Domain Analysis	Domain Ontology	Concepts from the domain to which the software is being built. Modeled in ODM, but converted to OWL for deployment.
Requirement Analysis	Conceptual Model	Concepts that are specific to the problem being solved. Modeled in ODM.
System Design	FrameWeb's Domain Model (FDM)	Same as above plus OR mappings. Modeled using S-FrameWeb's UML profile.
Coding	OWL code	OWL representation of FDM, without OR mappings.

Figure 17. Diagram of the structural part of the ontology for educational portals

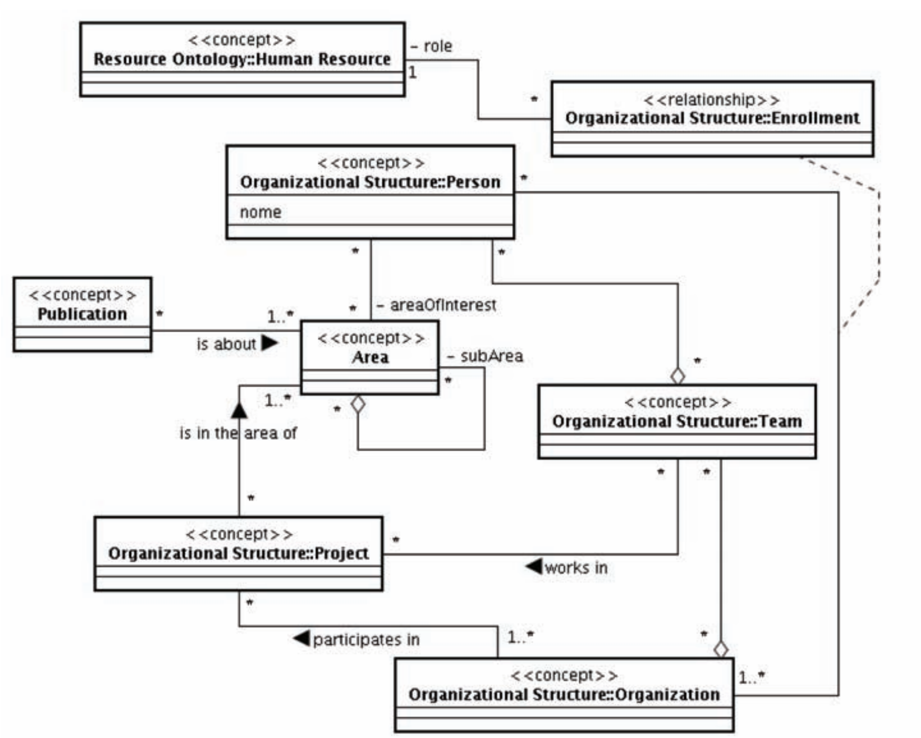
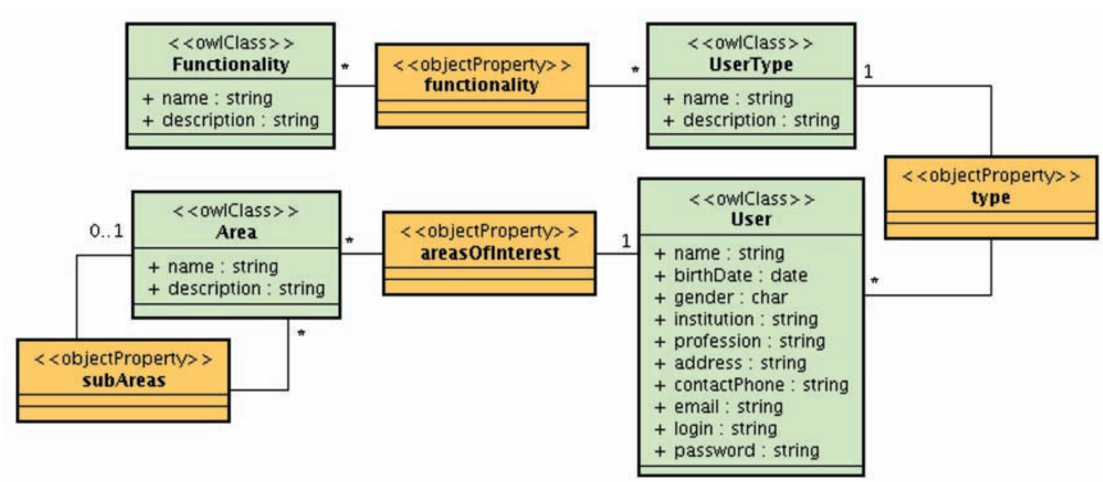


Figure 18. The conceptual model for the User Control module of LabES Portal, in ODM



hierarchy. This could also happen with attributes, for the same reasons. More on ODM's semantics can be found at (OMG, 2007).

In the cases where there is no need to represent associations or attributes as UML classes, S-FrameWeb suggests the conceptual model is simplified, such as the one shown in Figure 19. Notice that this diagram is very similar to the one in Figure 8.

Design

As discussed before, FrameWeb proposes the creation of four kinds of models during design: domain, persistence, navigation and application models. These models are still used with S-FrameWeb, although the domain model (FDM) should be adapted to a representation more suitable to the purposes of this semantic extension. Therefore, S-FrameWeb suggests a new UML profile for this diagram, mixing the profile defined by ODM with the one proposed by FrameWeb.

This new modeling language consists basically of the one defined by ODM, with the following adaptations:

1. Specification of association navigabilities for the implementation of the classes;
2. Addition of the O/R mappings for the configuration of the ORM framework;
3. Use of the data types of the implementation platform instead of those defined by the XML Schema Definition (XSD) standard[21];
4. Simplification of ODM's syntax when possible (if not already done previously).

Naturally, the construction of the FDM should be based on the conceptual model already built in previous activities. Figure 20 shows the FDM for the LabES Portal. We can see that, based on the simplified version of the conceptual model, association navigabilities were defined, data types were chosen among those of the implementation platform and that some O/R mappings were included. The result is very similar of that of Figure 11, due to the simplifications performed.

The representation of this model in a language that mixes profiles from both ODM and FrameWeb attempts to facilitate the implementation phase, when an OWL file representing the conceptual model should be created and the ORM framework should be configured.

Implementation, Testing and Deployment

During implementation, the classes that, integrated with the frameworks, provide a software solution to the problem at hand are developed. S-FrameWeb adds a new task to this activity: the construction of OWL files representing the domain ontology and the application conceptual model (based on the FDM). As stated before, this task is facilitated by the use of ODM in both models.

The OWL files should be used by the Front Controller framework to implement dynamic

Figure 19. The conceptual model for the User Control module of LabES Portal, in its simplified version

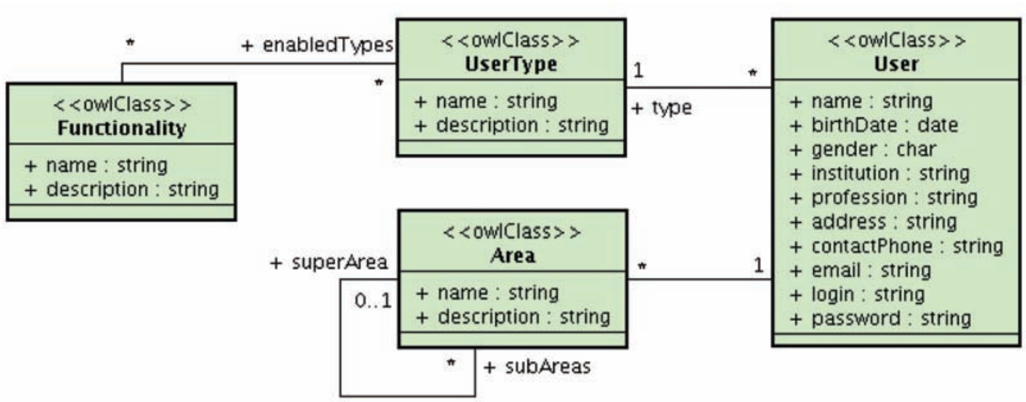
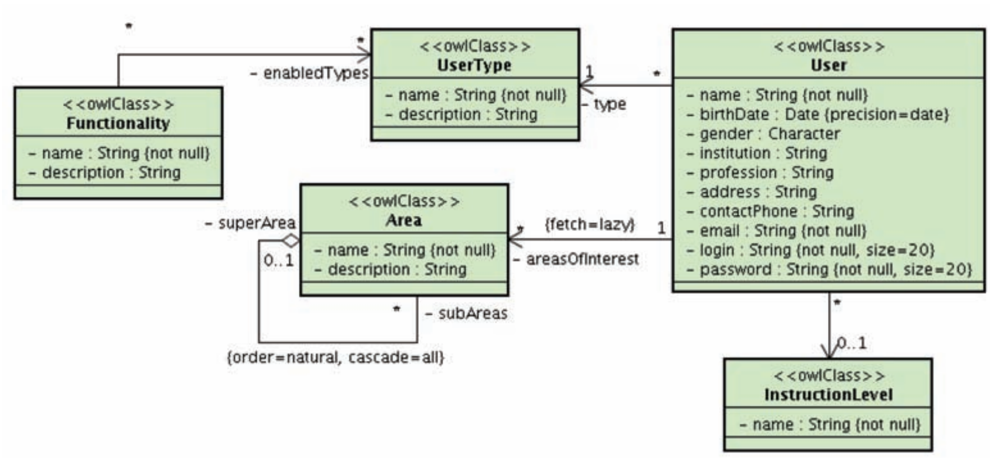


Figure 20. S-FrameWeb's Domain Model for the User Control module of LabES Portal



annotation on the Web pages. S-FrameWeb proposes an extension to this kind of framework that recognizes when a request comes from a human or from a software agent by analyzing a specific HTTP request parameter (e.g. owl=true). In the case of a software agent, the framework should respond with an OWL file that is based on the domain ontology and the conceptual model, and represents the data that would be shown in the human-readable version of the page.

To experiment this approach in practice, a prototype of an extension for the Struts2 framework was built. Figure 21 shows this extension and how it integrates with the framework. The client's web browser issues a request for an action to the framework. Before the action gets executed, the controller automatically dispatches the request through a stack of interceptors, following the pipes and filters architectural style.

An "OWL Interceptor" was developed and placed as first of the stack. When the request is made, this interceptor verifies the HTTP parameter and, if present, creates a pre-result listener that will deviate successful requests to the "OWL Result Class", another custom-made component that is responsible for producing this result.

The listing below is an excerpt of an OWL document produced by the search of publications with "FrameWeb" in their names. Publications that are returned by the applications are placed under the <results> tag, while objects associated with them are placed under <instancesList> tag. The association is made using the UUID of each object. See Figure 21.

Since this result should be based on the application ontology, it was necessary to use an ontology parser. For this purpose, we chose the Jena Ontology API, a framework that provides a programmatic environment for many ontology languages, including OWL. With Jena and Java's reflection mechanisms, the OWL Result Class reads all properties that are made available to the Web page by the action, produces an OWL document containing their information and delivers it to the software agent.

Testing should be conducted in order to check not only the source code, but also the ontologies codified in OWL. In the context of S-FrameWeb, however, this is still open to research and study. Deployment works as the same as other WISs, but should also include the OWL files in a specific place in order to be used by the Front Controller's extension.

Figure 21.


```
<results>
        <instance>
                <uuid>2a6304f5-34c9-4356-a1ce-baa1e7b99e04</uuid>
                <areasOfInterest>130c2f70-5a37-4ad3-815b-841922584cd9</areasOfInterest>
                <areasOfInterest>93fcbf36-cdfe-4fd3-be23-a0d6ab3b45e8</areasOfInterest>
                <publishDate>2007-06-20</publishDate>
                <summary>An Application of the S-FrameWeb Method</summary>
                <participants>e5242491-b2be-4a34-b7b4-b0d9b7537517</participants>
        </instance>
</results>
<instancesList>
        <instance>
                <uuid>130c2f70-5a37-4ad3-815b-841922584cd9</uuid>
                <name>Semantic Web</name>
        </instance>
        <instance>
                <uuid>93fcbf36-cdfe-4fd3-be23-a0d6ab3b45e8</uuid>
                <name>WebApps</name>
        </instance>
        <instance>
                <uuid>e5242491-b2be-4a34-b7b4-b0d9b7537517</uuid>
                <name>Vitor Souza</name>
                <birthDate>1981-06-15</birthDate>
                <gender>M</gender>
                <profession>Professor</profession>
                <institution>UFES</institution>
                <type>2e9c5b6e-0d0f-4da0-b99b-24178ca6873a</type>
        </instance>
</instancesList>
```


FUTURE TRENDS

Web Engineering is a relatively new field of research. New methods, languages and frameworks are proposed to provide practitioners with tools that can facilitate and increase the productivity when developing WebApps.

FrameWeb is a new tool, targeting WISs that have their architecture based on frameworks. By suggesting a standard architecture and bringing concepts from the frameworks to the design models, developers can translate models to code more easily and designer have more control on the outcome of the implementation.

FrameWeb was first applied in the development of the Portal of the Software Engineering Lab – LabES. First, developers were trained in general concepts of Web Engineering, in the use of FrameWeb and also in the following frameworks: WebWork2, FreeMarker (template engine), SiteMesh, Hibernate and Spring.

In general, the development went smoothly. The method allowed the developers to deliver the models mostly in time and few deadlines had to be extended. However, some developers had difficulties on capturing the idea of some frameworks, especially the MVC framework. All of them had some experience with the Java platform, but most did not have any experience with Web development.

At the end of the development, the developers were asked to provide feedback on the work done. This feedback can be summarized in the following items:

- Allowing to directly model aspects related to the use of frameworks is the biggest strength of FrameWeb;
- Implementing in Java what was modeled during design was very much facilitated by the clear understanding of the semantics of the four models (domain, persistence, navigation and application);
- The simplicity of the models facilitated the adoption of FrameWeb, except for the navigation model, which added some complexity to the method.

Two other case studies were conducted. The local Java User Group ESJUG[22] modeled a collaborative learning environment called JSchool[23] using FrameWeb for the same set of frameworks used in the LabES Portal project. This helped mature the method in its initial version.

Another case study reimplemented the LabES Portal changing the Front Controller framework. This helped identify some extensions that should be added to FrameWeb in order to cope with some characteristics of different frameworks. For instance, this work suggested the addition of the <<formBean>> stereotype for the navigation model to represent how the framework Struts sends data from the web page to the action class. It also reached the conclusion that the navigation model in FrameWeb is somewhat dependent on the instance of Front Controller frameworks used, and not generic as it was assumed before.

More case studies should be conducted to assess the effectiveness of the method and its appropriateness to different instances of frameworks. Many improvements can come from more practical experiences.

The use of framework-based architectures is becoming the standard for implementation of medium-to-large-sized WIS. Taking the Java platform as example, the definition of standards as JavaServer Faces (JSF)[24] for Web development and the new Enterprise JavaBeans (version 3.0)[25] for distributed components reinforce that conclusion. JSF defines a MVC-like architecture, and EJB 3.0 had all of its persistence model reconstructed based on Hibernate ORM framework and also makes heavy use of Dependency Injection.

The research on the Semantic Web points out to the future of the World Wide Web. Methods for the development of WISs should prepare for, or even help build, this new paradigm. S-FrameWeb suggests a software process that facilitates the development of Semantic WISs by automating certain tasks concerning the generation of semantic annotations on dynamic Web pages. Nonetheless, FrameWeb and S-FrameWeb are far from ideal: there are several opportunities to improve the method. Future work may include:

- Further research on the impact of the use of frameworks and FrameWeb on the activity of Testing. The current work provides no discussion on the subject of testing;
- Proposals on layout and interaction models. Complete methods for the design of WebApps should include models that model aesthetics and usability;
- Conduction of more formal experiments with the method, evaluating more precisely the gains in the productivity of the development team. Currently, only informal experiments have been conducted and conclusions have been reached by requesting developer's opinions;
- Tools could be developed to help create the models or to convert the models to code, automatically implementing much of the infrastructure code and configuration for the most used frameworks available;
- To make FrameWeb's models more generic, the development of an ontology on Web Applications and frameworks to guide the evolution of FrameWeb's modeling language. New concepts brought by new frameworks could be included in the ontology and, thus, taken to the modeling language;
- Continuation on the research on the Semantic Web and in-practice experiments on the construction of a Semantic WIS using S-FrameWeb;
- Deeper discussions on how to tackle specific Semantic Web issues such as: how will agents find the desired web page?, how will they know how to interact with it?, how will they know if a concept "table" refers to a piece of furniture or a systematic arrangement of data usually in rows and columns?, will a top-level ontology be used for all the Internet?

- Evaluation on how to use Semantic Web Services with S-FrameWeb instead of the dynamic page approach and a comparison of both solutions.

CONCLUSION

The amount of propositions in the Web Engineering area, including methods, frameworks and modeling languages, is quite vast, demonstrating that academics and practitioners haven't yet elected a standard when it comes to Web development.

Parallel to this, many frameworks and containers for the implementation of WISs were created, denoting the need for a basic infra-structure that helps on the quick development of reliable software with low future maintenance costs. With several ready-to-use and extensively tested components, frameworks promote reuse and good programming practices.

The large utilization of these frameworks and containers by practitioners and the absence of a design method directed to them has motivated the proposal of FrameWeb, a method based on frameworks for the design of WISs. The current research on the Semantic Web, with many efforts on bringing this idea to reality has impelled us to extend this method and create S-FrameWeb: a method based in frameworks to the construction of semantic WISs.

Given all of the options available, FrameWeb comes in as another one that targets a specific architecture, one based on the use of frameworks. In this case, FrameWeb excels for its agility, because models are directed towards the framework architectures and allow for quick understanding of the implementation. It also doesn't introduce much complexity, allowing organizations to use their own processes up to design with few adaptations, if any. Of all the proposed design models, the navigation model is the only one we consider a little bit complex, making FrameWeb very easy to learn and use.

S-FrameWeb complements FrameWeb, adding activities that promote the construction of Semantic WISs. Given that the Semantic Web vision will not come true unless Web authors add semantic to their websites, S-FrameWeb is a step in that direction, giving directives for WISs developers to follow in order to add Semantic to Web Applications.

REFERENCES

W3C (2007a). *OWL Web Ontology Language Guide, fev. 2004*.

W3C (2007b). *W3C Glossary and Dictionary*.

Ahmad, R., Li, Z., & Azam, F. (2005). Web Engineering: A New Emerging Discipline. In *Proceedings of the IEEE Symposium on Emerging Technologies* (pp.445-450). Catania, Italy: IEEE.

Alur, D., Malks, D., & Crupi, J. (2003). *Core J2EE Patterns: Best Practices and Design Strategies, 2nd edition*. Prentice Hall.

Ambler, S., & Jeffries, R. (2002). *Agile Modeling: Effective Practices for Extreme Programming and the Unified Process, 1st edition*. John Wiley & Sons.

Arch-Int, S., & Batanov, D. N. (2003). Development of industrial information systems on the Web using business components. [Elsevier.]. *Computers in Industry*, *50*(2), 231–250. doi:10.1016/S0166-3615(02)00122-7

Bauer, C., & King, G. (2004). *Hibernate in Action, 1st edition*. Manning.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*, 34–43.

Booch, G., Rumbaugh, J., & Jacobson, I. (2005). *The Unified Modeling Language User Guide. 2nd edition*. Addison-Wesley Professional.

Ceri, S., Fraternali, P., & Bongio, A. (2000). Web Modeling Language (WebML): a modeling language for designing Web sites. *Computer Networks*, 33 (1-6), 137-157, Elsevier.

Conallen, J. (2002). *Building Web Applications with UML, 2nd edition*. Addison-Wesley.

Díaz, P., Montero, S., & Aedo, I. (2005). Modelling hypermedia and web applications: the Ariadne Development Method. [Elsevier.]. *Information Systems*, *30*(8), 649–673. doi:10.1016/j.is.2004.09.001

Đurić, D. (2004). MDA-based Ontology Infrastructure. *Computer Science and Information Systems* 1 (1), ComSIS Consortium.

Falbo, R. A. (2004) Experiences in Using a Method for Building Domain Ontologies. In *Proceedings of the Sixteenth International Conference on Software Engineering and Knowledge Engineering* (pp. 474-477), Banff, Alberta, Canadá.

Falbo, R. A., Guizzardi, G., & Duarte, K. C. (2002). An Ontological Approach to Domain Engineering. In *Proceedings of the 14th International Conference on Software Engineering and Knowledge Engineering* (pp. 351- 358). Ischia, Italy: Springer.

Fons, J., Valderas, P., Ruiz, M., Rojas, G., & Pastor, O. (2003). OOWS: A Method to Develop Web Applications from Web-Oriented Conceptual Models. In *Proceedings of the International Workshop on Web Oriented Software Technology* (pp. 65-70), Oviedo, Spain.

Fowler, M. (2002). *Patterns of Enterprise Application Architecture*. Addison-Wesley.

Fowler, M. (2007). *Inversion of Control Containers and the Dependency Injection pattern*.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley.

Ginige, A., & Murugesan, S. (2001). Web Engineering: An Introduction. [IEEE.]. *IEEE MultiMedia*, *8*(1), 14–18. doi:10.1109/93.923949

Gomez-Perez, A., Corcho, O., & Fernandez-Lopez, M. (2005). *Ontological Engineering*. Springer.

Guarino, N. (1998). Formal Ontology and Information Systems. In *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems* (pp. 3-15), Trento, Italy: IOS Press.

Jacobson, I., Booch, G., & Rumbaugh, J. (1999). *The Unified Software Development Process*, Addison Wesley.

Jacyntho, M. D., Schwabe, D., & Rossi, G. (2002). A Software Architecture for Structuring Complex Web Applications. In *Proceedings of the 11th International World Wide Web Conference, Web Engineering Alternate Track*, Honolulu, EUA: ACM Press.

Koch, N., Baumeister, H., Hennicker, R., & Mandel, L. (2000). Extending UML to Model Navigation and Presentation in Web Applications. In *Proceedings of Modelling Web Applications in the UML Workshop*, York, UK.

Koch, N., & Kraus, A. (2002). The Expressive Power of UML-based Web Engineering. In D. Schwabe, O. Pastor, G. Rossi e L. Olsina (Ed.), *Proceedings of the Second International Workshop on Web-Oriented Software Technology* (pp. 105-119), CYTED.

Krutchen, P. (2000). *The Rational Unified Process: An Introduction, 2nd edition*. Addison-Wesley.

Lee, S. C., & Shirani, A. I. (2004). *A component based methodology for Web application development. Journal of Systems and Software*, 71 (1-2), 177-187, Elsevier. Lima, F., & Schwabe, D. (2003). Application Modeling for the Semantic Web. In Proceedings of the 1st Latin American Web Conference (pp. 93-102), Santiago, Chile: IEEE-CS Press.

McIlraith, S. A., Son, T. C., & Zeng, H. (2001). Semantic Web Services. [IEEE.]. *Intelligent Systems*, *16*(2), 46–53. doi:10.1109/5254.920599

Murugesan, S., Deshpande, Y., Hansen, S., & Ginige, A. (1999). Web Engineering: A New Discipline for Development of Web-based Systems. In *Proceedings of the 1st ICSE Workshop on Web Engineering* (pp. 3-13). Australia: Springer.

Narayanan, S., & McIlraith, S. A. (2002). Simulation, Verification and Automated Composition of Web Services. In *Proceedings of the 11th international conference on World Wide Web* (pp. 77-88), Hawaii, USA: ACM.

Neighbors, J. M. (1981). *Software Construction Using Components*. Ph.D. Thesis. Department of Information and Computer Science, University of California, Irvine.

OMG. (2007). *Ontology Definition Metamodel Specification*.

Pastor, O., Gómez, J., Insfrán, E., & Pelechano, V. (2001). The OO-Method Approach for Information Systems Modelling: From Objetct-Oriented Conceptual Modelling to Automated Programming. [Elsevier.]. *Information Systems*, *26*(7), 507–534. doi:10.1016/S0306-4379(01)00035-7

Pressman, R. S. (2005). *Software Engineering: A Practitioner's Approach, 6th edition*. McGraw Hill.

Reenskaug, T. (1979). *THING-MODEL-VIEW-EDITOR, an Example from a planning system*. Xerox PARC Technical Note.

Resende, A., & Silva, C. (2005). *Programação Orientada a Aspectos em Java*. Brasport.

Rosenberg, D., & Scott, K. (1999). *Use Case Driven Object Modeling with UML: A Practical Approach*. Addison-Wesley.

Schmidt, D. (2007). "*Programming Principles in Java: Architectures and Interfaces", chapter 9*.

Schwabe, D., & Rossi, G. (1998). An Object Oriented Approach to Web-Based Application Design. *Theory and Practice of Object Systems* 4 (4), Wiley and Sons.

Shannon, B. (2003). *JavaTM 2 Platform Enterprise Edition Specification, v1.4*. Sun Microsystems.

Souza, V. E. S., & Falbo, R. A. (2007). FrameWeb - A Framework-based Design Method for Web Engineering. In [Faro, Portugal: ACM Press.]. *Proceedings of the Euro American Conference on Telematics and Information Systems, 2007*, 17–24.

Souza, V. E. S., Lourenço, T. W., Falbo, R. A., & Guizzardi, G. (2007). S-FrameWeb – a Framework-based Design Method for Web Engineering with Semantic Web Support. In *Proceedings of Workshops and Doctoral Consortium of the 19th International Conference on Advanced Information Systems Engineering* (pp. 767-778), Trondheim, Norway.

Standing, C. (2002). Methodologies for developing Web applications. [Elsevier.]. *Information and Software Technology*, *44*(3), 151–159. doi:10.1016/S0950-5849(02)00002-2

ENDNOTES

1 http://struts.apache.org/2.x/index.html
2 http://www.springframework.org
3 http://tapestry.apache.org
4 http://www.opensymphony.com/sitemesh
5 http://struts.apache.org/struts-tiles
6 http://www.hibernate.org
7 http://java.sun.com/products/jdo
8 http://db.apache.org/ojb/
9 http://www.oracle.com/technology/products/ias/toplink
10 http://www.picocontainer.org
11 http://jakarta.apache.org/hivemind
12 http://www.eclipse.org/aspectj

[13] http://labs.jboss.com/portal/jbossaop

[14] http://www.acegisecurity.org

[15] http://cocoon.apache.org

[16] http://java.sun.com/products/jaas

[17] http://oiled.man.ac.uk/

[18] http://protege.stanford.edu/

[19] The relationship between an object's identity in memory and its primary key in the database raises several issues that are discussed in the article "Hibernate, null unsaved value and hashcode: A story of pain and suffering" from Jason Carreira (http://www.jroller.com/page/jcarreira?entry=hibernate_null_unsaved_value_and). The idea of using a Universal Unique Identifier (UUID) was taken from this article.

[20] Object Management Group – http://www.omg.org/ontology/

[21] The XML Schema standard can be found at http://www.w3.org/XML/Schema. Its data types are described in a specific page, at http://www.w3.org/TR/xmlschema-2.

[22] http://esjug.dev.java.net

[23] http://jschool.dev.java.net

[24] http://jcp.org/en/jsr/detail?id=127

[25] http://jcp.org/en/jsr/detail?id=220

This work was previously published in Innovations in Information Systems Modeling: Methods and Best Practices, edited by T. Halpin, J. Krogstie, and E. Proper, pp. 204-238, copyright 2009 by Medical Information Science Reference (an imprint of IGI Global).

Chapter 2.5

Focused Requirements Engineering Method for Web Application Development

Ala M. Abu-Samaha
Amman University, Jordan

Lana S. Al-Salem
SpecTec Ltd & MEP, Greece

INTRODUCTION

The requirements phase of the system/application development process typically involves the activities of requirements elicitation, analysis, validation, and specification. The main goal of such a process is "to develop a requirements specification document which defines the system to be procured and which can act as a basis for the system design" (Sawyer, Sommerville, & Viller, 1996). Hence the underpinning assumption of the requirements engineering (RE) process is to transform the operational needs of an organisation into complete, consistent, and unambiguous system/application specifications through an iterative process of definition and validation (Pohl, 1994).

DOI: 10.4018/978-1-60566-026-4.ch244

The Web engineering (WE) literature provides a limited number of methods and techniques that can be used to manage the RE process in a Web development context [*e*^{3}*-value* framework (Gordijn, Akkermans, & van Vliet, 2000), SOARE approach (Bleistein, Aurum, Cox, & Ray, 2004), e-prototyping (Bleek, Jeenicke, & Klischewski, 2002), AWARE (Bolchini & Paolini, 2004), and SSM/ICDT (Meldrum & Rose, 2004)]. Despite the availability of such a limited number of Web requirements engineering (WRE) methods, many researchers criticised such methods for their failure to address the necessity to align the Web application' requirements to the organisation's business strategy. Hence, the recommendation of many researchers (Al-Salem & Abu-Samaha, 2005a; Bleistein 2005; Bleistein, Cox, & Verner, 2004; Vidgen, Avison, Wood, & Wood-Harper, 2002) is to utilise a general WRE framework for the development of Web applications that can align the application's requirements to the organisation's business needs and its future vision. The objective of such a WRE framework is to incorporate the elicitation/analysis of business strategy as part of the application's RE process.

This chapter presents a WRE method that extends Sommerville and Kotonya's viewpoint-oriented requirements definition (VORD) and Kaplan and Norton's balanced scorecard (BSC) to elicit the Web application' requirements and to plan/analyze the business strategy, respectively. In addition, eWARE (extended Web application requirements engineering) deploys the concept of "requirements alignment" to attain business objectives during the requirements discovery, elicitation, and formalisation process to identify the services of the Web application that will achieve the business objectives in order to improve the organisation's profitability and competitiveness. The chapter is organised into a number of sections. The second section of this chapter provides a background to Web applications in terms of definition and differentiating characteristics. The third section provides a discussion of eWARE method in terms of phases and activities. This section is divided into two subsections to cover the activities of the two prominent phases of the eWARE process in more detail. The fourth and fifth sections provide a discussion of possible future trends in WRE and a number of concluding remarks.

BACKGROUND

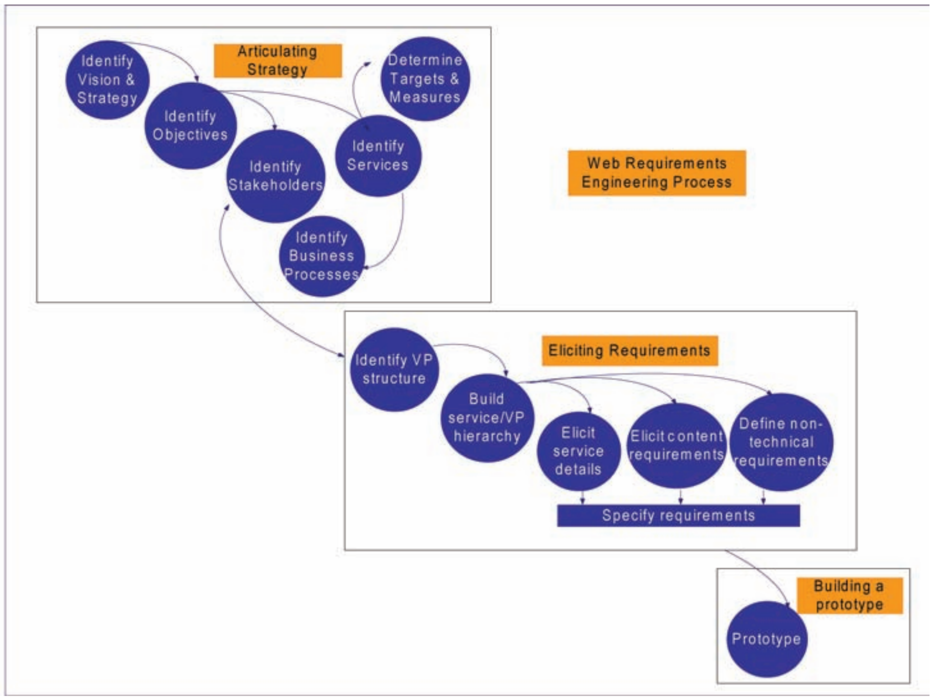
Web applications provide organisations an unprecedented chance to stretch their existence beyond the typical boundaries of an organisation to include customers, trading partners, and suppliers.Little attention has been paid to the process of RE for Web application development, in comparison to other areas of the development process [modelling, design, and coding] (Ginige & Murugesan, 2001). Hence, Web applications can be defined as "applications that tend to be used to integrate and streamline an organisation's business processes beyond organisational (customers, agents, suppliers, others) and geographical borders; provide an organisation with competitive products and services that give it a strategic advantage over its competitors in the marketplace; promote business innovation; and/or improve operational efficiency" (Al-Salem & Abu-Samaha, 2005a).

There is a pressing need in the WE discipline for RE approaches and techniques that (a) take into account the multiplicity of user profiles and the various stakeholders involved [a stakeholder is defined as "anyone who can share information about the system, its implementation constraints or the problem domain" (Potts, Takahashi, & Anton, 1994)]; (b) eliciting overall functionality and the business environment of the Web application; (c) specifying technical and nontechnical requirements of the Web application, and (d) aligning the Web application' requirements to the overall business strategy (Bleistein et al., 2005; Ceri, Fraternali, Bongio, Brambilla, Comai, & Matera, 2003; Ginige & Murugesan, 2001; Kautz & Madsen, 2003; Lowe, 2003; Meldrum & Rose, 2004; Nuseibeh & Easterbrook, 2000; Vidgen et al., 2002). More importantly, a Web application must be developed with an emphasis on how the services of such an application can achieve the business vision and strategy and fulfil the business processes (Haire, Henderson-Sellers, & Lowe, 2001).

eWARE Process

eWARE process can be best perceived as a series of activities grouped into three phases; strategy articulation via BSC, Web application' requirements elicitation via VORD, and prototype building; Figure (1) presents the phases and activities of eWARE. Such a process aims to develop a Web application requirements specification (WRS) document that is aligned with business strategy

Figure 1. The eWARE process



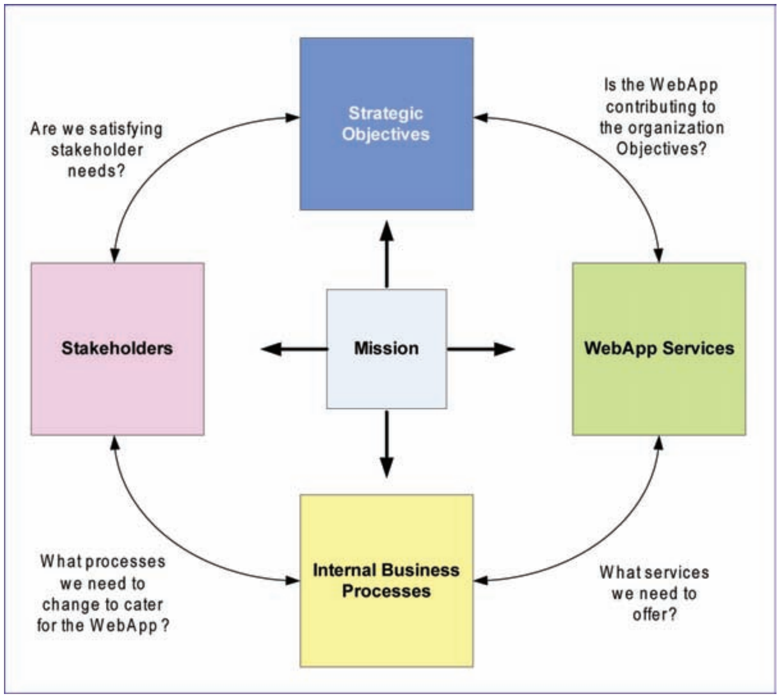
and detailed enough to be used for contractual purposes.

The strategy articulation phase of eWARE can be best thought of as a structure of many layers. The vision of the organisation is at the top of the structure, while the strategic objectives are presented in the next layer of the structure followed by the Web application services. The next layer of the structure contains the measurements and targets for meeting the strategic objectives. The level of detail tends to increase as we move down the structure of the strategy. This articulation of the organisation's vision, objectives, and measurements aims to translate the future vision of the organisation into detailed and prioritised Web application requirements (this will be fully covered in the coming subsections).

The requirements elicitation phase of the eWARE process is used to produce a WRS document based on requirements collected during the strategy articulation phase. Requirements elicitation relies on the identification of the relevant viewpoints (VPs), their sub-VPs, and requirements for each viewpoint (VP). Kotonya and Sommerville (1996) define a VP as anyone who may have some direct or indirect influence on the system/application requirements. Goals and objectives of the different stakeholder groups need to be identified to define success or failure measures for each stakeholder. Moreover, nonfunctional requirements {NFR) need to be identified; these include some of the "-ilities" of the Web application, such as reliability, supportability, maintainability, affordability, and so forth. Finally, in the prototyping phase of the eWARE process, the system/application stakeholders consider the unclear set of requirements, and agree on prototyping the ambiguous requirements to verify them. Moreover, the user interface (UI) and Web site structure are presented in a throw away prototype.

The mentioned WRE phases and activities are perceived to be iterative and incremental in nature where unmet targets are questioned. This cyclic view of the WRE process will trigger the strategy

Figure 2. eBSC perspectives



articulation phase to enter in a feedback loop in order to refine services and change requirements of the Web application in order to enhance the organisation's chance to achieve its set vision and strategy.

Strategy Articulation Phase

As mentioned earlier, the strategy articulation phase of eWARE process aims to align the Web application' requirements to the organisation's business strategy through a process of strategy analysis. eWARE delivers such alignment via eBSC (extended Balanced Score Card). According to eBSC, aligning the Web application' requirements to the business objectives yields four perspectives to focus upon (stakeholders, strategic objectives, internal processes, and Web application services). Figure 2 provides a diagrammatic representation of the perceived relationship between the four perspectives.

Strategic Objectives Perspective

Objectives are statements that clarify what the strategy aims to achieve. Hence, organisations pursue strategies in the belief that, when implemented, they will enable the organisation to better achieve its strategic objectives.

Web Application Services Perspective

The Web application services are the collection of functionality, quality, content, and all what the Web application must provide in order to achieve the strategic objectives of the organisation and to meet the expectations of its stakeholders. This perspective is considered as a precondition to process improvement, stakeholder satisfaction, and business objectives' realisation. Hence, the development team, comprising of requirement engineers and end user representatives, must be empowered to select which services (functionalities) of the Web application will be included, in order to deliver the optimal business solution. Decisions and trade-offs need to be made between new services elicited through the eBSC process, and less important original functionalities that may have to be excluded, being less strategically important.

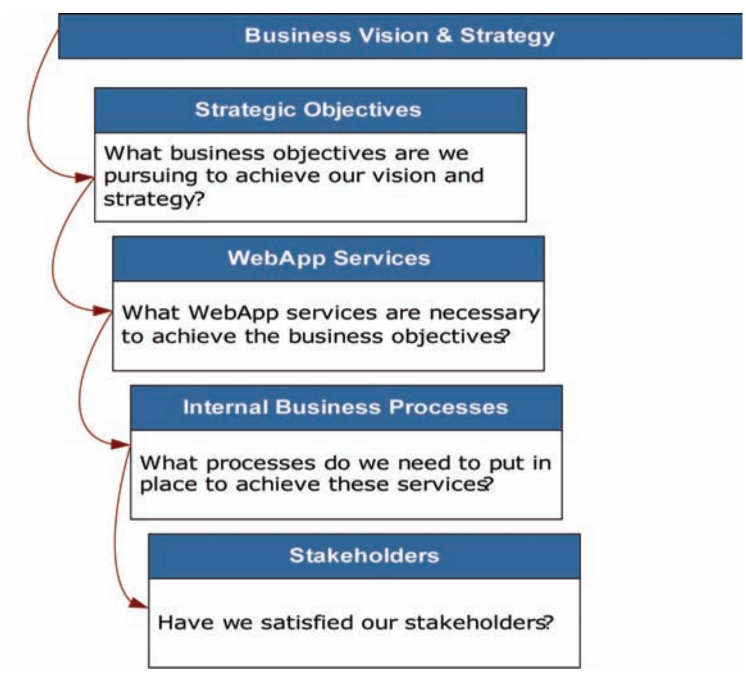
Stakeholders[1] Perspective

The stakeholder perspective is arguably the most important perspective in the Web application development process. As mentioned earlier, a stakeholder is defined by Potts et al. (1994) as "anyone who can share information about the system, its implementation constraints or the problem domain." According to Potts et al. (1994), the list of stakeholders will include end users, indirect users, other customer representatives, and developers. Since a stakeholder is, in essence, a requirements source; it can be an application user, a competitor, or even a third party. Hence, if the stakeholder represents a customer, then the requirements elicitation effort will focus on new customer acquisition, customer retention, and customer profitability. The goal of using eBSC is to classify the organisation's strategy by stakeholder, which will lead to introducing Web application services for each group to meet their strategy or objective. The organisation must determine whom it serves and how their requirements can be met.

Internal Process Perspective

Organisations' activities can be grouped into "business processes" that describe the way work is to be implemented. Some of the organisation's activities will be affected by the introduction of a Web application, since such applications have the potential to significantly change an organisation's work practices and procedures (Ginige & Murugesan, 2001; Pressman, 2004). The internal process perspective focuses on key processes at which the organisation must excel in order to add value to its stakeholders through the Web application.

Figure 3. Web application cause and effect relationships



When designing a score card, the starting point will be asking "what strategies do we have to put in place to satisfy the wants and needs of the key stakeholders?" The inclusion of BSC within the WRE process aims to help clarify, consolidate, and gain consensus around the business–Web application strategy of the organisation, and to translate the business strategy into Web application services (requirements) to ensure that the elicited requirements are strategy focused. Hence, a strategy can be best described as a series of cause and effect relationships that provide a translation from future vision to Web requirements, as shown in Figure (3).

eVORD for Eliciting Web Requirements

The second phase of the proposed WRE process presents the requirements elicitation phase of the Web application development process. eWARE delivers such elicitation via eVORD (extended viewpoint-oriented requirements definition). According to eVORD, templates are created to describe each viewpoint (VP), service, nonfunctional requirement (NFR), and content. As mentioned earlier, Kotonya and Sommerville (1996) define a VP as anyone who may have some direct or indirect influence on the system/application requirements. The VORD requirements engineering process includes activities concerned with VP identification, VP service description, cross-viewpoint analysis to discover inconsistencies, omissions, and conflicts, and developing an object-oriented model of the system/application from the VP analysis. In addition, a VP diagram is used to show the relationships among VPs, while sequence diagrams illustrate the interactions among VPs.

Web Application Viewpoints (WebVPs) Identification and Structuring

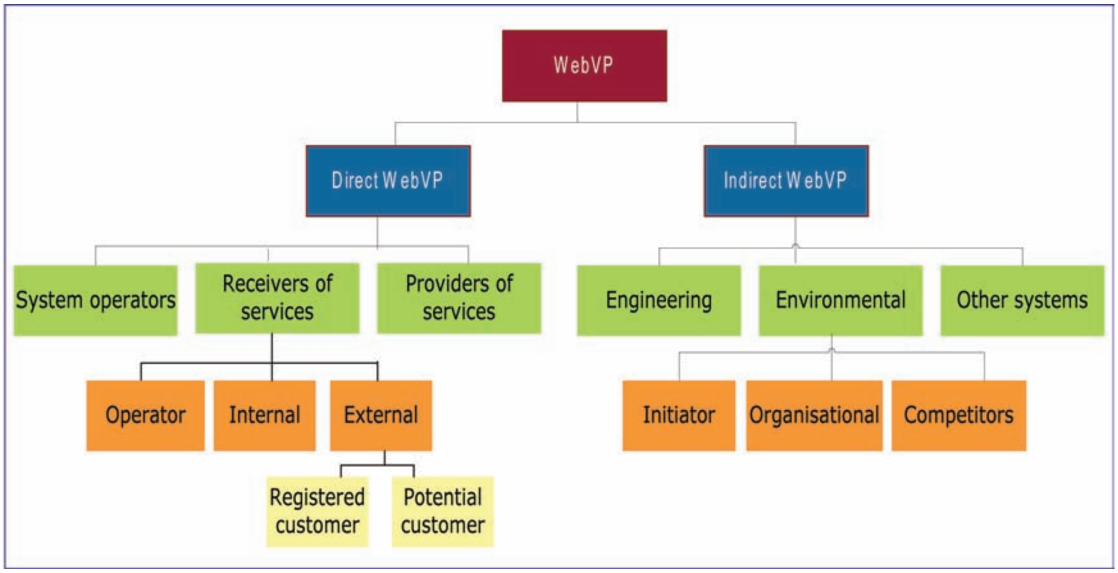
The construction of Web applications involves a great number of stakeholders who have different views of the Web application. These perspectives are partial or incomplete descriptions of the system/application, and reflect the environment in which the system/application will operate in. The integration of multiple views can contribute to augment the overall understanding of the Web application. The authors recommend the use of a new and amended list of VPs for developing Web applications. These WebVPs' abstracts have been identified as a starting point that acts as a template for WebVPs classes and hierarchy. Figure (4) shows a high-level abstract of WebVPs structure.

WebVPs can be classified into *direct VPs* that interact directly with the Web application and fall into two subclasses (receivers and providers of services), and *indirect VPs* that have "interest" in some or all of the services that are delivered by the system but do not interact directly with it; hence, they provide high-level organisational requirements and constraints. Indirect VPs fall into a number of subclasses: *environmental WebVPs*, which reflect the requirements of the business domain, that is, legalisation, localisation, taxation, and competitors; *engineering WebVPs*, which reflect the requirements of the development team, that is, software engineers, team leaders, and creative designers; and *system WebVPs*, which include all existing information systems that the application being analysed needs to interface to, that is, payment systems and supplier systems.

Eliciting Requirements Details

The second step of this phase is concerned with documenting the details of each WebVP (identified in the previous step) and its associated services. For every identified WebVP, a number of templates are used to elicit and specify its requirements. The authors have extended and adapted VORD templates to cater to the particularities of Web applications development (please refer to Al-Salem & Abu-Samaha, 2005a and Al-Salem & Abu-Samaha, 2005b for more details).

Figure 4. Abstract of WebVPs structure



Documenting Requirements

The objective of the requirements process is to deliver a requirements specification document that defines the system/application to be developed (Sawyer, Viller, & Sommerville, 1997). This document is used for contractual purposes, and can be used as a basis for facilitating a competitive tendering for the system/application design and implementation (IEEE, 2004). The authors have enhanced the typical software requirements specification (SRS) document to reflect the changes introduced to eVORD and eBSC, as depicted in Figure (5).

FUTURE WORK

The domain of RE, in general, and WRE, in particular, is evolving with the ever-changing contexts of software engineering and information systems development projects. Despite the availability of many requirements, engineering methods, processes, techniques, and tools, such artefacts are in need of constant extensions and enhancements to bring such artefacts to the changing contexts of the development projects. The presented eWARE method is an extension of two existent methods used to align requirements of a Web application to the organisation's business strategy. eWARE needs to be tested further in different industrial settings to validate the applicability of the method. As well, eWARE, like many other RE methods, needs to be supported through the development of a computer aided software engineering (CASE) tool to facilitate the generation of the WRS document. Compared to similar WRE methods (Bleek et al., 2002; Bleistein et al., 2004; Bolchini & Paolini, 2004; Gordijn et al., 2000; Meldrum & Rose, 2004), eWARE provides its users with a number of benefits: it extends the BSC approach to help with the formulation of Web application strategy as a stage in the WRE process in advance to requirements elicitation; it provides a prioritisation framework synchronised with business strategy; it extends VORD to elicit requirements for Web applications; and it creates a WRS document that specifies both the business strategy and requirements for the Web application. Such advantages proved to be valuable in Web applica-

Figure 5. Software requirements specification (SRS) document template

Table of Contents

0.1 Document History.....4
0.2 Changes From Last Issue.....4
0.3 Acknowledgements.....4
0.4 Distribution List.....4
0.5 Referenced Documents.....4
0.6 Abbreviations.....4
0.7 Glossary.....4

1 INTRODUCTION.....5
1.1 Purpose.....5
1.2 Documentation Conventions.....5
1.3 Scope.....5
1.4 Overview.....5

2 STRATEGY ARTICULATION.....6
2.1 Business Vision.....6
2.2 Strategic Objectives.....6
2.3 WebApp Services.....6
2.4 Business Processes.....6
2.5 Stakeholders.....6
2.6 Stakeholders' Services.....6
2.7 Measurements and Targets.....6

3 WEBVPS AND SERVICES.....7
3.1 WebVps Identification.....7
3.2 WebVp Services' Hierarchy.....7

4 WEBVP REQUIREMENTS.....8
4.1 WebVp 1.....8
4.1.1 Service 1.....8
4.1.2 Service 2.....8
4.1.3 Content 1.....8
4.1.4 Content 2.....8
4.2 WebVp 2.....8
4.2.1 Service 1.....8
4.2.2 Service 2.....8
4.2.3 Content.....8

5 HIGH-LEVEL ARCHITECTURE.....9
5.1 Component 1.....9
5.2 Component 2.....9
5.3 Component 3.....9

tion development projects, as Web applications are different in many aspects when compared to other applications. The differing characteristics of Web applications can be summarized as diverse and volatile requirements, vast and unknown end users, multiple stakeholders, adaptable architecture, short development life cycle, high visibility, heavy content, integration with backend databases and third party applications, Web applications' relevance and direct effect on business, and multidisciplinary development team.

CONCLUSIONS

There are two general opinions on whether current software engineering methods and techniques are applicable to face the challenges of developing a Web application. One opinion advocates the need for a new "software engineering" discipline that handles the Web application particularities (Ginige & Murugesan, 2001, Murugesan, 1999; Murugesan, Deshpande, Hansen, & Ginige, 1999). In contrast, the other opinion believes that current software engineering methods, tools, and techniques are applicable to Web application development. For example, McDonald and Welland (2001) recommended developing new methods, techniques, and approaches to address the challenges of developing Web applications in order to increase the possibility of their success. This implies the need for a new breed of WRE methods, tools,and working practices. In contrast, Pressman (2004) argues that Web applications are a natural evolution of existing applications/systems offering a solution to classical problems exhibited by previous information systems (IS). The authors of this chapter hold a middle position in between these two opinions. The authors perceive Web applications as a natural evolution of "traditional" information systems, yet they believe that such applications possess special characteristics that need to be provided for by the "traditional" RE method(s). Hence, existing RE methods are considered valid for WRE, though they need to be enhanced and extended to cater to the distinguishing features of Web applications.

The chapter has presented a WRE approach that enables businesses to develop/procure Web application(s) capable to achieve the organisations' business strategic objectives, and to effectively harness their business processes. The combination of BSC (Kaplan & Norton 1993, 1996a & b) and VORD (Kotonya & Sommerville 1996) within eWARE process provides the development team with the ability to translate strategy into Web application requirements and to incorporate views from different complementary perspectives. Hence, eWARE is a Web-specific RE process to cope with the complex aspects of requirements elicitation, alignment, and specification for Web applications development/procurement. eWARE views Web applications as organisational initiatives and as such, it takes into account the need to address strategic objectives, business processes issues, requirements details of services, NFR, and integration with other systems.

REFERENCES

Al-Salem, L. S., & Abu-Samaha, A. (2005a). *Assessing the usability of VORD for Web applications requirements engineering - An industrial case study* -. 1st International Workshop on Requirements Engineering for Business Need and IT Alignment (REBNITA 2005), Paris.

Al-Salem, L. S., & Abu-Samaha, A. (2005b). *Assessing the usability of VORD method for Web applications requirements elicitation*. International Conference on Internet Technologies and Applications (ITA 05), Wrexham, North Wales, UK.

Bleek, W.-G., Jeenicke, M., & Klischewski, R. (2002). *Developing Web-based applications through e-prototyping*. 26th Annual International Computer Software and Applications Conference (COMPSAC 2002), Oxford England.

Bleistein, S. J. (2005). *Requirements analysis framework for alignment of IT with competitive strategy of business organizations*. IEEE International Conference on Requirements Engineering - Doctoral Consortium, Paris.

Bleistein, S. J., Aurum, A., Cox, K., & Ray, P. (2004). Strategy-oriented alignment in requirements engineering: Linking business strategy to tequirements of e-business systems using the SOARE approach. *Journal of Research and Practice in Information Technology*, *36*, 259–276.

Bleistein, S. J., Cox, K., & Verner, J. (2005). Validating strategic alignment of organizational IT requirements using goal modelling and problem diagrams. *Journal of Software and Systems*, *79*, 362–378. doi:10.1016/j.jss.2005.04.033

Bolchini, D., & Paolini, P. (2004). Goal-driven requirements analysis for hypermedia-intensive Web applications. *Requirements Engineering Journal Special Issue.*

Ceri, S. Fraternali, P., Bongio, A., Brambilla M., Comai S., & Matera M. (2003). *Designing data-intensive Web applications*. Morgan Kaufman.

Ginige, A. (2002). Web engineering: Managing the complexity of Web systems development. In *14th International Conference on Software Engineering and Knowledge Engineering (SEKE '02),* Ischia, Italy: ACM Press.

Ginige, A., & Murugesan, S. (2001). The essence of Web engineering: Managing the diversity and complexity of Web application development. *IEEE MultiMedia*, *8*(2), 22–25. doi:10.1109/MMUL.2001.917968

Gordijn, J., Akkermans, H., & van Vliet, H. (2000). Value based requirements creation for electronic commerce applications. In *33rd Hawaii International Conference on System Sciences*, IEEE, Hawaii, USA.

Haire, B., Henderson-Sellers, B., & Lowe, D. (2001). Supporting Web development in the open process: Additional tasks. *COMPSAC'2001: International Computer Software and Applications Conference,* Chicago, Illinois, USA, IEEE Computer Society.

IEEE. (2004). *Guide to the software engineering body of knowledge*. IEEE.

Kaplan, R., & Norton, D. (1993). Putting the balanced scorecard to work. *Harvard Business Review*, (September/October): 134–147.

Kaplan, R., & Norton, D. (1996a). *The balanced scorecard: Translating strategy into action*. Boston: Harvard Business School Press.

Kaplan, R., & Norton, D. (1996b). Using the balanced scorecard as a strategic management system. *Harvard Business Review*, (January/February): 75–85.

Kautz, K., & Madsen, S. (2003). *Web development - The differences, similarities and in-betweens.* Melbourne, Australia: Information Systems Development Conference.

Kotonya, G., & Sommerville, I. (1996). Requirements engineering with viewpoints. *BCS/IEEE . Software Engineering Journal*, *11*(1), 5–18.

Lowe, D. (2003). Web system requirements: An overview. *Requirements Engineering Journal*, *8*, 102–113. doi:10.1007/s00766-002-0153-x

McDonald, A., & Welland, R. (2001). *Agile Web engineering (AWE) process.* Scotland: Department of Computing Science, University of Glasgow.

Meldrum, M., & Rose, J. (2004). *Activity based generation of requirements For Web-based information systems: The SSM/ICDT approach*. The 12th European Conference on Information Systems (ECIS 2004), Turku Finland.

Murugesan, S. (1999). Web engineering. *ACM SIGWEB Newsletter*, *8*(3), 28–32. doi:10.1145/951440.951445

Murugesan, S., Deshpande, Y., Hansen, S., & Ginige, A. (1999). Web engineering: A new discipline for development of Web-based systems. In *Proceeding of the Int'l Conf. Software Engineering (ICSE) Workshop on Web Engineering* (pp: 1-9). Sydney, Australia: IEEE Multimedia.

Nuseibeh, B., & Easterbrook, S. (2000). Requirements engineering: A roadmap. In *International Conference on Software Engineering (ICSE-2000),* Limerick, Ireland, ACM Press.

Pohl, K. (1994). The three dimensions of requirements engineering. *Information Systems*, *19*(3), 243–258. doi:10.1016/0306-4379(94)90044-2

Potts, C., Takahashi, K., & Anton, A. (1994). Inquiry-based requirements analysis. *IEEE Software*, *11*(2), 21–40. doi:10.1109/52.268952

Pressman, R. S. (2004). *Software engineering: A practitioner's approach*. McGraw-Hill.

Sawyer, P., Sommerville, I., & Viller, S. (1996). *PREview: Tackling the real concerns of requirements engineering*. Lancaster: Computing Department, Lancaster University.

Sawyer, P., Viller, S., & Sommerville, I. (1997). Requirements process improvement through the phased introduction of good practice. *Software Process Improvement and Practice*, *1*, 19–34. doi:10.1002/(SICI)1099-1670(199703)3:1<19::AID-SPIP66>3.0.CO;2-X

Sommerville, I. (1995). *Software engineering*. Addison Wesley.

Sommerville, I., & Sawyer, P. (1997). *Requirements engineering: A good practice guide*. John Wiley & Sons Lts.

Vidgen, R., Avison, D., Wood, J. R. G., & Wood-Harper, A. T. (2002). *Developing Web information systems*. Butterworth Heinemann.

KEY TERMS

Balanced Score Card (BSC): "A multidimensional framework for describing, implementing and managing strategy at all levels of an enterprise by linking objectives, initiatives and measures to an organisation's strategy" (Kaplan & Norton, 1993, 1996 a & b)

eWARE (extended Web application requirements engineering): "A strategy-focused requirements engineering method used to align Web application requirements to business strategy and to elicit legal, technological, business, marketing and content requirements"

A Requirement: "A condition or capability that must be met or fulfilled by a system to satisfy a contract, standard, specification, or other formally imposed documents." (IEEE Standard, 610,12-1990)

Requirements Engineering (RE): "The process of discovering that 'purpose' by identifying stakeholders and their needs, and documenting them in a form that is amenable to analysis, communication, and subsequent implementation" (Nuseibeh & Easterbrook, 2000)

Stakeholder: "Anyone who can share information about the system, its implementation constraints or the problem domain, including end users, indirect users, other customer representatives and developers" (Potts et al., 1994)

Viewpoint (VP): "Any one who may have some direct or indirect influence on the system requirements" (Kotonya & Sommerville, 1996)

Viewpoint Oriented Requirements Definition (VORD): "A software requirements engineering approach used to organise both the elicitation process and the requirements themselves into viewpoints" (Sommerville, 1995)

Web Business Application (WebApp): "An application that tends to be used to integrate and streamline an organisation's business processes beyond organisational (customers, agents, suppliers, others) and geographical borders, to provide an organisation with competitive products and services that give it a strategic advantage over its competitors in the marketplace

to promote business innovation and/or to improve operational efficiency" (Al-Salem & Abu-Samaha: 2005a)

ENDNOTE

[1] Throughout eWARE, a stakeholder /viewpoint is used interchangeably.

This work was previously published in Encyclopedia of Information Science and Technology, Second Edition, edited by M. Khosrow-Pour, pp. 1537-1545, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 2.6
A Logic Programming Based Framework for Intelligent Web Service Composition

Enrico Pontelli
New Mexico State University, USA

Tran Cao Son
New Mexico State University, USA

Chitta Baral
Arizona State University, USA

ABSTRACT

This chapter presents a comprehensive logic programming framework designed to support intelligent composition of Web services. The underlying model relies on the modeling of Web services as actions, each described by a logic programming theory. This view allows the use of logic-based planning to address the Web service composition problem, taking advantage of the fact that logic-based planning enables the elegant introduction of a number of extensions and generalizations (e.g., dealing with incomplete knowledge and preferences). The theory describing each Web service is encoded as a logic programming module, and different semantics are allowed within different modules, thus better reflecting the practical use of different service description formalisms and ontologies.

DOI: 10.4018/978-1-60566-042-4.ch009

INTRODUCTION

One of the main goals of the Semantic Web initiative (Berners-Lee, Hendler, & Lassila, 2001) is to extend the existing Web technology to support the development of intelligent agents, which can *automatically* and *unambiguously* process the information available in millions of Web pages. This led to numerous works on *Web services* and *Web service composition*. The primary goal of Web service composition is to determine an appropriate sequence of Web services to accomplish a user goal. The majority of the existing proposals dealing with Web service composition build on the principle of viewing Web services as *actions*, thus, representing the Web service composition problem as a *planning problem* that can be addressed using existing *planning* techniques. A second popular approach to Web service composition relies on techniques developed in the area of workflow development. The survey by Rao and Su (2004) provides a good overview of various proposals for Web service compositions.

McIlraith and Son (2002) propose to use *GOLOG* (Levesque, Reiter, Lesperance, Lin, & Scherl, 1997), a logic programming based language, for Web service composition. In such a proposal, each Web service is translated to a primitive action. GOLOG provides control-flow constructs, such as **if-then-else**, **while-do**, **sequence** (denoted by '**;**'), **procedure**, and **test** (denoted by '**?**'), which can be used to combine the primitive actions into programs. The resulting programs can be provided to a *GOLOG* interpreter for finding the sequence of Web services that need to be executed to achieve the goal of the user. Alternatively, the program can be given to an execution monitoring module, for direct execution. Sufficient conditions for a successful execution of a program are also provided. This direction of work has been adopted by Au, Kuter, and Nau (2005), Kuter, Sirin, Nau, Parsia, and Hendler (2005), and Wu, Parsia, Sirin, Hendler, and Nau (2003), where SHOP2, a hierarchical planning system, is used as the underlying system for automatic Web service composition. The latter work also addressed an important aspect of Web service composition, namely, the problem of incompleteness of information in Web service composition, by adding to the planning algorithm a module for gathering information during the planning process.

Viewing Web service composition as high-level planning is not only natural, but also advantageous for different reasons:

- AI planning has made remarkable progress in the last 10 years, and several robust and scalable planning systems have been developed and are available for use, such as *FF* (Hoffmann & Nebel, 2001), *SHOP* (Nau, Cao, Lotem, & Muñoz-Avila, 1999), SAT-based planners (Kautz & Selman, 1996), and logic programming based planners (Lifschitz, 1999). All these planners can be used as the backbone in the development of systems for Web service composition with an architecture similar to the one described by McIlraith and Son (2002).
- AI planning allows the Semantic Web research community to focus on the development of Web service representation and reasoning languages and tools for translating Web service representation into a planning language.

Indeed, this view of Web service composition has been embraced by many researchers, and a number of tools have been proposed, for example, translators to map Web services encoded using DAML-S or OWL-S to PDDL (PDDL Technical Committee, 1998), a well-known planning language used by many planning systems.

While the use of planning in Web service composition is advantageous, there are a number of issues that need to be addressed before this approach can be widely applied. The first problem, which can be termed as the *service selection problem*, derives from the huge number of available services that can be used to achieve the same goal. The second problem lies in the lack of information in service composition, which imposes additional requirements on the planning system, such as the ability to plan with incomplete knowledge and to handle knowledge producing actions. The third problem centers on the *quality* of the composed service, which can be translated into the problem of planning with preferences, an area of research that has only recently attracted interest from the planning community. The fourth problem derives from the fact that each Web service is often encoded within an ontology or a knowledge base, whose semantics is specified by the service provider, and hence, might be different from service to service. A final problem, that has been recognized very early in the development of the Semantic Web (Berners-Lee & Fischetti, 1999; Berners-Lee et al., 2001), is the need for mechanisms to encode *rules*, for example, rules for the description of Semantic Web services and business rules interchange in e-commerce applications.

Answer set programming (ASP) is a declarative programming framework, originally proposed by Marek and Truszczynski (1999) and Niemelä (1999). To solve a problem in ASP, we translate it into a logic program, whose answer sets correspond one-to-one to solutions of the original problems. ASP has found its way in several real-world applications, for example, diagnosis of components of the space shuttle (Balduccini, Gelfond, & Nogueira, 2006). ASP has been widely adopted in the area of reasoning about actions and planning; it has been applied to solve various forms of planning, such as classical planning, conformant planning, and conditional planning (e.g., Son, Tu, Gelfond, & Morales, 2005; Tu, Son, & Baral, 2006). ASP has been used to incorporate various forms of domain knowledge in planning (Son, Baral, Tran, & McIlraith, 2006a) and to construct plans satisfying multidimensional preferences (Son & Pontelli, 2006). This existing body of knowledge provides the foundations for the development of a logic programming-based framework for Web service composition that addresses all but the problem of dealing with Web services encoded in heterogeneous knowledge bases. This last problem is becoming more and more critical with the development of distinct standards and ontologies for Web service encoding and the wider availability of Web services in the Internet. *This problem will be our main concern in this chapter.*

In this chapter, we propose a logic programming framework for reasoning with distributed heterogeneous knowledge bases that contain rules and facts, that is, we concentrate on the fourth problem faced by the Web service composition problem. Our specific objectives are:

1. The design of a theoretical framework for *reasoning with heterogeneous knowledge bases*, which can be combined with logic programming-based planners for Web service composition. This framework supports:
   a. The interoperation between knowledge bases encoded using different rule markup languages, and
   b. The development and integration of different components that reason about knowledge bases.
2. The development of a prototype of the proposed framework.

The chapter is organized as follows. We start by briefly reviewing the representation of Web services and rules in the Semantic Web, the foundation of logic programming and answer set planning, the use of answer set planning in Web service composition, and discuss the challenges faced by the current architecture for Web service composition. In the successive sections, we describe the main contribution of this chapter, a framework for reasoning with distributed heterogeneous knowledge bases, and its implementation. Finally, we relate our work to other proposals and provide conclusions and indications for future developments.

WEB SERVICES AND LOGIC PROGRAMMING

We begin with a review of the common elements of Web services and the RuleML language, and the basics of logic programming. We then discuss how logic programming can be used in Web service composition and discuss the issues that need to be addressed.

Figure 1. CheckAvailability function for hotel reservation Web service

```
. . .
<types>
  <xs:schema
     xmlns:xs="http://www.w3.org/2001/XMLSchema"
     targetNamespace="http://greath.example.com/2004/schemas/resSvc"
        xmlns="http://greath.example.com/2004/schemas/resSvc">
     <xs:element name="checkAvailability" type="tCheckAvailability"/>
      <xs:complexType name="tCheckAvailability">
        <xs:sequence>
          <xs:element  name="checkInDate" type="xs:date"/>
          <xs:element  name="checkOutDate" type="xs:date"/>
          <xs:element  name="roomType" type="xs:string"/>
        </xs:sequence>
     </xs:complexType>
     <xs:element name="checkAvailabilityResponse" type="xs:double"/>
     <xs:element name="invalidDataError" type="xs:string"/>
  </xs:schema>
</types>
<interface  name = "reservationInterface" >
    <fault name = "invalidDataFault" element = "ghns:invalidDataError"/>
    <operation name="opCheckAvailability"
        pattern="http://www.w3.org/ns/wsdl/in-out"
        style="http://www.w3.org/ns/wsdl/style/iri" wsdlx:safe = "true">
        <input messageLabel="In"
              element="ghns:checkAvailability" />
        <output messageLabel="Out"
              element="ghns:checkAvailabilityResponse" />
        <outfault ref="tns:invalidDataFault" messageLabel="Out"/>
    </operation>
```

Web Services and RuleML

Many Web service description languages have been proposed (e.g., www.w3c.org). The most recent proposal, WSDL 2.0 (Chinnici, Moreau, Ryman, & Weerawarana, 2007), describes a Web service in terms of

- The kinds of messages that the service will send/ receive (using the **types** element).
- The functionalities of the service (**interface** element).
- The way to access the service (**binding** element).
- The location of the service (**service** element).

In essence, a service can be viewed as a collection of subroutines (or functions), and it is described by a valid WSDL 2.0 document. If a service requesting agent would like to use a service, it needs to place a request with the proper parameters. For example, Figure 1 shows an excerpt from Booth and Liu (2005), used to define the 'checkAvailability' function for a hotel reservation service, which takes three parameters (i.e., checkInDate, checkOutDate, and roomType) and returns either an error message indicating that the date is improper (invalidDataError), or a number indicating the number of available rooms (checkAvailabilityResponse).[1]

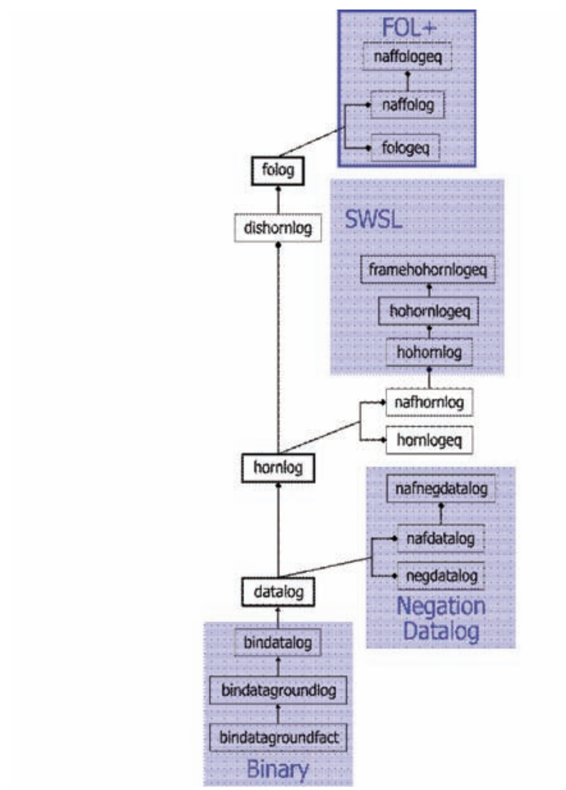
Many of the existing Web services are described within some ontology (e.g., an OWL description). More and more are associated with a set of rules (also called a *knowledge base*). Such a knowledge base is often written in one of the variants of the RuleML markup language.

The RuleML initiative[2] is a response to the need of a common XML-based rule markup language, which has precisely defined semantics and enables efficient implementations. In recent years, a significant amount of work has been devoted to develop knowledge representation languages suitable for the task, and a variety of languages for rule markup have been proposed. The initial design (Boley, Grosof, Sintek, Tabe, & Wagner, 2002) included a distinction (in terms of distinct DTDs) between *reaction* rules and *derivation* rules. The first type of rules is used for the encoding of event-condition-action (ECA) rules, while the second is meant for the encoding of implicational and inference rules.

The derivation rules component of the RuleML initiative has originated a family of languages. Figure 2, from Hirtle and Boley (2005), shows the most commonly referred languages; observe that Datalog plays the role of a core language, with simplified versions (e.g., unary and binary Datalog) developed for combining RuleML with OWL (as in SWRL) (Horrocks, Patel-Schneider, Boley, Tabet, Grosof, & Dean, 2004). Various sublanguages have been created to include features like explicit equality (e.g., fologeq), negation as failure (e.g., naffolog), and Hilog layers (e.g., hohornlog). Specific instances of RuleML for handling different aspects of Web services have been proposed (The Policy RuleML Technical Group, 2004).

Kifer, de Bruijn, Boley, and Fensel (2005) argue that any realistic architecture for the Semantic Web must be based on various independent but interoperable languages, one of them being the logic programming language with negation-as-failure. The use of rule-based languages requires the coexistence of different languages with different semantics and associated reasoning mechanisms, and the need to integrate reasoning across these languages. The need for these languages and their interactions have been extensively discussed (e.g., Kifer et al., 2005; May, Alferes, & Amador, 2005) and it is at the foundation of the most recent work of the Rule Interchange Format working group. It is also important to note that many of the sublanguages of RuleML have been implemented either through translators, for example, GEDCOM (Dean, 2001), which translates to XSB Prolog and JESS, or using independent inference engines, for example,

Figure 2. RuleML language modularization



- j-DREW (Spencer, 2002), a top-down engine for RuleML,
- DR-Device (Bassiliades, Antoniou, & Vlahavas, 2006), an engine supporting defeasible logic and both strong and default negation, and
- CommonRules (Chan & Grosof, 1999), a bottom-up engine for the Datalog sublanguage.

It should be noted that RuleML, WSDL 2.0, and the other markup languages employed when dealing with Web services are simply representation languages. An agent making use of Web services can be developed in any programming language (e.g., JAVA, C++, Perl, etc.). For this reason, previous proposals for Web service composition assume the existence of translators which map Web service descriptions to representations that are more adequate for manipulation and reasoning. For example, McIlraith and Son (2002) map Web service descriptions to situation calculus, while other recent proposals make use of encodings in PDDL; for example, translators of Web service descriptions (in DAML-S) to PDDL (McDermott, Dou, & Qi, 2002).

For the purpose of this chapter, we will assume the existence of translators that map Web service descriptions into a logic programming representation. Many of such tools are already available on the Web (e.g., Rainer, 2005).

Logic Programming

We will consider a logic programming language $\langle \Im, \Pi, \vartheta \rangle$, where $\Im$ is a denumerable collection of function symbols, $\Pi = \Pi_u \cup \Pi_d$ is a denumerable collection of predicate symbols, and ϑ is a collection of variables. Π_u are called user-defined predicates, while $\prod_d$ are called built-in predicates. We will assume that $\Pi_u \cap \Pi_d = \varnothing$. We will denote with $ar(\alpha)$ the arity of the symbol $\alpha \in \Pi \cup \Im$. We typically assume that Π_d contains at least the predicates assert, retract, and model.

A *term* is either a variable, an element of $\Im$ of arity 0 (i.e., a constant), or an expression of the form $f(t_1,\dots,t_n)$ where $f \in \Im$, $ar(f)$=n and $t_1,\dots,t_n$ are terms. We will say that a term t is ground if it does not contain variables. We will denote with H_P the Herbrand universe for this language, that is, the set of all ground terms.

An *atom* is a formula of the form $p(t_1,\dots,t_n)$ where $p \in \Pi$, $ar(p)$=n and $t_1,\dots,t_n$ are terms. The atom is ground if $t_1,\dots,t_n$ are ground. A *qualified atom* is a formula of the form $t{:}A$ where t is a ground term (called the *label* of the qualified atom) and A is an atom. In particular, if the predicate p of an atom belongs to Π_d, then the atom can only appear qualified in a rule. A literal is an atom, a qualified atom, or a formula *not* A, where A is an atom/qualified atom. *not* A is also referred to as a negative literal. We will denote with B_P the Herbrand base for this language (i.e., the set of all ground atoms). For an atom (qualified atom, negative literal) l, we denote with $\pi(l)$ the predicate symbol used in l.

A general rule is of the form

$A \leftarrow B_1,\dots,B_k$ (1)

where A is an atom and $B_1,\dots,B_k$ are literals. Intuitively, a Rule (1) states that if the literals $B_1,\dots,B_k$ are true, then we can conclude that A is also true. We view interpretations and models as subsets of B_P. For an atom A, the literal A (resp. *not* A) is true in a model M if $A \in M$ (resp. $A \notin M$).

Depending on the type of programs we wish to represent, different restrictions can be imposed on the rules:

a. **Datalog:** The B_i's in Rule (1) can be only atoms/qualified atoms and the terms used in the literals can be only variables or constants (i.e., of arity 0).
b. **Ground datalog:** The B_i's in Rule (1) can only be atoms/qualified atoms, and the only terms allowed are constants.
c. **Ground binary datalog:** The rules satisfy the conditions of Case (b), and in addition we require all predicates used to construct atoms to have arity at most 2.
d. **Datalog with negation:** The rules have the format as in Case (a) but negative literals (*not* A) are allowed in the body of the rule.

We will refer to a rule as a Ξ-rule (where Ξ is Datalog, ground Datalog, binary Datalog, etc.) to denote a rule that meets the corresponding requirements. A Ξ-program is a collection of Ξ-rules. Given a rule r, we denote with $used(r)$ the set of ground terms t such that t is a label of a qualified atom in r. Given a Ξ-program, we denote with

$used(P) = \{\ t \mid \exists\ r \in P,\ t \in used(r)\ \}$.

We also introduce

def(P) = { p | $p \in \Pi$, *ar*(p)=k, $\exists r \in P$. $\exists\, t_1,\ldots,t_k$, *head*(r) = $p(t_1,\ldots,t_k)$ }.

Intuitively, a Ξ-program encodes a knowledge base, whose semantics is defined by a class of models satisfying certain properties. It should be noted that the semantics of a program might be defined in different ways and depends on the program type. For Datalog programs, the least fixpoint semantics can be used (van Emden & Kowalski, 1976). However, for Datalog with negation programs, the well-founded (Van Gelder, Ross, & Schlipf, 1991) or the answer set semantics can be used (Gelfond & Lifschitz, 1988).

Planning in Logic Programming and Web Service Composition

A planning problem is specified by a domain, with its properties (called *fluents*) and *actions* with their preconditions and effects, a specification about the initial state of the world, and a formula describing the final state (*goal*). The objective is to determine a sequence of actions that can transform the initial state into a state satisfying the goal.

Logic programming under the answer set semantics was first used in planning by Subrahmanian and Zaniolo (1995) and is now known as answer set planning (Lifschitz, 2002). Answer set planning has gained popularity thanks to the development of fast answer set solvers (e.g., Smodels, Simons, Niemelä, & Soininen, 2002 and DLV Eiter, Leone, Mateis, Pfeifer, & Scarcello, 1998). Since our intention is to use logic programming for Web service composition, following the AI planning approach, we will give a brief overview of answer set planning by illustrating its key ideas through an example. As noted earlier, existing tools can be adapted to translate Web service descriptions into the representation discussed in this chapter.

Answer Set Planning

Let us consider the classical AI problem of going to the airport (McCarthy, 1959). In this problem, we have an action, named *drive_home_airport*; in the initial state, we are at home and have a car. We know that driving will bring us to the airport and our goal is to go be at the airport.

In answer set planning, this problem is encoded by a Datalog with negation program that consists of the different groups of rules described below. In each of these rules, T denotes a number between 0 and a predefined constant length, indicating the maximal length of the plan that we wish to compute. The rules

action(drive_home_airport) ←
fluent(at_home) ←
fluent(at_airport) ←
fluent(car_available) ←

specify the fluents and the actions of the domain. The rules

false← not holds(at_home,T), occ(drive_home_airport), T)
false← not holds(car_available,T), occ(drive_home_airport), T)

encode the precondition under which the action drive_home_airport can be executed. They state that the action drive_home_airport can only occur at the time T if at_home and car_available are true at the time T. The rules

holds(at_airport,T+1)←occ(drive_home_airport, T)
holds(neg(at_home),T+1)←occ(drive_home_airport, T)

say that if the drive_home_airport action occurs at the time T, at_airport and at_home will be true and false, respectively, at the time T+1. The two rules

occ(drive_home_airport,T)←not not_occ(drive_home_airport,T)
not_occ(drive_home_airport,T)←not occ(drive_home_airport,T)

are often referred to as *generation rules,* and are used to generate action occurrences. Intuitively, they state that the action either occurs or not. The rules

holds(at_airport,T+1)←holds(at_airport, T),

not holds(neg(at_airport), T+1)

holds(neg(at_airport),T+1)←holds(neg(at_airport), T),

not holds(at_airport, T+1)

known as *inertial rules*, encode the fact that a fluent normally does not change its value. (We omit the rules for at_home and car_available for brevity.)

Finally, the initial state is encoded by a set of facts of the form

holds(neg(at_airport),0) ←

holds(at_home, 0) ←

and the goal is expressed by the rule

false←not holds(at_airport,length)

Each answer set of the above program corresponds to a plan achieving the goal at_airport. For example, for length=1, the program yields an answer set corresponding to the plan [drive_home_airport]. For more on answer set planning, the interested reader is referred to Lifschitz's (2002) work.

Answer Set Planning and Web Service Composition

Answer set planning can be used for Web service composition, but it requires some modifications and extensions. For example, the functions of the hotel reservation Web service (Figure 1) can be encoded by the following actions:

action(checkAvailability(DateIn,DateOut,RoomType))←date(DateIn),

date(DaeOut),

type(RoomType)

action(makeReservation(DateIn,DateOut,RoomType))←date(DateIn),

date(DaeOut),

type(RoomType)

It is reasonable to consider the two following fluents:

- available: there are some rooms available
- reserved: the reservation has been successful

They can be encoded as follows

fluent(available)←

fluent(reserved)←

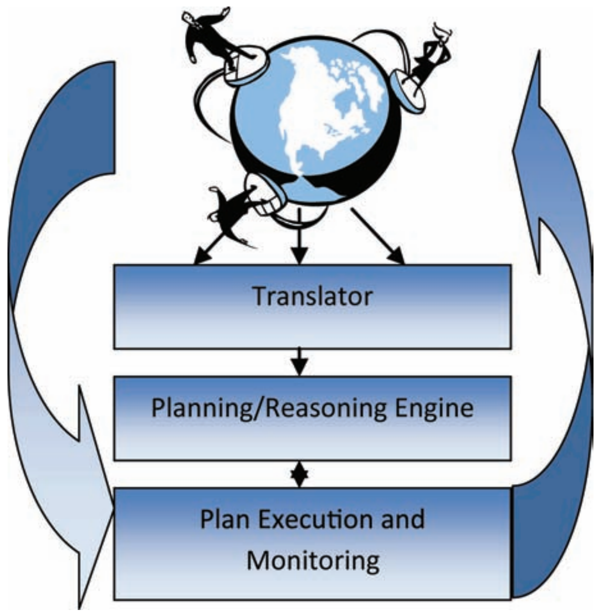
The effect of the action checkAvailability is different than the effect of the action makeReservation. The execution of checkAvailability is *either* available *or* ¬available, while the execution of makeReservation is reserved. The first action, called *knowledge producing action*, does not change the state of the world (it does not alter the number of available rooms) while the second one does (it reduces the number of available rooms by 1).

Knowledge producing actions are often needed when a reasoner does not have complete information about the state of the world. For example, an agent that uses the above service to place a reservation often does not know whether or not the hotel would have a room. In other words, neither available nor ¬available is true in the initial state. The agent knows about the value of available (or ¬available) after the execution of the action checkAvailability. In our previous work (i.e., Son et al., 2005; Tu, Son, & Baral, 2006), we discuss the use of logic programming in reasoning and planning with knowledge producing actions and incomplete information. We also extended answer set planning to incorporate different types of domain knowledge and user preferences (Son, Baral, Tran, & McIlraith, 2006b; Son & Pontelli, 2006).

Current Architecture and Challenges

The techniques developed so far can be used in the implementation of a Web service composition framework similar to other AI-planning Web service composition frameworks (Figure 3). The basic components of this approach are:

Figure 3. Schematic architecture



- A *translator* which reads Web service descriptions and translates them into an adequate logic programming theory (e.g., a Datalog with negation program).
- A *reasoning/planning engine* which accepts the output of the translator and creates a plan that achieves the goal of the service composition problem.
- An *execution and monitoring* module which is responsible for the execution of the plan.

Architectures analogous to the one depicted in Figure 3 have been successfully implemented, for example, in the development of systems for Web service composition of bioinformatics Web services (e.g., (Pan, Tu, Pontelli, & Son, 2004)). In this application, the results of the translation are Datalog, Datalog with negation, or a similar variant language. The reasoning/planning engine employs the answer set semantics for this program. Furthermore, it is assumed that all necessary information has been incorporated into the encodings of the corresponding services. Although the resulting system is adequate for the mentioned application, the new developments in Semantic Web present two technical challenges to this architecture.

The first problem lies in the fact that knowledge (and hence, Web services) often comes with *reasoning rules*, that were not the focus of earlier Web service description languages, and hence this aspect has not been addressed in previous implementations. The second challenge is the need to consider different semantics embedded in the encoding of a Web service, since the underlying semantics of various RuleML knowledge bases are specified by the knowledge engineers and not by the application developers. In dealing with this problem, we can identify a number of issues that need to be addressed:

1. *Reasoning within one knowledge base:* Being able to reason within a knowledge base implies the ability to interoperate with a computational framework capable of handling the type of knowledge present in the knowledge base (e.g., a Prolog system for hohorn rules, a Datalog system for Datalog rules).

2. *Reasoning across different knowledge bases:* The capability of combining knowledge is essential for intelligent agents (e.g., this is necessary in the context of services composition). This requires
    a. The ability to exchange inference results between different knowledge bases (e.g., the interoperability problem between rules and OWL described by Kifer et al. [2005]);
    b. The ability to combine reasoning results produced by different reasoning engines; and
    c. The ability to properly scope the reasoning w.r.t. a specific knowledge base (e.g., the scoped inference issue described by Kifer et al. [2005]).
3. *Utilizing available knowledge:* This requires the ability to use the results produced by different reasoning processes in the construction/implementation of complex Semantic Web applications.

A Motivating Example

Let us illustrate some of these issues within an example, drawn from the field of bioinformatics. This application domain is particularly interesting with respect to the issue of Web service composition; in recent years a large number of Web services have been developed, and they rely on a large collection of heterogeneous (and often redundant) ontologies and knowledge bases.

The problem at hand is to obtain reliable functional annotation, according to the Gene Ontology (Gene Ontology Consortium, 2007), of the proteins coded by the genes present in the genome of an organism. The input is represented by the NCBI RefSeq id of a genome (i.e., an id of the form *NC_xxxx*). It has been observed that functional annotations retrieved by AmiGO (the query system to the Gene Ontology Database) are occasionally unreliable (e.g., Andorf, Dobbs, & Honavar, 2007). In particular, it has been observed that some incorrect annotations in the Gene Ontology are characterized by a selection code equal to reviewed computational analysis (RCA). We consider the following services and knowledge bases:

- **NCBI GenBank Genome-to-Protein Search:** This service expects as input the RefSeq of a genome, and produces as results the NCBI ids of the proteins coded by the genes in the considered genome;
- **Protein Class Ontology:** Gene Ontology includes a taxonomy of functional classification of proteins (e.g., its OBO encoding can be found in “Gene Ontology, Edit 1.101, ” [2007]). The taxonomy can be directly encoded in Datalog; sample rules extracted are:

```
id(0016301)
name(0016301, ‘kinase activity’)
namespace(0016301,molecular_function)
is_a(0016301,0016772)
...
id(0004672)
name(0004672,’protein kinase activity’)
alt_id(0050222)
xref(reactome(4030))
is_a(0004672,0016301)
is_a(0004672,0016773)
...
```

Ontological reasoning allows us, for example, to discover subclasses and superclasses of a given one.

- **Functional Annotation Service:** The input is an id of a protein. If the protein belongs to a class whose functional annotation is not RCA, then the annotation is retrieved from the Gene Ontology annotation database and returned. Otherwise, the annotation is extracted from the UniProt database (and returned in UniProt format). The semantics of this service requires the use of more complex rules (encoded in RuleML using the nafdatalog language). Some of the Datalog with negation rules are:

suspect(ProteinClass) ←
amigo:evidence_code(ProteinClass,rca)
classification(ProteinClass,
FuncAnnotation,go)←
amigo:is_a(ProteinClass,FuncAnnotation),
not suspect(ProteinClass)
classification(ProteinClass,FuncAnnotation,
uniprot) ←
suspect(ProteinClass),
interpro2go:entity_xref(ProteinClass,X), uni
prot:is_a(X,FuncAnnotation)

- **Gene Ontology to InterPro Ontology Mapping:** This mapping, expressed in a ground Prolog facts database is available at the "OBO download matrix" (2007).

Let us assume that the initial state includes the RefSeq NC_003075 (Arabidopsis thaliana chromosome 4) and we are interested only in proteins that are enzyme inhibitors. The genome includes 4,817 coding genes. One of the proteins generated is AT4G00080. The Gene Ontology can be used (through a transitive closure of the is_a relationship) to determine that the protein belongs to the desired class. Accessing the functional annotation, we discover that the protein is annotated as a pectinesterase inhibitor (GO:0046910), but it is RCA. This means that the desired annotation will be extracted from UniProt, and this returns an annotation IPR006501. The mapping between ontologies contains the fact

metadata_db:entity_
xref('GO:0046910','IPR006501')

This will eventually produce the same Gene Ontology annotation.

A FRAMEWORK FOR INTEROPERATION AND COMPOSITION

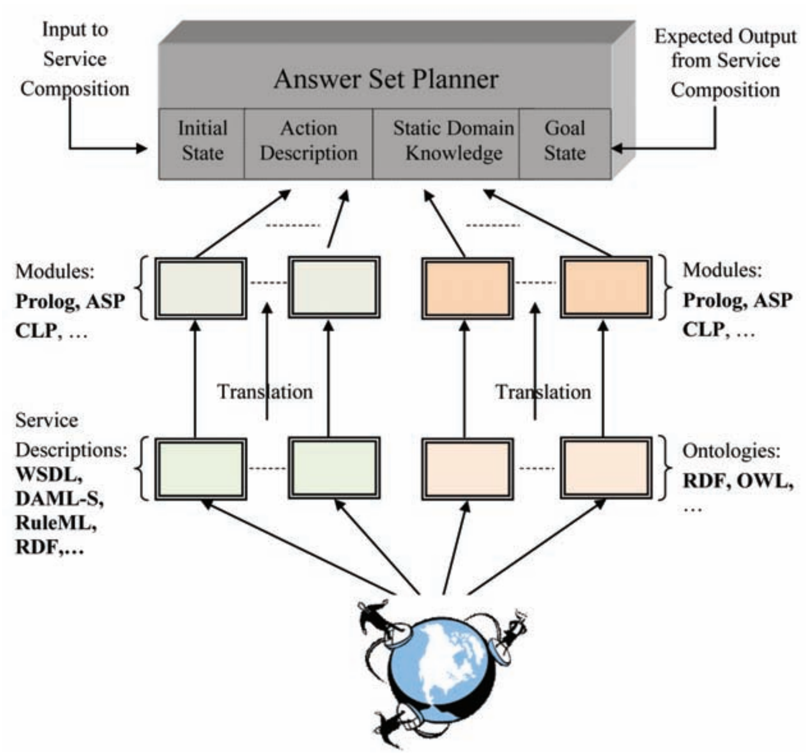
In this section, we describe our framework for interoperation and composition of distributed heterogeneous knowledge bases. We begin with an overview of the approach. We then present the architecture of the component needed to address the challenges faced by the current Web service composition methodology. The precise syntax and semantics of the framework are discussed next.

Overview of the Approach

The objective is to provide a formal logic-based framework that allows the development of the planning/reasoning component of Figure 3. The approach we follow is intuitively depicted in Figure 4. The framework is assembled as a collection of modules, each containing a logic programming theory, potentially requiring a different semantics (e.g., datalog/pure prolog, constraint logic programming, answer set programming, etc.) The modules are the result of a logic programming encoding of the semantics of the Web services description (left side of Figure 4) and of the underlying associated static knowledge (e.g., ontologies – shown on the right in Figure 4). The framework assumes the possibility of dependencies between the different modules, for example, ontologies can be hierarchically defined.

The answer set planner is encoded using a modular view of the action domain. Each service is represented by a logic programming module, which exports the description of the service as an action; in particular, the module exports the description of the action in terms of its executability conditions (predicate executable) and its dynamic effects (predicate causes). For example, in the case of the functional annotation service mentioned previously, these predicates could be defined as:

Figure 4. Framework architecture



executable([have_protein(X), have_protein_class(C)])
causes(go_classification(X,C)) ←
have_protein(X),
have_protein_class(Class),
classification(Class,C,go)
causes([validated_annotation,uniprot_classification(X,C)]) ←
have_protein(X),
have_protein_class(Class),
classification(Class,C,uniprot)

The planner will use these predicates to develop plans, for example, using rules like:

holds(F,T+1) ← action(A),
occ(A,T),
A:causes(F),
A:executable(List),
holds(List,T)

where the execution of the action A requires recovering from the corresponding module (which we assume has the same name as the action) the executability conditions (which are tested in the state of the world at time T) and the effects of the action.

Static causal laws will be employed to allow the planner to make use of background knowledge used by the Web services (e.g., ontologies, databases). For example, the protein class ontology, imported as a Prolog module, would allow us to derive rules like:

holds(have_protein_class(C),T) ←
holds(have_protein_class(C1),T),
gene_ontology:is_a(T1,T)

Architecture

The approach adopted in this work relies on the use of a core logic programming framework to address the issues of integration and interoperation. In particular, the spirit of our approach relies on the following beliefs:

- The natural semantics of various languages for Web service description, ontology description, and various levels of the RuleML deduction rules hierarchy can be naturally captured by different flavors of logic programming.
- Modern logic programming systems provide foreign interfaces that allow declarative interfacing to other programming paradigms.

The idea is to combine the *ASP-Prolog* framework of Elkhatib, Pontelli, and Son, (2006)—which allows CIAO Prolog (Bueno, Cabeza, Carro, Hermenegildo, López-García, & Puebla, 1997) programs to access and modify modules containing answer set programming (ASP) code (Niemelä, 1999)—with the notation for modularization of answer set programming of Answar, Baral, and Tari (2005) and Baral, Dzifcak, and Takahashi (2006). The result is a logic programming framework, where modules responding to different logic programming semantics (e.g., Herbrand minimal model, well-founded semantics, and answer set semantics) can coexist and interoperate. The framework provides a natural answer to the problems of use and interoperation of RuleML and other Web service semantic descriptions described earlier. The overall structure is depicted in Figure 4. Most of the emphasis is on the use of answer set programming to handle some of the RuleML sublanguages (e.g., datalog, ur-datalog, nafdatalog, and negdatalog), even though the core framework will naturally support most of the languages (e.g., hornlog and hohornlog).

The problems mentioned in the previous section are addressed by the proposed framework as follows:

- *Issue 1*: CIAO Prolog offers direct access to a collection of modules that support different forms of logic programming reasoning, for example, traditional Prolog, constraint logic programming (over finite domains and reals), fuzzy Prolog, and a declarative ODBC interface. In addition, CIAO Prolog provides a mechanism that allows Prolog programs to invoke Java methods, offering a bidirectional communication and a reflection of Java objects into Prolog. This provides, for example, a natural way to execute Java-based engines (e.g., Jess) and communicate between the core framework and external Java packages. Furthermore, CIAO Prolog includes the PiLLoW library, a standardized Prolog library for Web programming, which provides the framework with capabilities for Web access (e.g., management of URLs) and parsing of HTML and XML documents to Prolog terms. Thus, we envision the core framework as the bridge between distinct execution models for heterogeneous knowledge bases.
- *Issue 2*:
  - *Issue 2(a)*: This issue will be addressed through the introduction of a module system, where different knowledge bases can be encoded (directly or indirectly) as distinct modules. The original import/export of CIAO Prolog can be combined with the languages for answer set modules of Baral et al. (2006) to allow forms of bidirectional communication between the core framework and the modules representing the knowledge bases.
  - *Issue 2(b)*: The core framework will provide the full computational power of Prolog, constraint logic programming, and answer set programming, combined through a sophisticated module and class system. Module interfaces will allow extraction of semantic information from the various knowledge bases (e.g., result of queries, models of knowledge bases) and reason with them.

- ◦ *Issue 2(c)*: The scoped inference is naturally supported by the module system of ASP-Prolog, for example, skeptical and credulous reasoning w.r.t. answer set modules.
- *Issue 3*: This aspect can be handled thanks to the combination, in ASP-Prolog and CIAO Prolog, of Web access capabilities along with the full computational power of Prolog.

We will now introduce the basic structure of our framework for modules and module interoperation.

Module Structure

As mentioned in the introduction to logic programming, we will assume that assert,retract,model,… are elements of $\prod_d$. Given a Ξ-program *P*, we say that *P* is pure if neither assert nor retract appear in the rules of *P*; otherwise we say that P is impure.

A module is composed of two parts: a module interface and a module body. A module interface has the form

← module: *t*
← import: $t_1,\ldots,t_k$
← export: $q_1/k_1,\ldots,q_m/k_m$
where

- *t* is a ground term, called the name of the module,
- t1,…,tk are ground terms, representing names of other modules, and
- q1,…,qm are predicates, and are k1,…,km non-negative integers, such that $ar(q_i)=k_i$.

The body of a module is a Ξ-program, for a given Ξ. In that case, we will say that the module is a Ξ-module. Given a module named *t*, we identify with the *export set of t* (denoted by *exp(t)*) the predicates $q_1,\ldots,q_m$ exported by *t*. We also identify with *imp*(*p*) the import set of *t*, that is, the names of the modules imported by *t*.

A program $P=\{M_{t1}, \ldots, M_{tk}\}$ is a collection of modules named $t_1,\ldots,t_k$. The *graph* of *P* (denoted by *graph*(*P*)) is a graph (N,E), where the set of nodes *N* is $\{t_1,\ldots,t_k\}$and

A program *P* is *admissible* if it satisfies the following properties:

- for each t_j we have that $imp(t_j) \subseteq \{t_1,\ldots,t_k\}$; and
- the graph *graph*(*P*) is acyclic.

The module structure can be expanded by allowing cyclic dependencies (i.e., two-way communications between modules) as well as OO-style organization of modules (e.g., as described by Baral et al. [2006]). We omit this discussion due to lack of space.

Example: Let us consider the modules associated to the bioinformatics services described earlier. One module, *gene_ontology*, is derived from the Gene Ontology. It will have an overall structure of the form:

← module: gene_ontology
← export: is_a/2, evidence_code/2, …
id(0016301)
is_a(0016301,0016772)
is_a(0004672,0016301)
is_a(0004674,0004672)
…
is_a(X,Y) ← is_a(X,Z), is_a(Z,Y)
…
evidence_code(0016772, rca)
…

A second module is obtained from the RuleML semantics of the functional annotation service:

← module: validated_annotation
← import: gene_ontology, …
← export: executable/1, causes/1, …
suspect(ProteinClass) ←
gene_ontology:evidence_
code(ProteinClass,rca)
classification(ProteinClass,
FuncAnnotation,go) ←
gene_ontology:is_a(ProteinClass,FuncAnno
tation),

not suspect(ProteinClass)

...

The resulting program P will be a set of modules { *planner, gene_ontology, validated_annotation*, ...}. □

General Semantics

In order to understand the framework, it is important to clearly define the semantics of the module-based system. We start by considering pure programs, and then we extend the discussion to impure programs.

Pure Programs

In this section, we propose a model-theoretic semantics for programs that do not contain any impure modules. Given a program, a *model naming* function τ is an onto function $\tau: H_P \rightarrow 2^{B}$P. We will use this function to assign *distinct* names to the models of the different modules. In the rest of this work, we will assume that the function τ is fixed.

Given a program $P = \{M_{t1}, \ldots, M_{tk}\}$, the acyclic nature of *graph*(P) guarantees the ability to construct a topological sort of $\{t_1,\ldots,t_k\}$, say $\eta_1,\ldots,\eta_k$, such that if (η_i,η_j) is an edge of the graph, then $i < j$.

Given the program P and a topological sorting of the modules $\eta_1,\ldots,\eta_k$, we construct the semantics module by module, following the topological sort order. The semantics of each module M_i will be given by a collection of models $\Phi^\tau(M_i)$, where $\Phi^\tau(M_i) \subseteq 2^{B}$P. Given a Ξ-program T, not containing any qualified atoms and not containing any occurrence of predicates from Π_d, we assume that its semantics NAT(T) is given. For example, if T is a Datalog with negation program meeting these conditions, then NAT(T) will be the set of answer sets of T, while if T is a pure Prolog program (i.e., definite logic program without extra-logical predicates such as assert and retract), then NAT(T) contains the least Herbrand model of T.

This suggests a natural way to handle the semantics Φ^τ of a program P. Φ^τ is a mapping of the form $\Phi^\tau: P \rightarrow 2^{B}$P. Once the topological sort $\eta_1,\ldots,\eta_k$ of the modules is given, we can construct Φ^τ as follows:

- the semantics of $M_{\eta 1}$ is given, since it does not import any other modules, and

$\Phi^\tau(M_{\eta 1}) = \mathrm{NAT}(M_{\eta 1})$

- the semantics of $M_{\eta i}$ can be constructed by computing the natural semantics of a "reduct" of the module,

$\Phi^\tau(M_{\eta i}) = \mathrm{NAT}(MR(M_{\eta i}, \Phi^\tau))$

as defined next.

Let us consider a module $M_{\eta i}$ of P. Then:

- If t: A is a ground qualified atom and $t \in$ *imp*$(M_{\eta i})$, then

$M_{\eta i} \models_\Phi \tau$ t: A iff for each model M $\in \Phi^\tau(M_t)$ we have that M $\models$ A.

- If t: A is a ground qualified atom and $t \notin$ *imp*$(M_{\eta i})$, then $M_{\eta i} \models_\Phi \tau$ t: A if there exists $x \in$ *imp*$(M_{\eta i})$, M $\in \Phi^\tau(M_x)$ such that π(A) ∈ exp(x), τ(t)=M and M $\models$ A.
- $M_{\eta i} \models_\Phi \tau$ t: model(s) if t ∈ *imp* $(M_{\eta i})$ and τ(s) ∈ $\Phi^\tau(M_t)$.
- If *not t: A* is a ground qualified literal and $t \in$ *imp*$(M_{\eta i})$, then

$M_{\eta i} \models_\Phi \tau$ not t: A if $M_{\eta i} \not\models_\Phi \tau$ t: A.

The *model reduct* of M_i w.r.t. Φ^τ, denoted $MR(M_i, \Phi^\tau)$, is defined as follows:

- Remove from M_i all rules that contain in the body a qualified element l such that $M_i \not\models_\Phi \tau l$.
- Remove from the remaining rules all occurrences of qualified elements.

One can easily see that $MR(M_i, \Phi^\tau)$ is a program without qualified atoms whose semantics is defined by NAT. This allows us to set

$\Phi^\tau(M_{\eta i}) = \mathrm{NAT}(MR(M_{\eta i}, \Phi^\tau))$

From now on, we will denote with $\Phi^{\tau/P}$ the semantics of a program P.

Example: Let us continue the example we started earlier. The module *gene_ontology* is a Datalog program. Thus, NAT(*gene_ontology*) = { M_0 } where M_0 is its least Herbrand model; some of the elements in M_0 are:

is_a(0016301,0016772), is_a(0004672,0016301),

is_a(0004674,0004672), is_a(0004672,0016772),

is_a(0004674,0016772), is_a(0004674,0016301),...

The module *validated_annotation* is a Datalog with negation program, and its reduct will be performed with respect to the semantics of *gene_ontology* (and the other imported modules). For example, the rule

classification(ProteinClass, FuncAnnotation, go) ←

gene_ontology:is_a(ProteinClass, FuncAnnotation),

not suspect(ProteinClass)

will be grounded and only the instances for which the is_a atom is true in M_0 will be kept, for example,

classification(0004674, 0016772, go) ← not suspect(0004674)

Only the answer sets of the resulting program will be maintained as NAT(*validated_annotation*). □

Impure Programs

We say that a program is impure if it contains impure modules and/or it contains modules that are not based on logic programming. The use of impure programs allows the planner to provide the Web services with knowledge about the state of the world the services will be executed in. It also allows the reasoner to temporarily modify the content of a module (e.g., a module describing an action) to support the planning process. This is particularly important when dealing with very complex plans (Son & Pontelli, 2007) or incomplete knowledge (Tu et al., 2006).

Let $P=\{M_{t1}, \ldots, M_{tk}\}$ be a program; for the sake of simplicity, we assume that $t_1,\ldots,t_k$ is a topological sort of *graph*(P). We also consider impure programs under the following restrictions:

- The planner is itself a module, encoded using Prolog (pure or impure), and it is represented by the module M_{tk}.
- The impure predicates assert and retract are allowed to appear only in Prolog modules.

Because of the nonlogical nature of the impure predicates, we rely on an operational semantics to chaacterize the meaning of programs.

The *state* of a computation is given by a tuple $\langle G,\theta, P\rangle$, where G is a Prolog goal, θ is a substitution, and P is a program. The operational semantics for a goal executed in the module named t_i is defined through a state transition system $\langle G, \theta, P\rangle \rightarrow_i \langle G', \theta', P'\rangle$ where

- If $G = (A \wedge Rest)$ and $\pi(A) \in def(M_{ti})$, and $h \leftarrow body$ is a variant of a rule in M_{ti} such that σ is the most general unifier of $A\theta$ and h, then we have that $G'=body\wedge Rest$, $\theta'=\theta\sigma$,[3] and $P'=P$.
- If $G = (t{:}\ A \wedge Rest)$, $t_j \in imp(M_{ti})$, $\pi(A) \in def(M_{tj})$, t_j is a Prolog module, and $\langle A, \theta, P\rangle \rightarrow_j \langle \square, \theta', P'\rangle$ then we have $G'=Rest$
- If $G = (t{:}\ A \wedge Rest)$, $t_j \in imp(M_{ti})$, $\pi(A) \in def(M_{tj})$, t_j is an ASP/datalog module, and let σ be a ground substitution fpor $A\theta$ such that $A\theta\sigma$ is true in each model in $\Phi^{\tau/P}(M_{tj})$ then we have that $G'=Rest$, $\theta'=\theta\sigma$, and $P'=P$.
- If $G = t{:}\ model(t') \wedge Rest$ and $t \in imp(M_{ti})$, and σ is a ground substitution such that $t'\theta$ is ground and $\tau(t'\theta) \in \Phi^{\tau/P}(M_t)$, then $G'=Rest$, $\theta'=\theta\sigma$, and $P'=P$.
- If $G = t{:}\ A \wedge Rest$ and there exists $t' \in imp(M_{ti})$, and σ is a substitution such that $(t{:}\ A)\theta\sigma$ is ground, $\tau(t\sigma) \in \Phi^{\tau/P}(M_{t'})$, and $A\theta\sigma$ is true in $\tau(t\sigma)$, then we have that $G'=Rest$, $\theta'=\theta\sigma$, and $P'=P$.

- If $G = t$: *assert*(*Head*, *Body*) $\wedge$ *Rest* and σ is a substitution such that $t\sigma \in imp(M_{ti})$, then we have that $G' = Rest$, $\theta' = \theta\sigma$, and $P' = (P \setminus \{M_{t\sigma}\}) \cup N_{t\sigma}$ where $N_{t\sigma} = M_{t\sigma} \cup \{ (Head \leftarrow Body)\, \sigma \}$.
- If $G = t$: *retract*(*Head*, *Body*) $\wedge$ *Rest* and σ is a ground substitution such that $t\sigma \in imp(M_{ti})$ and $(Head \leftarrow Body)\sigma \in M_{t\sigma}$, then we have that $G' = Rest$, $\theta' = \theta\sigma$, and $P' = (P \setminus \{M_{t\sigma}\}) \cup N_{t\sigma}$ where $N_{t\sigma} = M_{t\sigma} \setminus \{(Head \leftarrow Body)\sigma\}$.

Given a goal *G* and a program *P* with main module M_{tk}, we say that θ is a solution of *G* if

where $\rightarrow^*$ denotes an arbitrarily long sequence of transitions.

Example: Let us continue with our example. The ability to perform service composition using logic-based planning provides the added advantage that the plan and the complete trajectory are represented by logic statements, and it is possible to write additional modules to reason about the trajectory. For example:

- Given a list of proteins L we can write a simple query that will provide the functional class to which the majority of the proteins belong to:

```
majority(L, Class) ←
findall(C,(member(X,L),annotate(X,C)),Funs),
findall(K-F, (member(A,Funs),
findall(A,member(A,Funs),S),
length(S,K)
),
Res),
keysort(Res,Result),
last(Result,Class)
annotate(P,F) ←
assert(planner:initially(have_protein(P))),
planner:holds(go_classification(P,F),length)
```

- Let us assume that the service description includes an estimate of the reliability of the service, provided by the service provider and measuring the ability of the service to quickly respond to a request. This is translated in a fact of the form low_load or high_load. The logic programming module allows the action to successfully return only in case of low load, that is, the definition of the action in the service description module will include

```
executable(…):- low_load.
```

If the resulting plan is sufficiently long, we would like to replace the load information with a pessimistic value, which may reflect the fact that the load figure is not reliable any longer. We could modify the rule used by the planner:

```
holds(F,T+1) ← action(A), occ(A,T),
(T > Threshold -> retract(A:low_load); true),
A:causes(F), A:executable(List), holds(List,T)
```

□

FRAMEWORK IMPLEMENTATION

Logical Core Implementation

The implementation of the logical core is based on the combination of two logic programming systems: *CIAO Prolog* (Bueno et al., 1997) and *Smodels* (Niemelä & Simons, 1997). *CIAO Prolog* is a full-fledged Prolog system, with a sophisticated module system, and designed to handle a variety of flavors of logic programming, including constraint logic programming (over reals and finite domains), fuzzy logic programming, and concurrent logic programming. *Smodels* is a logic programming engine which supports computation of the well-founded and answer set semantics for NAF-datalog programs. The proposed system is composed of two parts: a *preprocessor* and the actual CIAO Prolog system.

The input to the preprocessor is composed of

- The main planner module (*Pr*), typically a Prolog module, which extracts information from a planner encoded in ASP;

- A collection of CIAO Prolog modules ($m_1,\ldots, m_n$);
- A collection of ASP modules ($e_1,\ldots, e_m$).

The main task of the preprocessor is to transform the ASP modules to enable their use in the CIAO Prolog system and to manage the interaction with the *Smodels* solver.

The transformation of each ASP module leads to the creation of two entities that will be employed during the actual program execution, that is, an *interface module* and a *model class*. These are described in the following subsections. The preprocessor will also automatically invoke the CIAO Prolog top-level and load all the appropriate modules for execution. The interaction with the user is the same as the standard CIAO Prolog top-level.

Interface Modules

The preprocessor generates one interface module for each ASP module present in the original input program. The interface module is implemented as a standard CIAO Prolog module and it provides the client Prolog modules with the predicates used to access and manage the ASP module. The interface module is created for each ASP module by instantiating a generic module skeleton

The overall structure of the interface module is illustrated in Figure 5. The module has an export list which includes all the predicates used to manipulate ASP modules (e.g., assert, retract, model) as well as all the predicates that are defined within the ASP module.

Each module has an initialization part, which is in charge of setting up the internal data structures and invoke the ASP solver (*Smodels*) for the first time on the ASP module. The result of the computation of the models will be encoded as a collection of *model objects* (see the description of the model classes in the next subsection). The module will maintain a number of internal data structures, including a representation of the ASP code, a representation of the parameters to be used for the computation of the answer sets (e.g., values of constants), a list containing the objects representing the models of the ASP module, a counter of the number of answer sets currently present, and so forth.

The preprocessor creates a graph-module structure which represents the call hierarchy of modules. If a Prolog/ASP module u_p calling another prolog/ASP module u_c then u_p is called the parent module and u_c is one of its children. The tree-module structure is stored in each interface module.

Each interface provides also a timestamp predicate, which is used to inform of the time at which the module's semantics have been last computed (recorded as a discrete system time); each interface module will recompute the local semantics whenever the timestamp of one of the imported modules changes. This allows the system to propagate the effect of changes (e.g., assert/retract) to all modules that depend on the modified one.

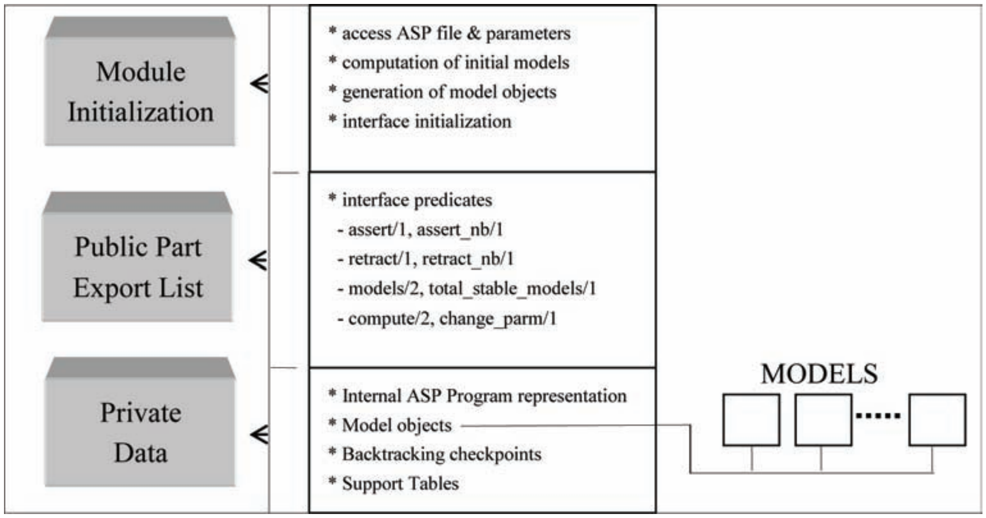
Finally, the preprocessing will add code to handle the different types of goal qualification.

Model Classes

The preprocessor generates a CIAO class definition for each module. The objects obtained from the instantiation of such class will be used to represent the individual models of the module. In particular, for an ASP module we will have one instance for each answer set, while for a Prolog module we will have a single instance, corresponding to the least Herbrand model.

Prolog and ASP modules can obtain reference to these objects (e.g., using the model predicate supplied by the interface module) and use them to directly query the content of one model. The definition of the class is obtained through a straightforward parsing of the export declaration of each module, to collect the names of the predicates defined in it; the class will provide a public method

Figure 5. Structure of an interface module



for each of the predicates present in the module's export list. The class also defines a public method add/1 which is used by the interface module to initialize the content of the model.

To facilitate computation, the preprocessor generates an additional module for each ASP module, called skeptical module. The module is instantiated to a collection of facts representing the elements that have the same truth value in all answer sets of the corresponding ASP module.

Further Implementation Details

Further details about the implementation of this framework can be found by Pontelli, Son, and Baral (2006). Let us discuss only some significant aspects here.

ASP-State: Every ASP-module has a set of models, atoms, and skeptical model called ASP-states. A stack is used to store the ASP-states which will be used for backtracking. Any change in the ASP-state of an ASP module m, requires that all the ASP ancestor modules of m change their own ASP-states. This can be done with the help of the tree-module structure.

Partial Grounder: In order for the Smodels system to ground the ASP-modules, all qualification of the form (α: *t*) have to be computed and stored as new facts with a newly created ASP-variable. Therefore, if we have a positive qualification literal (α: *t*), then the partial grounder will evaluate the predicate *t* in module α. If α is a Prolog module, then the built-in predicate findall is used to compute all possible answers to the goal ?-t. If α is an ASP-module, then the predicate *t* is evaluated against the skeptical answers of the ASP-module α. If α is an ASP-model name, then predicate *t* is computed against the corresponding answer set named α. After that, all ground values of *t* are added as facts into the ASP-module.

The same rewriting rules are needed for negative qualification literals (*not* α: *t*), domain qualification literals, cardinality qualification literals, and weight qualification literals.

RuleML Specific Issues: RuleML knowledge bases are retrieved and encoded as modules to support the reasoning activities. The translation process relies on the PiLLoW library (which supports HTTP protocol and basic XML manipulation) and the sophisticated XML library provided by CIAO Prolog (which allows XML parsing, URIs management, and even XPath queries). The translation process is performed in two steps. During the first step, the RuleML document is parsed and converted into a Prolog XML representation (as a compound Prolog term). In the second phase, the Prolog XML representation is parsed and translated into logical rules and collected into a module.

The import component of the module is automatically derived by retrieving those atoms used in the program and linking (through URIs) to external components (e.g., used in the rel elements). By default, the export list will contain all the rel that appear as heads of rules/facts in the knowledge base.

Another specific issue related to RuleML is the inclusion of ECA rules. They can be effectively included in the proposed framework, as discussed by Pontelli, Son, and Baral (2006).

RELATED WORK

The importance of developing languages and frameworks to integrate different aspects of semantic Web reasoning has been highlighted in the literature. Most of the existing focus has been on integrating rule-based reasoning with ontology reasoning. Two relevant contributions in this field are represented by the work of Golbreich (2004) and Laera, Tamma, Bench-Capon, and Semeraro (2004). Golbreich (2004) describes a combination of reasoning about ontologies (encoded in OWL) with rule-based reasoning (with rules encoded in SWRL and processed by the Jess system). Different from our framework, the system is not based on logic programming (relying on Java's Jess platform) and limited to SWRL's rule format.

A wider-breadth approach, and closer to the spirit of ASP-Prolog, is SweetProlog (Laera, Tamma, Bench-Capon, & Semeraro, 2004). SweetProlog offers a platform for integrating rules and OWL ontologies build using logic programming. Rules are encoded using OWLRuleML, which represents an OWL ontology for the RuleML dialect captured by the courteous logic programming scheme (Grosof, 1999). Both rules and OWL ontologies are mapped to Prolog (specifically, the Java-Internet Prolog engine) for interrogation and reasoning. This project has a greater emphasis on integration between a fixed-rule language (the fragment of RuleML captured by OWLRuleML) and OWL ontologies, providing an orthogonal contribution to the integration problem.

A remarkable direction has been explored in the context of the DLV project; dlvhex (Eiter, Ianni, Schindlauer, & Tompits, 2006) is an extension of DLV:

- It allows answer set programs to invoke external source of computation, that is, truth value of selected predicates can be achieved through external deduction systems, for example, accessing RDF statements; and
- It allows answer set programs to contain higher-order atoms.

This language has been motivated by similar reasons as those in this chapter, to support semantic Web reasoning (Eiter et al., 2006; Eiter, Lukasiewicz, Schindlauer, & Tompits, 2004), with particular emphasis on integration of ontologies.

Relatively few proposals have appeared in the literature dealing with the high-level integration of different forms of logic programming reasoning, specifically top-down goal-oriented Prolog and bottom-up answer set semantics. ASP-Prolog (Elkhatib et al., 2006), on which the work described in this chapter builds, is a system that provides Prolog programs with the ability to seamlessly access modules processed under answer set semantics. A simplified interface, between the *Smodels* system and XSB Prolog, has been described by Castro, Swift, and Warren (2002).

Lower level interfaces between answer set systems (DLV and *Smodels*) and traditional imperative languages have been developed (Calimeri & Ianni, 2005; Ricca, 2003; Syrjanen, 1998).

CONCLUSION AND FUTURE WORK

In this chapter, we addressed the issue of dealing with heterogeneous Web services in the context of the Web service composition problem. We presented a logic programming framework for intelligent Web service composition and a preliminary prototype of parts of the framework within the ASP-Prolog system. The key idea in our framework is to view Web services as actions and Web service composition as a planning problem. Under this view, Web service documents (e.g., WSDL or RuleML documents) are encoded as logic programs which can be reasoned about and combined using logic programming techniques. The variety of semantics of logic programs supports a natural way to deal with heterogeneous Web service description and ontologies. Furthermore, the use of answer set planning in Web service composition allows us to easily incorporate users' preferences and effectively handle incomplete knowledge.

As future work, we propose to demonstrate the framework on real-world applications, with particular focus on applications on description and manipulation of bioinformatics Web services. The proposed framework is also expected to have a significant role in facilitating the deployment of and reasoning about rule-bases constructed according to the guidelines of the rule interchange format (RIF) initiative (Ginsberg, Hirtle, McCabe, & Patranjan, 2006). Recent developments in RIF have highlighted the importance of being able to merge rule sets, and the requirement of supporting standard ways to characterize rule sets dialects. The proposed framework could easily avail of such standardized identification of dialects to guide the automated translation of RIF rule sets to ASP-Prolog modules.

REFERENCES

Andorf, C., Dobbs, D., & Honavar, V. (2007). Exploring inconsistencies in genome-wide protein function annotations. *BMC Bioinformatics*.

Answar, S., Baral, C., & Tari, L. (2005). *A language for modular ASP: Application to ACC tournament scheduling*. Paper presented at the ASP Workshops.

Au, T. C., Kuter, U., & Nau, D. (2005). *Web service composition with volatile information.* Paper presented at the International Semantic Web Conference (ISWC).

Balduccini, M., Gelfond, M., & Nogueira, M. (2006). Answer set based design of knowledge systems. *Annals of Mathematics and Artificial Intelligence*.

Baral, C., Dzifcak, J., & Takahashi, H. (2006). *Macros, macro calls, and use of ensembles in modular answer set programming*. Paper presented at the International Conference on Logic Programming.

Bassiliades, N., Antoniou, G., & Vlahavas, I. (2006). DR-DEVICE, a defeasible logic reasoner for the Semantic Web. *International Journal on Semantic Web and Information Systems*, *2*(1), 1–41.

Berners-Lee, T., & Fischetti, M. (1999). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. Harper.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantics Web. *Scientific American*, *284*(5), 34–43.

Boley, H., Grosof, B., Sintek, M., Tabe, S., & Wagner, G. (2002). *RuleML design, version 0.8*. Retrieved May 30, 2008, from www.ruleml.org/indesign.html

Booth, D., & Liu, C. K. (2005). *Web service description language primer*.

Bueno, F., Cabeza, D., Carro, M., Hermenegildo, M., López-García, P., & Puebla, G. (1997). *The Ciao Prolog system.* Unpublished manuscript.

Calimeri, F., & Ianni, G. (2005). *External sources of computation for answer set solvers*. Paper presented at the Logic Programming and Non-Monotonic Reasoning.

Castro, L., Swift, T., & Warren, D. (2002). *XASP: Answer set programming with XSB and smodels*. Retrieved May 30, 2008, from xsb.sourceforge.net/manual2/node149.html

Chan, H., & Grosof, B. (1999). *CommonRules.* Retrieved May 30, 2008, from www.alphaworks.ibm.com/tech/commonrules

Chinnici, R., Moreau, J.-J., Ryman, A., & Weerawarana, S. (2007). *Web service description language, version 2.0*.

Dean, M. (2001). *RuleML experiments with GEDCOM.* Retrieved May 30, 2008, from www.daml.org/2001/02/gedcom-ruleml

Eiter, T., Ianni, G., Schindlauer, R., & Tompits, H. (2006). *Effective integration of declarative rules with external evaluations for Semantic-Web reasoning*. Paper presented at the European Semantic Web Conference.

Eiter, T., Leone, N., Mateis, C., Pfeifer, G., & Scarcello, F. (1998). *The KR system dlv: Progress report, comparisons, and benchmarks.* Paper presented at the International Conference on Principles of Knowledge Representation and Reasoning (pp. 406-417).

Eiter, T., Lukasiewicz, T., Schindlauer, R., & Tompits, H. (2004). *Combining ASP with description logics for the Semantic Web*. Paper presented at the Principles of Knowledge Representation and Reasoning

Elkhatib, O., Pontelli, E., & Son, T. C. (2006). *A tool for knowledge base integration and querying*. Paper presented at the AAAI Spring Symposium.

Gelfond, M., & Lifschitz, V. (1988). *The stable model semantics for logic programming.* Paper presented at the Logic Programming: Proceedings of the Fifth International Conf1 and Symp. (pp. 1070-1080).

Gene Ontology (Ed.). *1.101*. (2007). Retrieved May 30, 2008, from http://www.geneontology.org/ontology/gene_ontology_edit.obo

Gene Ontology Consortium. (2007). *The gene ontology*. Retrieved May 30, 2008, from www.geneontology.org

Ginsberg, A., Hirtle, D., McCabe, F., & Patranjan, P.-L. (2006). *RIF use cases and requirements*. W3C.

Golbreich, C. (2004). *Combining rule and ontology reasoners for the Semantic Web*. Paper presented at the RuleML.

Grosof, B. (1999). *A courteous compiler from generalized courteous logic programs to ordinary logic programs*. IBM T.J. Watson Research Center.

Hirtle, D., & Boley, H. (2005). *The modularization of RuleML*. Retrieved May 30, 2008, from www.ruleml.org/modularization

Hoffmann, J., & Nebel, B. (2001). The FF planning system: Fast plan generation through heuristic search. *Journal of Artificial Intelligence Research, 14*, 253–302.

Horrocks, I., Patel-Schneider, P., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004). *SWRL: A Semantic Web rule language combining OWL and RuleML* (No. SUBM-SWRL-20040521). W3C.

Kautz, H., & Selman, B. (1996). Pushing the envelope: Planning, propositional logic, and stochastic search. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96)* (pp. 1194-1199).

Kifer, M., de Bruijn, J., Boley, H., & Fensel, D. (2005). *A realistic architecture for the Semantic Web*. Paper presented at the RuleML.

Kuter, U., Sirin, E., Nau, D., Parsia, B., & Hendler, J. (2005). Information gathering during planning for Web service composition. [JWS]. *Journal of Web Semantics*, *3*(2-3), 183–205. doi:10.1016/j.websem.2005.07.001

Laera, L., Tamma, V., Bench-Capon, T., & Semeraro, G. (2004). *SweetProlog: A system to integrate ontologies and rules*. Paper presented at the RuleML.

Levesque, H., Reiter, R., Lesperance, Y., Lin, F., & Scherl, R. (1997). GOLOG: A logic programming language for dynamic domains. *The Journal of Logic Programming*, *31*(1-3), 59–84. doi:10.1016/S0743-1066(96)00121-5

Lifschitz, V. (1999). *Answer set planning*. Paper presented at the International Conference on Logic Programming (pp. 23-37).

Lifschitz, V. (2002). Answer set programming and plan generation. *Artificial Intelligence*, *138*(1-2), 39–54. doi:10.1016/S0004-3702(02)00186-8

Marek, V., & Truszczynski, M. (1999). Stable models and an alternative logic programming paradigm. *The logic programming paradigm: A 25-year perspective* (pp. 375-398).

May, W., Alferes, J., & Amador, R. (2005). *Active rules in the Semantic Web: Dealing with language heterogeneity*. Paper presented at the RuleML.

McCarthy, J. (1959). Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes* (pp. 75-91).

McDermott, D., Dou, D., & Qi, P. (2002). *PDDAML: An automatic translator between PDDL and DAML*.

McIlraith, S., & Son, T. C. (2002). Adapting GOLOG for composition of Semantic Web services. In *Proceedings of the Eighth International Conference on Principles of Knowledge Representation and Reasoning (KR'2002)* (pp. 482-493).

Nau, D., Cao, Y., Lotem, A., & Muñoz-Avila, H. (1999). SHOP: Simple hierarchical ordered planner. In *Proceedings of the 16th International Conference on Artificial Intelligence* (pp. 968-973).

Niemelä, I. (1999). Logic programming with stable model semantics as a constraint programming paradigm. *Annals of Mathematics and Artificial Intelligence*, *25*(3,4), 241-273.

Niemelä, I., & Simons, P. (1997). Smodels: An implementation of the stable model and well-founded semantics for normal logic programs. In *Proceedings ICLP & LPNMR* (pp. 420-429).

OBO download matrix. (2007). Retrieved May 30, 2008, from http://www.berkeleybop.org/ontologies/

Pan, Y., Tu, P. H., Pontelli, E., & Son, T. C. (2004). *Construction of an agent-based framework for evolutionary biology: A progress report*. Paper presented at the Declarative Agent Languages and Technologies (DALT).

PDDL Technical Committee. (1998). *Planning domain definition language*. Author.

Pontelli, E., Son, T. C., & Baral, C. (2006). *A framework for composition and inter-operation of rules in the Semantic Web*. Paper presented at the RuleML.

Rainer, W. (2005). *Web service composition using answer set programming*. Paper presented at the Workshop on Planning, Scheduling, and Configuration.

Rao, J., & Su, X. (2004). *A survey of automated Web service composition methods*. Paper presented at the SWSWPC.

Ricca, F. (2003). *The DLV Java wrapper*. Paper presented at the Italian Congress on Computational Logic (CILC).

Simons, P., Niemelä, N., & Soininen, T. (2002). Extending and implementing the stable model semantics. *Artificial Intelligence*, *138*(1-2), 181–234. doi:10.1016/S0004-3702(02)00187-X

Son, T., & Pontelli, E. (2007). *Planning for biochemical pathways: A case study of answer set planning in large planning problem instances*. Paper presented at the Software Engineering for Answer Set Programming.

Son, T. C., Baral, C., Tran, N., & McIlraith, S. (2006a). Domain-dependent knowledge in answer set planning. *ACM Transactions on Computational Logic*, *7*(4).

Son, T. C., Baral, C., Tran, N., & McIlraith, S. (2006b). Domain-dependent knowledge in answer set planning. *ACM Transactions on Computational Logic*, *7*(4), 613–657.

Son, T. C., & Pontelli, E. (2006). Planning with preferences using logic programming. *Theory and Practice of Logic Programming*, *6*, 559–607. doi:10.1017/S1471068406002717

Son, T. C., Tu, P. H., Gelfond, M., & Morales, R. (2005). Conformant planning for domains with constraints: A new approach. In *Proceedings of the the Twentieth National Conference on Artificial Intelligence* (pp. 1211-1216).

Spencer, B. (2002). *The design of j-DREW, a deductive reasoning engine for the Semantic Web*. Uni. Pol. Madrid.

Subrahmanian, V. S., & Zaniolo, C. (1995). Relating stable models and AI planning domains. In *Proceedings of the International Conference on Logic Programming* (pp. 233-247).

Syrjanen, T. (1998). *Lparse 1.0: User's manual*. Helsinki University of Technology.

The Policy RuleML Technical Group. (2004). *RuleML-powered policy specification and interchange*. Author.

Tu, P. H., Son, T. C., & Baral, C. (2006). Reasoning and planning with sensing actions, incomplete information, and static causal laws using logic programming. *Theory and Practice of Logic Programming*, *7*, 1–74.

van Emden, M., & Kowalski, R. (1976). The semantics of predicate logic as a programming language. *Journal of the ACM*, *23*(4), 733–742. doi:10.1145/321978.321991

Van Gelder, A., Ross, K., & Schlipf, J. (1991). The well-founded semantics for general logic programs. *Journal of the ACM*, *38*(3), 620–650. doi:10.1145/116825.116838

Wu, D., Parsia, B., Sirin, E., Hendler, J., & Nau, D. (2003). *Automating DAML-S Web services composition using SHOP2*. Paper presented at the The SemanticWeb – ISWC.

ENDNOTES

1. For simplicity, we omit the **binding** and **service** elements and the XML namespace definitions.
2. www.ruleml.org
3. We use the notation $\theta\sigma$ to denote the composition of the two substitutions; we also denote with $A\theta$ to denote the application of the substitution θ to the entity A.

This work was previously published in Managing Web Services Quality: Measuring Outcomes and Effectiveness, edited by K. Khan, pp. 193-221, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 2.7
The Effectiveness of Scaffolding in a Web-Based, Adaptive Learning System

Mei-Yu Chang
National Hsinchu University of Education, Taiwan

Wernhuar Tarng
National Hsinchu University of Education, Taiwan

Fu-Yu Shin
Chien-Kuo Elementary School, Taiwan

ABSTRACT

This study combined ideas from learning hierarchy and scaffolding theory to design a web-based, adaptive learning system to investigate the effectiveness of scaffolding for elementary school students having different levels of learning achievement. The topic chosen for learning was the Three States of Water. A quasi-experiment was conducted. In this experiment, students were divided into three groups: control group (without scaffolds), experimental group A (scaffolds providing by on-line conversation) and experimental group B (scaffolds providing by face-to-face conversation). The experimental results showed significant improvement for students after they had studied using the web-based, adaptive learning system. Specifically, scaffolds in the form of face-to-face conversations greatly enhanced the learning of high-achievement students. However, there were no significant differences between the low-achievement students with or without the provision of scaffolds. It was also discovered that the web-based, adaptive learning system could help students develop their learning responsibility.

INTRODUCTION

The purpose of science education is to provide students with scientific knowledge, concepts, attitudes and methods for application in their daily lives. The role of a teacher is to assist students in learning and solving problems. At the beginning, students are usually interested in learning science, but to some extent, they lose it or become confused, especially when learning abstract concepts. The Three States of Water is a learning unit containing abstract and complicated concepts for elementary students.

Use of the Internet and instructional technology can help teachers and students in many ways. It is easier for students to understand abstract concepts if their learning process is assisted by instructional technology. This study combined with ideas of learning hierarchy and scaffolding theory to the design of a web-based, adaptive learning system to improve the quality of web-based learning.

The major purpose of this study was to investigate students' learning via a web-based, adaptive learning system where scaffolds were provided to help the students to study the concept of the Three States of Water. This study also tried to find out whether students' learning responsibility was being developed during their learning process.

The concept of *scaffolding* is based on Vygotsky's social constructivist view of learning (1978). Vygotsky proposed that there were two major factors, i.e., culture and social context, which influence learning. He claimed that every mental function in a child's development first came from the social interaction with an adult. This kind of interaction provides a supportive environment in which children can extend their current knowledge and skills. The supported situation occurs in what Vygotsky referred to as the *zone of proximal development*. That is the area between what children can do independently and what they can do with assistance, such as they get from teachers and other students. The assistance that other people provide is a *scaffold* for a child. Given repeated experiences, children can internalize the supported situation of the mental processes, and can engage in them in new contexts (Clark & Graves, 2005).

Two sources of knowledge are suggested by Vygotsky. The first is *everyday knowledge*, i.e., gut knowledge, instinctive knowledge and spontaneous knowledge. This type of knowledge is influenced by peer interaction, language and experience of the individual who tries to understand his or her environment. The second is from formal education in the classroom, which is called *formal knowledge* and it possesses strict logic and clear definitions. Therefore, learners construct meaningful knowledge through both their daily lives and school experiences.

However, some scientific concepts are very abstract and difficult for students to grasp. In addition, they may often be influenced by different cultural and social environments. For example, the concept of "The sun rises in the east and falls in the west" generally exists in textbooks and daily communication, and students have this *misconception* before they are educated with *formal* knowledge. Students may think the experiences they have in daily life constitute their full knowledge base. Therefore, it is important for them to know if the knowledge they have received is formally right or not.

According to scaffolding theory, teachers should hold continuous and active conversations with students to find out the possible levels of their potential development (where they are) and to control their learning environments (where they should be) by providing proper support to make the concepts they acquire consistent with scientists' current definitions.

Empirical Studies on Related Topics

Osbome and Cosgrove (1983) discovered in their study that students lacked the support of substantial scientific concepts when they explained changes in the states of water. The finding was that they had only superficial knowledge of the terms and expressions. For example, some students would think that a solid changing into a liquid will lose weight or condensation will make particles more compacted. Bar and Travis (1991) investigated the concepts of liquid and gas as possessed by Israeli children and they found that most children had misconceptions about evaporation and condensation, suggesting that it was more difficult for the children to understand these abstract concepts.

The curriculum for science and technology in elementary schools in Taiwan includes the learning about substances and energy, natural environments, ecological conservation and information technology. The learning unit chosen in this study was focused on the important concepts of the Three States of Water and changes such as solidification, melting, condensation and evaporation, because it is related directly or indirectly to most of the subjects mentioned above. Although most of the phenomena can widely be seen in our daily lives, students may still have misconceptions due to their own mis-understanding and mis-reasoning.

After reviewing related research conducted in this area, the authors found that most children do have misconceptions about the states of water. Only a few of them think that vapor can also become water at a low temperature. Very few of them discover that wind blowing can make water evaporate more easily. Senior students in elementary schools know much about evaporation, but they know little about condensation as a formal scientific concept. Many children think condensation is solidification.

In general, condensation is a more difficult concept than evaporation for students to understand. Most students do not know the white smoke above boiling water is a gas or vapor; they think of it as smoke. Again, some students believe the weight of water changes after it becomes ice. This study therefore designed diagnostic tools and adaptive learning mechanisms to correct the students' misconceptions that often occur.

Current Status of Adaptive Learning

Electronic Learning (or e-Learning), sometimes provided as online learning or web-based learning, means learning through digital content to reach the goal of learning anytime and anywhere. In general, e-Learning includes any means of delivering teaching materials through networks (Whittington, 2000). Since 1990, the Internet has developed rapidly. After the establishment of the World Wide Web (WWW), many educational websites emerged (Berge & Collins, 1995; Cahoon, 1998; Collis, 1996; Porter, 1997).

Digital content may consist of multimedia as well as online interactions, allowing learners to play a more active role in the process of learning (Aivazidis, Lazaridou & Hellden, 2006). Learners' locations are no longer limited to classrooms, and they can learn anytime they want. In addition to a variety of teaching materials, e-Learning has gradually become the most convenient way for students to access new information and knowledge.

Since traditional ways of learning can not meeting the needs of individual learners, an increasing number of researches in adaptive learning systems have been conducted in recent years. An adaptive learning system can provide suitable contents for different learners according to their backgrounds, prior knowledge, individual demands and learning statuses (Papanikolaou *et al.* 2002).

Atkinson (1976) suggested that an efficient instructional strategy must be adaptive. Therefore, an adaptive learning system must change instructional strategy according to learners' situations during a learning process under teacher's control to enhance learning and achieve the expected instructional objectives. Besides, the system can collect some features of learners and store them in certain learning modules, which can be used to provide suitable contents for different learners (Brusilovsky,1996).

Web-based learning has become more and more popular today. A web-based learning system is accessible by a large number of students at the same time according to individual necessity and learning pace, and it is not limited by time or space. Therefore, it is suitable for the development of adaptive learning environments (Chang, 2005). According to Lin's study (1998), an adaptive learning system can achieve the goal of teaching students according to their aptitudes in a normal class grouping environment, and it makes learning more active and efficient.

A personalized web-based learning system was proposed by Chang *et al.* (2006) based on item response theory (IRT). The system was aimed at providing a suitable learning environment by considering the learning portfolios of content difficulty, learner's ability and conceptual coherence. Lin and Kuo (2005) developed a virtual learning environment based on the theory of learning objects and constructivism. The learning contents and statuses of learners on the system can be individualized to achieve the goal of adaptive learning. Also, learners can cooperate in the inhabited virtual world to increase learning effectiveness during their learning processes.

Because e-Learning systems are usually developed in accordance with *learners*' requirement in terms of platforms, materials, presentation styles and virtual communities, it often occurs that some systems are only suitable for certain users. If a system is transferred to another environment, for some other people, tremendous manpower and costs must be spent in making changes (Dodds, 2001a; 2001b). To solve this problem, international standard organizations have specified e-Learning standards. E-Learning systems such as SCORM, AICC, IEEE, IMS, IEEELOM, ARIADNE, and Microsoft promulgate these standards (Hodgins & Conner, 2000). Among them, SCORM has now become an important standard for e-Learning content (Chang, Hsu, Smith & Wang, 2004).

Learning contents that conform to SCORM standards have reusable and sharable features. However, there are still some disadvantages, such as complicated definitions for the rules of learning activity and huge frameworks for learning contents, making it difficult to manage, reuse and combine these with traditional educational theories. E-Learning does need a good system to support online services, such as course design, learning resource delivery, integration of courses and teaching methods (Suthers, Johnson & Tillinghast, 2002), etc. For this reason, the Graduate School of Online Learning Applications, National Chiao Tung University, Hsinchu, Taiwan developed their Object Oriented Learning Activity System (OOLAS) based on the standards of SCORM, Learning Design Specification, Knowledge Tree and Multibook. In the SCORM standards, most metadata models were used to define learning materials (Su, Tseng, Lin & Lu, 2005). Chang *et al.* proposed in their study (2004) a metadata model to define online assessment. OOLAS is composed of an authoring system and a learning system. In addition to a simple process for editing materials, it presents the contents of curricula easily and allows teachers to plan related programs to assist students with learning effectively. OOLAS also allows teachers to design the related tests of concepts.

In this study, the authors used OOLAS to develop a web-based, adaptive learning system that was integrated with scaffolding to improve the learning efficiency of students. They also developed teaching materials for the concepts of the Three States of Water. The system modules were: diagnostic tools, a test-item database, adaptive learning modules and learning contents for the scientific concepts in the Three States of Water (Figure 1).

The diagnostic tools were designed using two-tiered test items to identify the prior misconceptions of students. After detecting their misconceptions, an adaptive learning module was initiated by a rule-based inference engine to generate individualized learning content for students according to the particular misconceptions of the particular student.

RESEARCH DESIGN

This study was conducted using a quasi-experimental design. A total of 200 students belonging to two elementary schools in Taipei County were chosen for the experiment. The authors divided the students into a control group and experimental groups to investigate the learning of students with or without scaffolds. The authors also measured the changes in their learning responsibility after the e-Learning. For the control group, the students studied using the web-based, adaptive learning system by themselves without the provision of any scaffolds. For the experimental groups, the students could ask teachers for help if they had questions during the learning processes and after the diagnostic tests. There were two experimental groups, provided with scaffolds in different ways. For experimental group A, the scaffolds were provided in the form of online conversation between the teachers and students. For experimental group B, the scaffolds were provided in the form of face-to-face conversations between the teachers and students. For both media of communication, teachers provided scaffolds by discussing with the students to find out their problems and correct their misconceptions.

Research Procedure

This study selected the Three States of Water as the topic of learning. Therefore, the authors first analyzed the related concepts covered by textbooks of elementary schools, and then developed a learning hierarchy according to these concepts. Based on the learning hierarchy, the authors used the diagnostic tools of the web-based, adaptive learning system to find out the misconceptions of students. The authors developed the adaptive learning modules by using a rule-based inference engine to provide remedial instructions for students according to a categorization of the types of their misconceptions. After that, the authors conducted a quasi-experiment to find out the influence of scaffolding on students when they were learning using the system.

Research Samples

The samples in this study were 5th and 6th grade students, randomly selected from two elementary schools in Taipei County. Because these students had already taken computer courses during their 3rd grade year, by the time of this experiment, they had two to three years of experience in using computers and the Internet. This reduced the influence of computer skills on the effectiveness of e-Learning. As stated, the students were divided into the control group (without scaffolds) and the experimental groups (with scaffolds). The experimental groups consisted of group A (providing scaffolds via online conversation) and group B (providing scaffolds via face-to-face conversation).

According to a pretest, students with scores above the first 50% were classified as high achievement students while the remaining 50% were classified as low achievement students. In the control group, there were 33 high-achievement students and 33 low-achievement students, making

Figure 1. The operation of the web-based, adaptive learning system

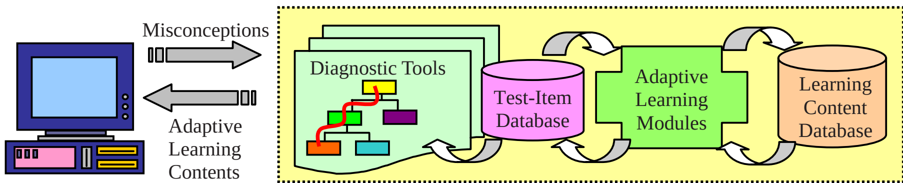


Table 1. The number of students in each group

Content	Groups	Scaffolds	Achievement	Number of Students
Three States of Water	Control Group	None	High	33
			Low	33
	Experimental Group A	Online Conversation	High	33
			Low	32
	Experimental Group B	Face-to-Face Conversation	High	35
			Low	34
Total Number of Students				200

a total of 66 students. In the experimental group A, there were 33 high-achievement students and 32 low-achievement students, making a total of 65 students. In the experimental group B, there were 35 high-achievement students and 34 low-achievement students, making a total of 69 students (as listed in Table 1).

Diagnostic Tools

The questions to diagnose misconceptions about the Three States of Water in this study were modified from the paper-based test questions designed by Chang (2003), who analyzed the misconceptions that students in elementary schools often have. Since two-tier test items have been used widely and efficiently in discovering misconceptions and conceptual changes (Tsai & Chou, 2002), the authors thus used them to develop the diagnostic tools based on the empirical study of Lin (1995). The authors consulted three science teachers and two experts in science education about the validity of the test items and diagnostic tools. Finally, the two-tier test items were uploaded and stored in the test-item bank by using authoring tools of the system. An example of the two-tiered question follows:

There are two identical wet towels. One is spread and the other is folded. Which one do you think will dry first?

Answer: ____(A) Folded one. (B) Both at the same time. (C) Spread one. (D) I don't know. Explain your reason:.

Pretest and Posttest

Limited by the sampling method and number of samples, we adopted a non-equivalent group pretest-posttest design in our experiment, and adjusted the difference in students' background using an analysis of covariance (ANCOVA) to investigate the influence of scaffolding on the students with different levels of achievement. The questions in the pretest and posttest were designed based on the learning contents and related concepts of the Three States of Water. The reliability of the achievement test was verified by an internal-consistency reliability test with Cronbach α=0.71.

After removing the inappropriate questions, there were 113 multiple-choice questions left. Among them, 37 questions were selected to form the pretest and posttest, and the remaining questions were used as diagnostic test items. The scores of the pretest were used to measure the students' levels of achievement. The questions in the posttest were similar to that of the pretest, enabling the authors to compare the experimental and control group results. An example of a question follows:

Which of the following is the correct description for the speed of evaporation? (A) The more humid the environment is, the faster the evaporation is. (B) The hotter the environment is, the faster the evaporation is. (C) The more airtight the environment is, the faster the evaporation is.

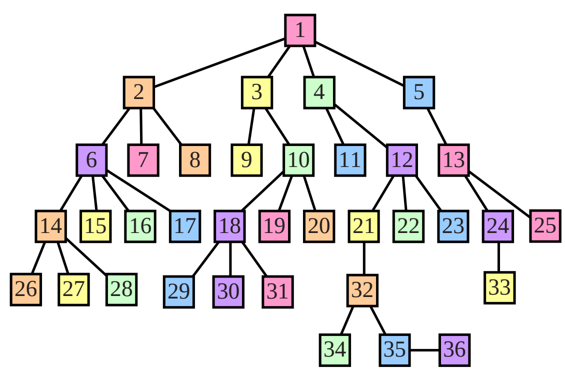
Measurement of Learning Responsibility

In this study, the authors used a questionnaire to measure the development of self-learning responsibility after the web-based learning. The questionnaire was developed by the researchers and verified by three elementary school teachers and two experts in the research area of science education. It contained 18 questions. Each question had five scales in its answer: 5 points for very likely, 4 points for likely, 3 points for no opinion, 2 points for not likely, and 1 point for not likely at all (as shown in Appendix I). This is an indicator for the degree of development of the students' learning responsibility. All questions in this measurement were positive-response designed. Thus, the more points a student gained, the more responsible and active he or she was in learning. The reliability of this measurement was verified by testing on 32 6th grade students and the total internal consistency was calculated as Cronbach α=0.89.

INSTRUCTIONAL DESIGN

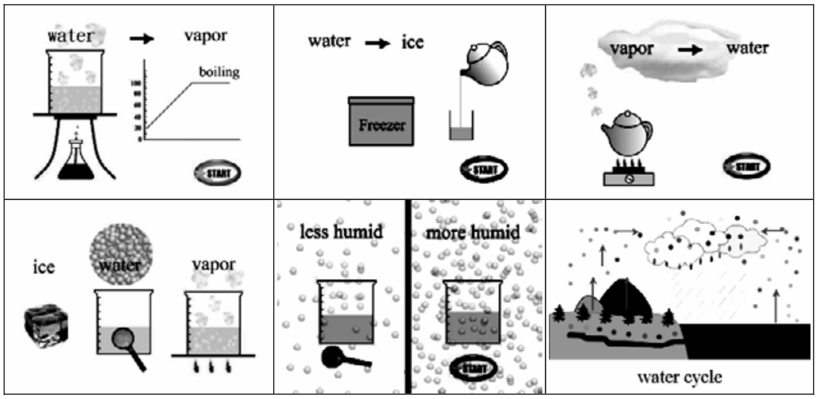
The principle of instructional design in this study was based on Gagne's learning hierarchy (1985). The major concepts of the Three States of Water were analyzed to form a hierarchical diagram. The learning hierarchy was constructed according to competence indicators for the curriculum of science and technology in elementary schools in Taiwan. For example, the concepts of the Three States of Water are based on solidification, melting, evaporation and condensation. Deeper, the concept of solidification contains the lower-level concepts of water changing from liquid to solid, increasing volume, and mass conservation. Because condensation, solidification, melting and evaporation are the major concepts for the Three

Figure 2. Hierarchy of scientific concepts in the three states of water



1. three states of water 2. solidification 3. melting 4. evaporation 5. condensation 6. liquid to solid 7. mass conservation 8. increasing volume 9. mass conservation 10. solid to liquid 11. liquid to gas	12. evaporation speed 13. gas to liquid 14. liquid water 15. solid water 16. decreasing temperature 17. freezing point 18. liquid water 19. solid water 20. increasing temperature	21. temperature 22. convection 23. relative temperature 24. Gaseous water is invisible. 25. decreasing temperature 26. achromatic 27. tasteless 28. without fixed shape 29. achromatic	30. tasteless 31. without fixed shape 32. evaporation absorbing heat 33. steam in the air 34. evaporation at room temperature 35. Heating speeds up evaporation. 36. Heating causes water to evaporate.

Figure 3. The online experiments incorporated in the learning content



States of Water in elementary science courses, a number of teaching activities were designed to introduce these concepts.

The authors encoded all concepts in the learning hierarchy by number to simplify the diagnostic sequences. A student with a misconception at a high level implies that he or she may have misconceptions at lower levels and thus requires further diagnostic processes. Generally, the higher the hierarchical level is, the smaller the number is. For example, the highest-level concept is “three states of water”, which was encoded as 1. The second-level concepts are “solidification”, “melting”, “evaporation”, “condensation”, and which were encoded as 2, 3, 4, and 5. The concept of “solidification” (2) contains the lower-level concepts of water changing from “liquid to solid” (6), “mass conservation” (7), and “increasing volume” (8) as shown in Figure 2.

Because experiments and observation are very important for the establishment of scientific concepts, the authors designed several online experiments to simulate the phenomena of the Three States of Water to help students develop their scientific concepts (Figure 3). The experiments were designed using Flash to provide an interactive user interface, animations for the simulation of the phenomena and concepts in the Three States of Water.

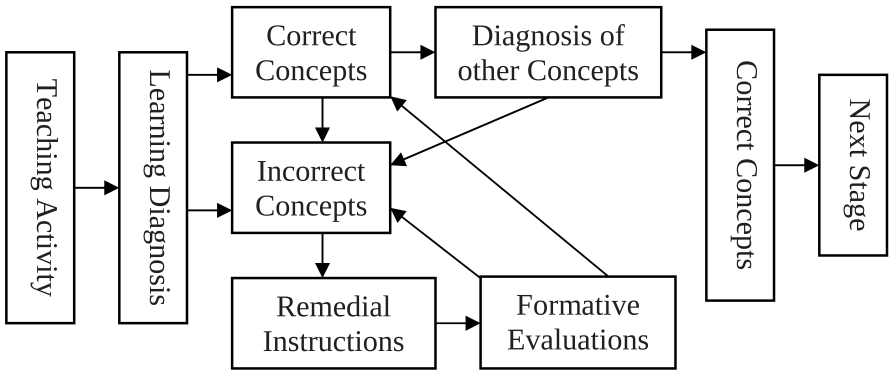
In this study, there were four lessons for all students to complete. After each one, the students did a diagnostic test to see if they had completely understood the scientific concepts. The design of diagnostic tests followed the sequence of solidification, melting, evaporation and condensation. After students had completed a learning activity, they were directed to the diagnostic tests to see if their understanding about the related concepts was correct. If the concepts of students were valid according to the diagnostic results, they could keep on learning the remaining concepts. Otherwise, the students had to proceed with remedial instruction to correct their misconceptions.

For example, if students passed the tests on the concept of solidification, they could move to the next stage to start the learning activities on the concept of condensation. Otherwise, they were directed to the remedial instructions for the correction of previous misconceptions, followed by formative evaluations to guarantee the conceptual changes had been internalized. When all misconceptions were corrected after remedial learning activities, they could start learning the contents for the next stage (Figure 4).

RESULTS AND DISCUSSIONS

The major purpose of this experiment was to investigate students’ learning enhancement on the web-based, adaptive learning system and the

Figure 4. Learning diagnosis design of the system



development of learning responsibility after the e-Learning. In addition, the learning statuses of students, such as misconceptions and conceptual changes, were also studied by analyzing learning records. Students' scores in the pretest and posttest were analyzed using statistical software SPSS to calculate the mean and standard deviation for each group.

Learning Effectiveness

Based on the analysis of the T-test, all students performed very well on the web-based, adaptive learning system, no matter whether they were high-achievement (t=7.27, p<0.01) or low-achievement (t=13.63, p<0.01) students. This revealed that the system was efficient for students in learning the concepts in the Three States of Water.

Based on the analysis of distribution in students' misconceptions, over 70% of the students had valid concepts about melting before learning activities. However, there were also 65% of the students with misconceptions about condensation so that they spent more time in the related remedial instruction. According to the results of the T-test, we found the students made significant progress in the areas of solidification (t=2.37, p<0.05), evaporation (t=2.98, p<0.05), and condensation (t=2.57, p<0.05). Besides, the results showed that more students understood the concept of evaporation than that of condensation, which agreed with the results of the studies by Lin (1995), Lai (1994) and Chang (1997).

Because the teaching materials were designed based on Gagne's notion of learning hierarchy (1985), it was effective for learning the concepts in the Three States of Water. Also, the web-based, adaptive learning system was useful for teachers in reducing the workload, while increasing the learning effectiveness of the students. According to the diagnostic results, students could go immediately through remedial instructions to correct their misconceptions. It could also solve the problem of insufficient teachers to provide feedback for individual students. Besides, through the records of learning processes, it was easy to know the status of each student and provide appropriate support.

High-Achievement Students

The effective samples of the high-achievement students contained 29 students in the control group, 31 students in the experimental group A, and 32 students in the experimental group B. Using ANCOVA, the homogeneity of the regression slope was tested (f=0.32, p>0.05), indicating that the homogeneity assumption was met. After the ANCOVA analysis, the results showed that there were significant differences among these three

Table 2. ANCOVA summary data of the high-achievement students

Source	S.S.	DF	MS	*F*	Sig.
Pretest	76.73	1	76.73	12.04**	0.01
Deviation	105.14	2	52.57	8.25**	0.01
Error	560.68	88	6.37		
Sum	75114.00	92			

***p<0.01*

Table 3. Post-hoc analysis on the posttest of high-achievement students

Comparison			Mean Difference	Standard Error
Experimental Group B	vs.	Experimental Group A	2.59	0.65
Experimental Group B	vs.	Control Group	1.46	0.64
Experimental Group A	vs.	Control Group	-1.12	0.67

groups. Table 2 shows the provision of scaffolds had a significant impact on the learning effectiveness of high-achievement students (f=12.04, p<0.01). It also reveals that the high-achievement students had different learning effectiveness in the three groups with deviation in the provision of scaffolds (f=8.25, p<0.01).

The post-hoc analysis revealed that the high-achievement students in the experimental group B had better learning effectiveness than those in the control group and experimental group A. In addition, Table 3 shows no significant difference between the high-achievement students in the control group and experimental group A.

According to the above results, it can be discovered that providing the scaffold of face-to-face interactions between teachers and students was better than that of online interactions for the high-achievement students

For the students in experimental group A, the support of teachers did not assist them too much because the scaffold provided was in the form of online conversation. It was not so effective because the teachers might not understand the student s' difficulties during the processes of web-based learning, especially when the students had some misconceptions. However, in the face-to-face conversation, the teachers could understand the students' misconceptions and then decide what kind of support to provide. That is why the learning effectiveness was significantly increased.

Low-Achievement Students

The effective samples of the low-achievement students consisted of 28 students in the control group, 31 students in the experimental group A, and 31 students in the experimental group B. Using ANCOVA, the homogeneity of the regression slope was tested (f=0.27, p>0.05), indicating that the homogeneity assumption was met. Through the analysis of ANCOVA, the authors were able to investigate the influences of different approaches on low-achievement students. The deviation F value caused by the independent variables (scaffold-provision difference) was 0.51 (p>0.05), which did not reach the degree of significance. The results in Table 4 show that the learning of scientific concepts by low-achievement students was not affected by the provision of scaffolds.

From the above results, it can be implied that the learning effect of low-achievement students was not influenced by the provision of scaffolds in e-Learning processes, meaning that no matter whether scaffolds were provided or not and regardless of the forms of scaffolds, there was no significant difference in the learning effectiveness of low-achievement students.

According to the above results, it is better for teachers to provide the high-achievement students with face-to-face conversations instead of online discussions. However, providing either scaffold could not help low-achievement students improve further because they all made a remarkable progress in the posttest, meaning that the web-based, adaptive learning system was very effective for low-achievement students and scaffolding was not a factor in influencing their learning.

Development of Learning Responsibility

Another objective of this study was to find out if the web-based, adaptive learning system could help students develop their responsibility for self-learning. There were 18 questions to measure the development of learning responsibility, and the score for each question was from 1 to 5 according to the agreement with the answer. A higher score meant more responsibility was established by the student, and vice versa.

The results showed that both levels of students (t=2.82, p<0.01), no matter whether they were high-achievement (t=2.93, p<0.01) or low-achievement (t=2.71, p<0.01), reached statistical significance in the comparison of the pretest and posttest. More precisely, the students raised their sense of self-responsibility as a result of e-Learning.

E-Learning is different from traditional ways of learning. Given well-organized teaching materials and learning activities, students were more aware of their learning conditions and therefore could do better by themselves to obtain knowledge. Therefore, the e-Learning process was helpful for the development of students' learning responsibility. They could follow the directions to study by themselves and make their schedules to complete the learning activities. For both types of students, the difference was that the low-achievement students preferred to find the answers by themselves, while the high-achievement students wanted to ask teachers for help when they had problems. In both ways, they performed well on the web-based, adaptive learning system.

After the experiment, the authors conducted some semi-structural interviews with a few students to understand their ideas about the web-based, adaptive learning system and its performance. The interview results are summarized as:

I like the way of web-based learning, because I can also do it by myself at home.

I can stop somewhere and repeat studying for several times until I can fully understand the concepts. However, I cannot do it that way in classrooms because I am afraid to ask teachers to slow down their teaching.

Web-based learning is more efficient in raising my learning motivation than traditional classroom learning does.

Table 4. ANCOVA summary data of the low-achievement students

Source	S.S.	DF	MS	*F*	Sig.
Pretest	299.26	1	299.26	24.84	0.01
Deviation	12.33	2	6.17	0.51	0.60
Error	1036.33	86	12.05		
Sum	65606.00	90			

The web-based, adaptive learning system can point out my misconceptions and help me correct them. I can operate the system easily, so I am not afraid of it.

I can learn the concepts of the three states of water efficiently on the system.

Learning on the system is very interesting. I wish teachers in other classes (could) also use the same way in their teaching.

The authors also asked the teachers who were involved in this experiment for their opinions, and they were satisfied with the system because it provided a convenient interface for teachers to design the teaching activities and edit the learning contents. Besides, the system could help teachers identify the misconceptions of students and direct them to suitable learning content to enhance their learning effectiveness. Therefore, it could save teachers' time and effort in providing remedial instructions for individual students.

CONCLUSION

In this study, a web-based, adaptive learning system was designed to investigate the effectiveness of scaffolding on elementary school students with different levels of learning achievement. This study was conducted using a quasi-experimental design and the contents for learning were the scientific concepts of the Three States of Water. According to the experimental results, the findings are:

The web-based, adaptive learning system is effective in helping students acquire the concepts of the Three States of Water. For the high-achievement students, providing scaffolds in the form of face-to-face conversation could significantly enhance the learning effectiveness, which is better than using online interactions or without provision of scaffolds. However, there was no significant difference for the low-achievement students with or without the provision of scaffolds, since they all performed well on the web-based, adaptive learning system. For both types of students, their learning responsibility was significantly improved in the web-based learning environment.

The design of the web-based, adaptive learning system in this study was based on the notion of scaffolding theory and Gagne's learning hierarchy. Because the system provided diagnostic mechanisms in the entire learning processes, it was assumed that the students passing the diagnostic tests would not have any misconceptions. However, the results in posttest showed that some students still had misconceptions about the material. Therefore, teachers should not rely on the diagnostic tools completely. They should use some other assessment tools to see if the students' concepts were correct or not, and then provide appropriate feedback to correct the misconceptions.

If the use of teaching materials and diagnostic tools on the system can be easily transferred to other systems, it will increase the sharing among schools to help more students learn effectively. Moreover, advances in computer and network technologies make it easier for students to learn through networks, and this reduces the workload of teachers in helping students solve their problems. This study was quantitatively oriented so that it might not be enough to discover the details of students' learning statuses and the process of conceptual changes. Future studies can be focused on qualitative methods to understand students' behavior while learning.

ACKNOWLEDGMENT

The authors would like to thank for the financial support of the National Science Council of the Republic of China under the contract number NSC 92-2524-S-134-006.

REFERENCES

Aivazidis, C., Lazaridou, M. & Hellden, G. F. (2006). A comparison between a traditional and an online environmental educational program. *Journal of Environmental Education*, 37, 45-54.

Atkinson, R. C. (1976). *Adaptive instructional systems: some attempts to optimize the learning process.* In D. Klahr (Ed.), *Cognition and instruction*. New York: Wiley.

Bar, V. & Travis, A. S. (1991). Children's views concerning phase changes. *Journal of Research in Science Education*, 18, 363-382.

Berge, Z. & Collins, M. P. (Eds.) (1995). *Computer-mediated communications and the online classroom, Vol. III: Distance Learning.* Cresskill, New Jersey: Hampton Press.

Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. *User Modeling and User Adapted Interaction*, 6(2-3), 87-129.

Cahoon, B. (1998). *New directions for adult and continuing education, Summer, No. 78. Adult Learning and the Internet*. San Francisco: Jossey-Bass.

Chang, C. (1997). A study on the concepts of evaporation, condensation and boiling by senior students in elementary schools. *Chinese Journal of Science Education*, 5(3), 321-346.

Chang, K. (2003). *A research on the concepts of states of water.* Master thesis, National Pingtung Teachers College, Pingtung, Taiwan.

Chang, S. L.(2005). *Applying characteristics described system to diagnosis of e-learning.* Master thesis, Department of Information and Computer Engineering, Chung Yuan Christian University, Chung-Li, Taiwan.

Chang, W., Hsu, H., Smith, T. K. & Wang, C. (2004). Enhancing SCORM metadata for assessment authoring in e-Learning. *Journal of Computer Assisted Learning*, 20, 305-316.

Clark, K. F. & Graves, M. F. (2005). Scaffolding students' comprehension of text. *The Reading Teacher*, 58(6), 570-580.

Chen, C. M, Liu, C. Y. & Chang, M. H. (2006). Personalized curriculum sequencing utilizing modified item response theory for web-based instruction. *Expert Systems with Applications*, 30(2), 378-396.

Collis, B. (1996). *Tele-learning in digital world: The future of distance learning.* New York: International Thomson Computer Press.

Dodds, P. (2001a). *Sharable Content Object Reference Model (SCORM) Version 1.2 Technical Report*. The SCORM content aggregation Model, Advanced Distributed Learning Initiative, URL: http://www.adlnet.org/index.cfm?fuseaction=SCOR-Down [last accessed 20 November 2003]

Dodds, P. (2001b). *Sharable Content Object Reference Model (SCORM) Version 1.2 Technical Report*. The SCORM Run-Time Environment, Advanced Distributed Learning Initiative, URL: http://www.adlnet.org/index.cfm?fuseaction=SCOR-Down [last accessed 20 November 2003]

Gagne, R. M. (1985). *The conditions of learning and theory of instruction*. New York: Holt, Rinehart, & Winston.

Hodgins, W. & Conner, M. (2000). *Everything you ever wanted to know about learning standards but were afraid to ask*. LiNE Zine, URL:

http://www.adlnet.org/index.cfm [last accessed on 20 November 2003],

http://www.adlnet.org/index.cfm [last accessed on 20 November 2003], and http://www.linezine.com/2.1/features/wheyewtkls.htm [last accessed on 9 Feb. 2004].

Lai, H. C. (1994). *Children's concepts of water cycle and the forms of substances.* Master thesis, National Taiwan Normal University, Taipei, Taiwan.

Lin, C. H. (1998). The study of individual differences and automated learning strategies in hypertext learning environment. *Journal of National Hsinchu Teachers College*, 11, 1-14.

Lin, C. S. & Kuo, M. S. (2005, July 5-8). Adaptive networked learning environments using learning objects, learner profiles and inhabited virtual learning worlds. *The 5th IEEE International Conference on Advanced Learning Technologies*, Kaohsiung, Taiwan.

Lin, S. (1995). *A study on students' concepts of evaporation and condensation in elementary schools.* Technical report, National Science Council, Taipei, Taiwan.

Osbome, R. J. & Cosgrove, M. M. (1983). Children's conception of changes of state of water. *Journal of Research in Science Teaching*, 27, 825-838.

Papanikolaou, K., Grigoriadou, M., Magoulas, G. D. & Kornilakis, H. (2002). Towards new forms of knowledge communication: the adaptive dimension of a web-based learning environment. *Computers and Education*, 39, 333-360.

Porter, L. R. (1997). *Creating the virtual classroom: Distance learning with the Internet*. New York: John Wiley.

Su, J. M., Tseng, S. S., Lin, H. Y. & Lu, C. Y. (2005, May 6-7). The development of an object-oriented learning activities system. *Taiwan e-Learning Forum 2005*, Taipei, Taiwan.

Suthers, D., Johnson, S. & Tillinghast, B. (2002). Learning object meta-data for a database of elementary and secondary school resources. *Interactive Learning Environments*. 9, 273-289.

Tsai, C. & Chou, C. (2002). Diagnosing students' alternative conceptions in science through a networked two-tier test system. *International Journal of Computer Assisted Learning*, 18 (2), 157-165.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, Massachusetts: Harvard University Press.

Whittington, D. (2000). Evaluating three years' use of virtual university. *Quality Assurance in Education*, 8(1), 48-52.

APPENDIX: A QUESTIONNAIRE TO MEASURE THE DEVELOPMENT OF LEARNING RESPONSIBILITY

Dear students, The following is a questionnaire to measure the establishment of learning responsibility after web-based, adaptive learning. It is not a test and there are no standard answers. Please read the following questions carefully, and put a check mark inside the space to represent your opinion regarding each question. Grade: □5th Grade □6th Grade — Sex: □Boy □Girl		1. Very likely	2. Likely	3. No opinion	4. Not likely	5. Not likely at all
1.	In the process of learning, I often finish all learning activities by myself.					
2.	In the process of learning, I know clearly whether or not I have understood all learning contents taught in class.					
3.	In the process of learning, I would look for the answers actively to prove that my ideas are correct.					
4.	In the process of learning, I would join the discussion among my classmates actively.					
5.	When discussing with my classmates, I could judge whose ideas are correct.					
6.	I consider that learning is my own responsibility no matter if the result is good or bad.					
7.	In the process of learning, I usually doubt if my own ideas are correct or not.					
8.	If I have difficulty in learning, I will ask for help actively when my teacher does not assist me.					
9.	In the process of learning, I would check on myself to see if my ideas are correct.					
10.	I know clearly how to learn in class.					
11.	If I have difficulty in learning, I will ask for help actively when my classmates do not assist me.					
12.	In the process of learning, I would often request myself to learn more conscientiously.					
13.	During the in-class discussion, I often doubt if the ideas of teachers or classmates are correct or not.					
14.	When I have doubts about the ideas of teachers or classmates, I would search for the answers to prove if they are correct or not.					
15.	I would rather find the answers to my questions than have my teachers or classmates help me.					
16.	After school, I do a self test to see if I have learned whatever was taught in class.					
17.	I think I have to learn actively and positively in order to understand the materials taught in class.					
18.	I make my own learning schedule and follow it accurately.					

This work was previously published in International Journal of Web-Based Learning and Teaching Technologies, Vol. 14, Issue 1, edited by E.M.W Ng, N. Karacapilidis, and M.S. Raisinghani, pp. 1-15, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 2.8
WebSphere Portal 6.1: An Agile Development Approach

Thomas Stober
IBM Germany Research and Development, Germany

Uwe Hansmann
IBM Germany Research and Development, Germany

ABSTRACT

IBM's Portal technology continues to evolve as a powerful infrastructure for integrating the IT landscape, by presenting it as a consolidated view to the user community. The new capabilities of WebSphere Portal 6.1 are the outcome of a world-wide development team, which focused on this release for the past 2 years. During that time major architectural enhancements have been introduced and a significant amount of code was written. In this article the authors will describe how developers and testers have adopted agile principles to collaborate across the globe. In detail, aspects like an iterative approach, test driven development, budget based prioritization and cross-organization teaming will be discussed. The authors will also cover how "tiger teams" interact with customers by making early code drops available and responding to feedback.

REACHING THE LIMITS OF THE CLASSICAL WATERFALL APPROACH

It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so. Mark Twain

Developing market-leading Enterprise Portal products, like WebSphere Portal requires a first-class development team. Far more than 300 developers and test engineers are working for different organizational units and collaborate in very far apart time zones. There are 8 major development sites across the world. The product has dependencies on other IBM products, such as WebSphere Application Server, and is the base for other products, like Lotus Quickr. Further dependencies arise from customer requirements and commitments.

IBM WebSphere Portal leads this wave of innovation, combining the latest user-centric functionality with reliable security and manageability features to meet the needs of the business. The software incorporates extensive Web 2.0 capabilities, allowing companies to fuel social interaction by delivering high-performing, intuitive applications through a rich Web interface. This new release adopts the latest industry-driven standards. It also introduces flexible ways to create and manage Portal sites and content. Many more enhancements emphasize increased utility and flexibility, such as web site management, integration of non-Portal pages as well as step up authentication.

Up to Portal 6.0, the Portal team used to work a classical waterfall approach. Product management captured the requirements and work items for a particular release project, and prioritized them. The project management team assembled a complex project plan with a break down of distinct task assignments for individual developers and testers. There were milestones, test phases, and fixed target dates to achieve the well defined goals. In general there was an analysis and design phase in which content, architecture, and project plan were established. In addition, there was distinct development phase executing the plan, followed by the distinct test phase. In major releases the project plan covered a period of up to 1.5 years.

However the complexity of the technology and especially the growing complexity of the team and time constraints have made it more and more difficult to execute the established plan as scheduled. Future needs and issues are difficult to predict. Each of these distinct phases turned out to be not that distinct and isolated. Instead there were dependencies, loops circulating back to earlier phases in order to adjust. Communication and interfaces between different organizational units are a challenge in large distributed teams. It is extremely difficult to make sure that the right information is made available to the right set of people. Bringing the independently developed pieces together in order to assemble a complex use case requires a significant integration effort, before the overall system reaches a satisfying level of stability. Development and testing were done by separate organizations. While the developers owned the responsibility for design, coding and unit testing, the test organization covered functional and system verification testing.

Typically, unforeseen issues, like a design flaw, or a growing number of bugs beyond the expected, or redirection of resources to other activities, made it necessary to rework the plan. Typically the problems are getting really pressing at the end in the final test phase. At that time, content removal isn't really an option, as the code is already done, although not stable. Delaying the shipment is not a good option, as customers do rely on the promised delivery dates. And sacrificing the quality is not acceptable either. And obviously the costs of fixing problems increase significantly the later the issue is detected. It took a tremendous, costly team effort to solve the situation and ensure that a solid product is still being shipped on time.

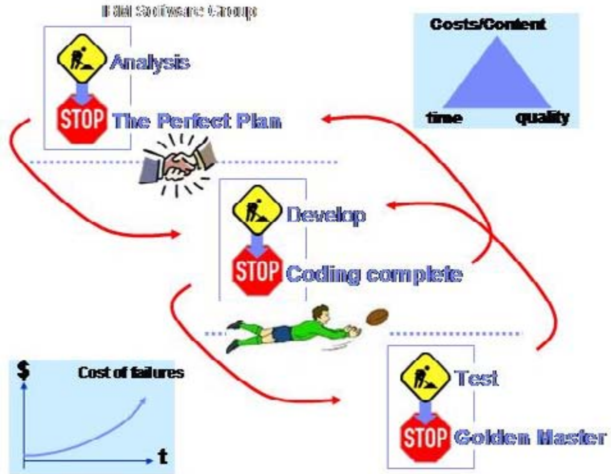
As a result of these experiences, the limits of such a pre-planned waterfall approach became obvious: the classical approach it is too inflexible to react quickly enough to the highly dynamic constraints of a complex product within a large organization (Figure 1).

MOVING TO AGILE SOFTWARE DEVELOPMENT

One of the key reasons, why the Portal team has moved away from a classical waterfall approach used for Portal 6.0, is to gain more flexibility and improve the ability to react to changing constraints. Within a release project, the content needs to be decided as late as possible, while tested and usable pieces of functionality are to be made available as early as feasible.

Another goal is to optimize the flow of human interaction. Intensive collaboration should be

Figure 1. Waterfall approach: it is getting tougher towards the end



fostered across organizational structure - especially between test and development teams and the number of dependencies between different teams should be minimized.

The agile approach, which the Portal team has introduced for Portal 6.1, features 4 key concepts:

- Budget based prioritization and content definition
- Cross-organization teaming Structure
- Iterative development approach
- Test driven development

BUDGET BASED PRIORITIZATION AND CONTENT DEFINITION: REQUIREMENTS AND USE CASES

It is not the strongest of the species that survives, nor the most intelligent; it is the one that is most adaptable to change. Charles Darwin

Which features should be included into a release? Who decides the release content with which level of granularity? For Portal, the product management and lead architects do come up with a first cut of release content by aggregating the customer feedback and high-level requirement into rough focus areas. Each of these focus areas is associated with a budget reflecting the approximate number of developers supposed to be working on that area throughout the upcoming release project. In Portal the designated developers pursuing the same focus area are grouped into a "*Tiger Team*". Each team can span multiple geographical locations and organizational units. The worldwide team acknowledges that the developers of each tiger team are the subject matter experts, and that the "center of gravity" is lowered to allow quicker, better and more optimized decisions directly by the tiger team. This distributed approach scales much better than a central release management deciding on all detailed use cases of the entire release and maintaining a complex overall project plan for all worldwide developers.

For example, it is the responsibility of the developers in each tiger team to translate the given high-level requirements into specific use cases. Each tiger team prioritizes the use cases it intends to deliver in a team charter document. The charter lists highly prioritized use cases, which the team is committing to deliver under all circumstances.

Other items have the disclaimer to be run-at, and will be implemented in the order of priority as time permits. Further items are listed for completeness, but are marked as out-of-scope for the release project. The final scope and timeline of the release project is not yet fully defined when team is being established. To leave sufficient room for agility, it is extremely important that each team is committing only use cases to a limited amount of its team capacity.

Within the constraints of staffing budget, overall release timeline and given high-level requirements, the team can proceed very autonomously with the planning and execution of their activities. Often tiger teams start their work independently of a specific release and drive towards early deliverables outside of an official product. One example is the tiger team, which developed the web 2.0 capabilities for Portal. The team has been inaugurated long before there where any planning activities for a release 6.1. Without knowing yet in which product release their code will eventually be shipped, they delivered early prototypes and presented them at the LotusSphere conference, on IBM's greenhouse site, and as well as part of the Portal 6.1. Beta. Most important, the received customer feedback has been used to improve the web 2.0 capability throughout the project.

Each team is continuously updating and adapting their charter document as well as any other planning document to reflect changing overall constraints, like modified release schedules, dependencies on other teams, or different high-level requirements. Each team's project plan is only specific and confirmed for the next few weeks (one "iteration"), while the rest of the project's duration is tentatively proposed, but subject to change.

The initial budget defined for a Tiger Team is only the first rough guess to help the Tiger Team with some guidelines. During the initial planning and the composition of the overall Release Plan, this budget may very well get adjusted to ensure that the focus items for the release get in with a solid staffing and the required skills. Usually the market is demanding certain features at a certain timeframe and even an agile plan has to commit some minimum of enhancements early up in the cycle.

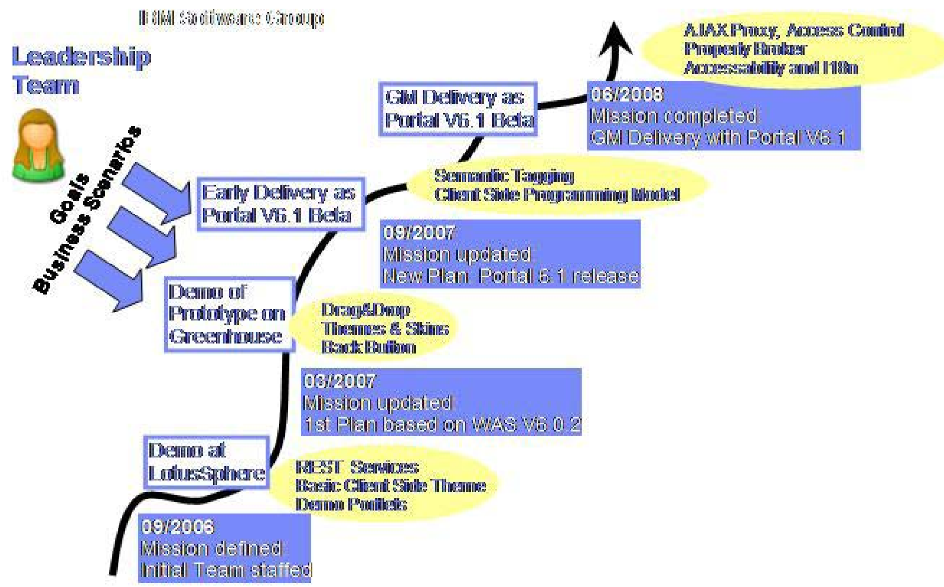
And change needs to be anticipated. For instance, after 5 months of Portal 6.1 development, a change of the underlying WAS version has been decided to meet customer requests: While the original plan has been to ship with WAS 6.0.2, the final product moved to WAS 6.1. This modification affected design documents and implementation significantly, since WebSphere 6.1 introduced an entirely new user management component and had different non-functional requirements. The project duration had to be extended. As the Web 2.0 tiger team did only have little impact by this decision, the team added more use cases to their charter document and spent the extra time until shipment on creating more functionality. In parallel another tiger team began driving the move towards WAS 6.1 (Figure 2).

CROSS-ORGANIZATION TEAMING STRUCTURE: TIGER TEAMS DRIVE INNOVATION

People fail to get along because they fear each other; they fear each other because they don't know each other; they don't know each other because they have not communicated with each other. Martin Luther King Jr.

For each of the focus areas, a tiger team is established with the right set of skills (especially test and development) to fulfill its mission. While the Portal overall organization is traditionally structured into product components, the new tiger teams, which are founded to address certain focus areas, are virtual teams. They focus on their deliverable rather than on existing organizational structures. They span multiple organizations and components, if necessary and possible. Each

Figure 2. Teaming: adjusting goals



tiger team owns the responsibility for planning, designing, coding, documenting and testing of their deliverables. Use cases are implemented by a single team end-to-end and across component boundaries, rather than coordinating multiple parallel changes done by multiple teams. This approach minimizes the cross-team dependencies, hand-offs and task switching. It supports continuous communication and focus on the deliverables.

Most importantly the functional verification testing of the team's deliverables is the team's own responsibility. Continuous testing and direct collaboration between developers and testers improve the process of troubleshooting and bug fixing tremendously as well as creates a very efficient bridge between the development and test organization (Figure 3).

ITERATIVE DEVELOPMENT APPROACH: EVOLVE A SOLUTION AND AVOID THE BIG BANG

Life is what happens to you while you're busy making other plans. John Lennon

It is hard to integrate a significant amount of code at a certain milestone date without causing a major disruption. To avoid painful integration struggle, Portal has adopted an iterative development model. Code is continuously integrated into a common code stream and functionality is brought forward in multiple, small iterations.

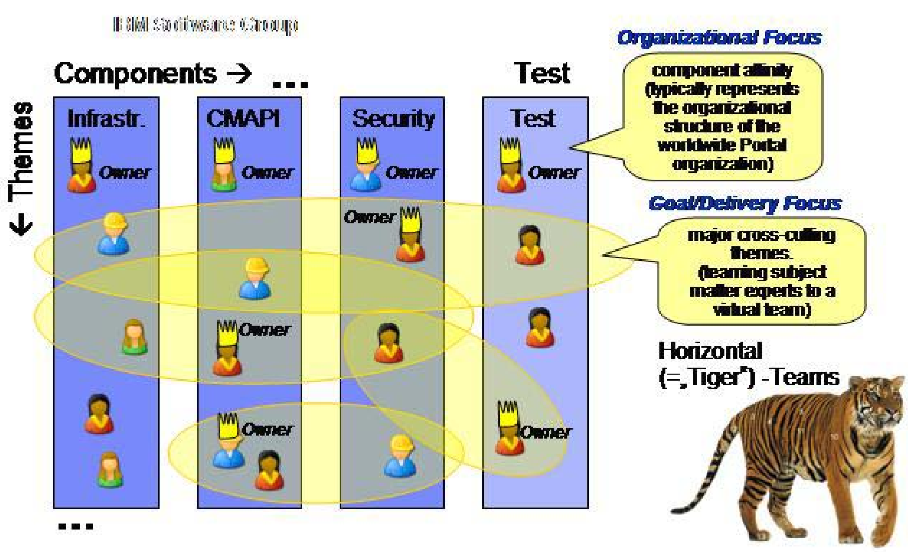
The purpose of an iterative development is nicely summarized in IBM's Agile Manifesto:

Agile Software Development uses continuous stakeholder feedback to deliver high-quality and consumable code through use cases and a series of short, time-boxed iterations.

This implies a few key assumptions:

- The **duration** of an iteration varies between 4 and 6 weeks.
- The **content** of each iteration is defined at the beginning of each iteration. A tiger team picks the top use cases from their prioritized team charter and starts designing and coding those items.
- Throughout the iteration the teams **continuously integrate** their code, documentation and automated test cases into a common code stream. There are daily builds of the entire product. Continuous integration with immediate testing is extremely important to avoid the destabilization which typically arises, when multiple developers add code back into the code stream, which they have accumulated over quite some time.
- Everyone's major goal is to ensure the **stability** of each build. Disruptive changes are to be avoided by all means. Thorough unit testing and automated regression testing is the responsibility of every single developer. The build environment gives some support, by running each code change through a pre-build to surface potential issues prior to the regular production build. It is a mandate to focus on any open issues and bugs first, before proceeding with the development of new functionality ("Stop-the-line" concept).
- Part of the iteration is **functional verification testing** within the tiger team. Only tested and working use cases are accepted as a delivered achievement.
- **Performance** and **documentation** are further aspects to be covered within the iteration.
- In order to be able to implement a use case within an iteration it is extremely important to break larger user stories into **smaller, digestible chunks**.
- Iterations are **time-boxed**. They have a defined start and end date. Usually all tiger teams operate on the same iteration schedule.
- At the end of each iteration, all tiger teams are jointly demonstrating their deliverables

Figure 3. Teaming: theme and component matrix

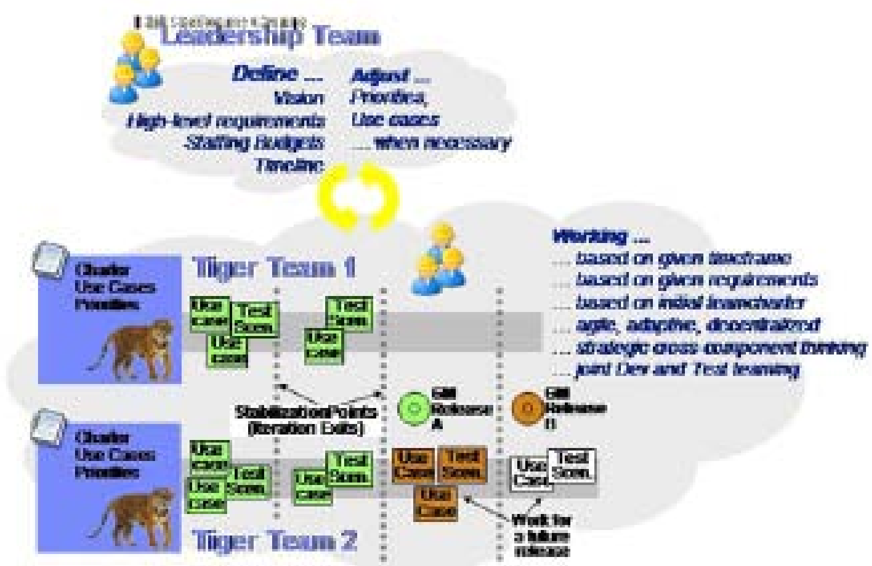


to the worldwide team. This demo event is referred to as "**Integration Fest**" and is performed using a regular build. The demo should proof that a stable, tested, and usable version of the product has been accomplished and that this version can be delivered to exploiters of WebSphere Portal.

- At the end of an iteration, the tiger team needs to proof that it has executed the agreed test cases, and that no defects are open. This will also be validated by Performance and System Tests that immediately follow each iteration.
- **Feedback** by the exploiters is incorporated into the next revision of design and plan.

Throughout the release, the teams maintain their prioritized list of use cases which they tentatively want to address in the foreseeable future, and they have elaborated a rough high level design outlining all items of their focus area. But only the current iteration is being precisely planned and detailed into work items as well as low level use case descriptions (Figure 4).

Figure 4. Iterations: executing the release project



TEST DRIVEN DEVELOPMENT: GAIN TRUST AND QUALITY FROM DAY ONE

If anything can go wrong, it will. Murphy's Law

As mentioned above, the collaboration between development and test is absolutely key for achieving a usable level of quality throughout the release project.

There are several different test considerations:

- **Build Verification** and Smoke Testing is done for each product build. It is partially automated and covers the very basic product capabilities.
- **Functional verification testing** is done within each Tiger team and is an integrated part of ANY development activities. Design documents with use case descriptions already consider test scenarios and test planning. The testing is includes automated regression testing.

- **Extended Functional Verification Testing** is executed by a dedicated test team in parallel to the ongoing iterations. Main focus is extending platform coverage and adding more regression testing.
- **Performance testing** is going on throughout the release project. Goal is to identify performance regression, but also to help the development teams with analyzing the performance bottle necks in the implementation.
- **System Verification Testing** is done in parallel to the ongoing iterations as well. This testing covers long running test scenarios as well as complex environment such as cluster setups. Goal is to identify potential memory leaks or instabilities early.
- Once the last iteration is completed, a "**Release Closure**" phase is appended, in which the test teams conduct further regression testing. There is especially a focus on system verification testing.

THE ISSUES AND PAIN POINTS

No man ever steps in the same river twice, for it is not the same river. Heraklit

In summary, agile development practices proofed to improve the flexibility of the release project and especially the efficiency within the development and test organization. Nevertheless the move from the waterfall model towards tiger teams and iterations has been a challenging journey. Portal has mixed and matched suitable ideas from the palette of known agile techniques and concepts. Part of this journey has always been to reflect the applied techniques and adapt the progress. Learn as you go. Each project is different. Each team has different needs to perform. Continuous feedback from the overall team triggered the continuous improvement of Portal's agile approach.

There have been key challenges to resolve, but there is no generic solution, which fits all. The experience made by WebSphere Portal can be summarized as a set of guidelines and considerations:

Finding the right **balance between flexibility and planning** is difficult. Do not waste time with planning exercises on a too granular level of detail. The project status 6 months from now is not predictable. Making early commitments will narrow down the possible options for your future.

Flexibility is good, but no for free! Deferring decisions will require additional investment, as some simplifying assumptions cannot be made yet. Planning and documentation are needed to document agreements and commitments, to describe the architecture, and to specify interfaces. Don't use "Agility" as an excuse for lack of planning and design preparation. This is quite similar to purchasing plane tickets: If you decide late and want full flexibility, the airfare will be higher compared to early bookings for a fixed date.

In the context of the Portal product, key architectural decisions, like choosing the underlying WAS version or selection of code streams in the source library system, are constraints which are rather expensive to hold off. Defining the major focus areas of investment within the release project and a rough timeline is crucial. It needs to be done early, whereas finalizing the exact feature scope or the shaping the exact out-of-the-box functionality can be done rather late. The first iterations will give a well-founded assessment of progress, status, quality and the remaining capacity of the teams. That kind of information will be a much more reliable baseline for an ongoing fine grained planning.

It is important to **break down big pieces of work in smaller use cases**, which can be implemented and tested within a single iteration. This helps to create frequent and measurable results, instead of piling up a lot of unfinished work. It also helps to monitor and understand progress – an essential information for fine tuning the planning of the upcoming iteration.

You need to **plan for change**. Therefore any planning or commitment needs to ensure, that there is sufficient buffer to accommodate the unforeseeable events. It is very useful to only commit to few, most important use cases initially. The team can add more, once the product is getting shape and there is less ambiguity in the planning constraints.

Test automation is crucial. If there is no extensive coverage of automated functional verification tests, it is impossible to avoid regression problems, when the next set of use cases is implemented in a subsequent iteration.

Agile Software development is demanding openness for change and flexibility from everyone. Agile needs commitment by the entire organization in order to work out:

Each **Tiger team** is accountable for its results. Everyone needs to understand what other team members do. Teams need to keep risks in mind, take over responsibility and be prepared to come up with mitigation plans quickly. Tiger teams span organizational structures and location. They drive customer oriented, end-to-end use cases with less gaps in between. They think out of the box rather than being cramped by a component centric point of view. This is the key strength of Portal's agile approach.

The **leadership team** needs to trust and empower the teams. They need to give guidance by communicating a well defined overall Vision, high level requirements and a rough timeline in which deliverables are expected. And they need to encourage a common team spirit within each tiger team.

Agile Software Development is a lot about philosophy. As mentioned above, there is no general rule, no generic process. Mix and Match of various techniques is the right way to go, as long as you accept that change is part of your plan. Based on the experiences outlined in this article, the authors have published a comprehensive book covering the basics of agile software development. In particular, this book emphasizes best practices for large software development projects.

ACKNOWLEDGMENT

SUMMARY

The adaptation of agile software development has been a quite challenging and disruptive exercise. Most important has been the ability to break down complex development items into manageable use cases, which can be designed, scheduled, implemented, tested and tracked easily. The major challenge has been to get the ongoing support from the key players in the organization. As described above, the move towards a more distributed leadership model, as well as letting go of centralized control has been controversial throughout the project. But in the end, agile development has been a convincing approach, despite all the issues and struggles which had to be resolved during the project. Its flexibility turned out to be especially supportive for such a large development organization. Future Portal releases will continue to evolve agile development and further improve and adapt the process to meet the team's requirements.

REFERENCES

Beck, K. (2002). *Test Driven Development*. Addison-Wesley.

Kessler, C., & Sweitzer, J. (2007). Outside-in Software Development. Addison-Wesley.

Poppendieck, M., & Poppendieck, T. (2006). *Implementing Lean Software Development*. Addison-Wesley.

Schwaber, K. (2004). *Agile Project Management with Scrum*. Microsoft Press.

Stober, T., Hansmann, U. (2009). *Agile Software Development: Best Practices for Large Software Development Projects.* Springer.

LINKS

Agile Forum: http://ibmforums.ibm.com/forums/forum.jspa?forumID=2710

Poppendieck Lean website: http://www.poppendieck.com/

ScrumAlliance.org: http://www.scrumalliance.org/

This work was previously published in International Journal of Web Portals, Vol. 1, Issue 3, edited by J. Polgar, G. Adamson, pp. 44-55, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 2.9
Adaptation and Recommendation in Modern Web 2.0 Portals

Andreas Nauerz
IBM Research and Development, Germany

Rich Thompson
IBM T.J. Watson Research Center, USA

ABSTRACT

In this paper, we propose a generic recommender framework that allows transparently integrating different recommender engines into a Portal. The framework comes with a number of preinstalled recommender engines and can be extended by adding further such components. Recommendations are computed by each engine and then transparently merged. This ensures that neither the Portal vendor, nor the Portal operator, nor the user is burdened with choosing an appropriate engine and still high quality recommendations can be made. Furthermore we present means to automatically adapt the Portal system to better suit users needs.

INTRODUCTION

In recent years Enterprise Information Portals have gained importance in many companies. As a single point of access they integrate various applications and processes into one homogeneous user interface. Today, typical Portals contain thousands of pages. They are no longer exclusively maintained by an IT department, instead, Web 2.0 techniques are used increasingly, allowing user generated content to be added to Portal pages. This tremendous popularity and success of Portals, has its downsides: Their continuous growth makes access to relevant information increasingly difficult. Users need to find task- and role-specific information quickly, but face information overload and feel lost in hyperspace. The huge amount of content results in complex structures designed to satisfy the majority of users. However, those superimposed structures, defined by Portal authors and administrators are not necessarily compliant to the users' mental models and therefore result in long navigation paths and significant effort to find the information needed. The likelihood of a mismatch between a user's mental model and the administrator's mental model increases as more users access the Portal. This becomes even worse, once user generated content is added, where the structure may not follow the design the administrator had in mind. In addition, the more content a Portal offers, the more likely it becomes that users are no longer aware of all the resources available within it. They might thus miss out on resources that are potentially relevant to their tasks, simply because they never come across them. Thus, on the one hand, users obtain too much information that is not relevant to their current task, on the other hand, it becomes cumbersome to find the right information and they do not obtain all the information that would be relevant. Users therefore need the Portal to assist them in finding relevant information in an efficient manner.

Generally this type of problem falls in the domain of recommender systems and numerous such systems have been proposed in recent years. Each of these can recommend relevant items for specific applications or when certain data characteristics are met, but none meet the breadth needed to address assisting a Portal user. In this paper we will outline a generic recommender framework into which specific recommendation engines can be installed. The framework decides which engines are likely to produce relevant recommendations for any particular situation and how multiple results sets are combined when multiple engines are invoked. The framework comes with a number of preinstalled engines and basic configuration for using them. This alleviates the burden on the Portal administrator relative to the initial configuration and transparently leverages the best engines for assisting the user in accomplishing their task.

In addition to providing users with recommendations we also adapt the Portal's structure automatically to better satisfy users needs.

Most of our solutions for adapting and recommending content are based on user and context models that reflect users' interest and preferences and on annotations of resources provided by users. For instance, we adapt a Portal's structure (e.g. navigation) and provide recommendations to be able to reach content being of interest easier. We recommend background information, experts and users with similar interests.

In the following we will first give an overview of related work. Next, we outline which information is needed to achieve our goals, and, more importantly, how to obtain the necessary information. Here the focus lies on collecting information about users (and their interests, preferences and thus needs) and the resources they interact with. We will show that a mixture of automated information extraction and user provided information is currently the most realistic approach. Afterwards we demonstrate how this information can be exploited to either adapt the Portal or to issue reasonable recommendations.

RELATED WORK

The explosive growth of information on the Web has led to the development of recommender systems [Resnick and Varian, 1997]. Recommender systems are a personalized information filtering technology used to either predict whether a particular user will like a particular item (prediction problem) or to identify a set of N items that will be of interest to a certain user (top-N recommendation problem). In recent years, recommender systems have been used ina number of different applications such as recommending products a customer will most likely buy; movies, TV programs, or music a user will find enjoyable. An excellent survey of different recommender systems for various applications can be found in [Schafer et al., 1999].

Over the years, various approaches for building recommender systems have been developed [Ramezania et al.,2008]: Collaborative Filtering (CF) recommenders use social knowledge - typically ratings of items by a community of users to generate recommendations. Content-based (CB) recommenders use item features to recommend items similar to those in which the user has expressed interest. Knowledge-based (KB) recommenders use domain knowledge to generate recommendations. Hybrid recommender systems combine two or more techniques to gain better results with fewer drawbacks.

The recent competition to improve the recommendation system employed by Netflix has shown the value of applying multiple recommendation engines (a combination of an algorithm and a specific configuration) to be a problem. The Progress winner employed many variants [Robert M. Bell and Volinsky, 2007], each of which was designed to do best in certain circumstances, such that the overall quality of the generated recommendations was improved. This demonstrates the value of multiple engines within a single domain and the effect is multiplied when multiple application domains are accessible through a single system, as is common for Portals.

More specifically, regarding the recommendation of expertise, systems that help to find experts are called expertise finders or expertise location engines [Zhang and Ackerman, 2005]. A general architecture for recommendation systems that allow locating experts is described in [McDonald and Ackerman, 2000]. More specifically Streeter et al.present Who Knows, a system which recommends experts having knowledge in specific topics based on profiles created from observing the documents they have selected and worked with previously [Streeter and Lochbaum, 1988].Newer systems that use information about social networks to find experts are e.g. [Kautz et al., 1997].

Providing background information or interlinking information pieces is based on the ability to either allow users or programmatic, automated, annotators to annotate information pieces. We have described the first approach in [Nauerz and Welsch, 2007] already. The second approach is based on information extraction from unstructured machine-readable documents. Although the approach to perform the extraction often differs, most papers in this area regard information extraction as a proper way to automatically extract semantic annotations from web content. Most of these systems are based on machine learning techniques, e.g. [Dill et al., 2003].

With respect to adaptation systems, systems that build and apply user and usage models to adapt web sites to the user's context (interests, preferences, needs, goals, etc.), a lot of research has already been performed in the field of adaptive hypermedia [Brusilovsky, 2001]. One possible approach to derive those models and enable adaptation is to analyze user access data, as Perkowitz and Etzioni [Perkowitz and Etzioni, 1997] propose. Projects in this context include PageGather [Perkowitz and Etzioni, 2000], Letizia [Lieberman, 1995] and WebWatcher [Joachims et al., 1997]. Especially with respect to navigation adaptation Smyth and Cotter [Smyth and Cotter, 2003] describe an approach to improve navigation in mobile Portals significantly.

INFORMATION ABOUT USERS, BEHAVIOR, AND RESOURCES

From a conceptual point of view (cp. fig. 1) Portals are comprised of various resources such as pages and portlets (artifacts residing on pages delivering content). These resources are arranged based on Portal models, often initially created by some administrator with the aim to satisfy the largest set of users and not the preferences of each single user. We therefore need information about individual users (or groups of users) and their behavior as a basis for both adaptation and recommendations. We apply different techniques such as web mining to construct user models reflecting users interests and preferences; we use information from their static profile (native language, home country, working location, age, etc.), their interaction behavior (pages and portlets they work with; tags they apply to resources), and their social network to derive knowledge about their needs. We observe the context (date, time, location, ...) in which they interact to partition the user model in so called profiles like private or business. Additionally, we need enriched information about the resources available in the system. We illustrate how we extract information pieces of certain type in order to provide background information by connecting to external sources and to interlink them in order to issue recommendations.

Extracting Information about Users

User Model

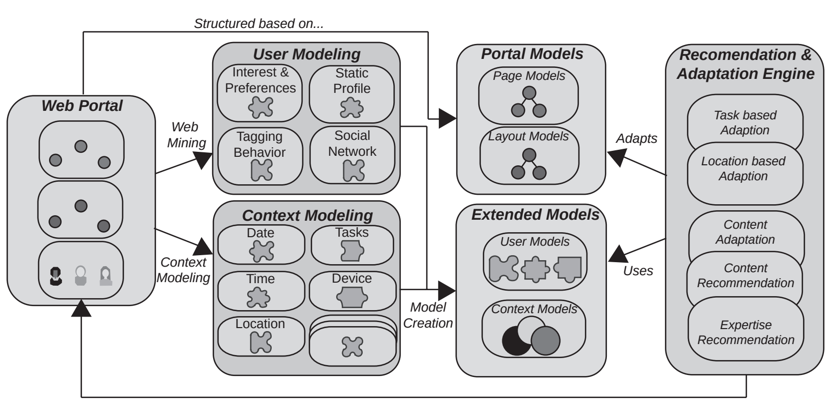
In order to perform reasonable adaptations, or to provide users with recommendations, we need to understand users' interests and preferences. Therefore we construct user models reflecting their behavior. We use static information from users' profiles (describing their age, native language, etc.), as well as dynamical information which we retrieve via web usage mining.

Web Mining [Liu, 2006] is the application of data mining techniques to discover (usage)-patterns within web data. Web usage mining is the extraction of usage patterns from access log data to model certain aspects of the behavior of users. Our system has been incorporated into IBM's WebSphere Portal. Analyzing its logs reveals information about, among other things, several interesting events, e.g. when pages (or portlets) are created, read, updated or deleted, when pages (or portlets) are requested,when users are created, updated, deleted and many more.

Analyzing the log allows to understand which pages and portlets a user typically works with. Obviously, the user model must allow the calculation of the utilization of pages and portlets from the historical data available. We do this by measuring how often a user interacts with certain pages and portlets. Of course, we also consider interactions that occurred recently to be more important than interactions that occurred in the past and we hence apply time-weighting factors when calculating the utilization of pages and portlets based on the target hits they received.

More generally, we apply techniques from the area of frequent set mining [Liu, 2006] to analyze the usage of pages and portlets. We use the Apriori algorithm [Agrawal and R., 1994], a standard association rule mining algorithm, to determine items, such as pages and portlets that co-occur frequently. We apply the GSP algorithm [Srikant and Agrawal, 1996], a standard sequential

Figure 1. Conceptual overview



pattern mining algorithm, to determine sequences of items, such as pages and portlets, that co-occur frequently. Comparing the itemsets even allows to find users behaving similarly.

Tagging Behavior Analysis. In our previous work we have incorporated a tag engine into IBM's WebSphere Portal allowing users to annotate resources such as pages, portlets and users. Analyzing (and comparing) the tagging behavior of users allows for refining the user model. The general assumption is that tagging expresses interest in a resource.

Social Network Analysis. Finally, the analysis of users' explicit contacts allows to determine users' interests and preferences, too. The assumption is that the fact that users directly know each other can be an indication for related job roles and hence for sharing similar knowledge.

Context Model

Focusing on user models only neglects the context users are currently acting in. Hence, these could be regarded suitable models, only, if the role, the interests and preferences of users will not change too much over time. In reality, a user's needs usually change if their context changes. For example, a user in the process of planning a business trip will need resources that provide information about hotels, rental cars, and flights. When the same user returns to their tasks as a project manager, a completely different set of resources is needed. Of course, interests and preferences will be different in these roles and access to a different set of resources (pages, portlets, etc.) will be needed.

Our solution allows single users to have several context profiles between which either the system switches automatically, based on context attributes being observed (date, time, location, etc.), or the user can manually switch. New profiles can be defined using a profile management portlet which allows to specify the initial settings of a profile (which theme to use, which skin to use, etc.) and to associate it with a set of context attributes (date, time, location, etc.) which define when it should become active.

Our adaptation and recommendation components utilize both the information stored in the user and context model, to perform their operations (i.e. to adapt structures such as the navigation). Technically, the user model is partitioned in a separate partition for each context profile available in the context model. To determine the best matching profile, the system continuously observes a set of defined context attributes. Users always have the option to outvote the system's decision and to manually switch to another profile. As only one context profile can be active at one specific point in time, whatever people do only influences the user model partition associated to the currently active profile. For example, if the currently active profile is trip planning, all the navigation behavior, does not influence the user model partition associated to the profile project management.

Extracting Enriched Information about Resources

To extract enriched information about the resources, we currently allow for the usage of three different mechanisms:

Automated Tagging. Here the system analyzes markup generated by the Portal to find occurrences of identifiable information pieces of certain types such as persons, locations, etc., and wraps these into semantic tags. We have integrated the UIMA framework 1 and written customized analysis engines able to identify such information pieces. Semi-automated Tagging If the system cannot unambiguously identify the type of an information piece it still allows users to mark it and tell the system of what type it is. We call this process semi-automated tagging. For instance, if we find a fragment "Paris H. was sighted leaving a Hotel in Paris" it becomes difficult for the system to determine whether Paris is a name or a location. The user can then mark the corresponding information pieces and tell the system their type.

The information pieces are then wrapped into a semantic tag exactly as outlined before.

Manual Annotating Moreover, our system allows semantically tagged information pieces to be annotated manually again. For example, if the name of three persons Alice, Bob, and Charly often appear somewhere in the Portal system, e.g. in blog- or wiki portlets, our system automatically determines these fragments to be of type person, wraps them into semantic tags and allows for advanced interaction with these information pieces. Our tag engine allows these enriched fragments to be annotated e.g. with the term project-x which indicates that all three persons are somehow related to this project. This means that the options for manual annotating allow for a finer-grain categorization of information pieces.

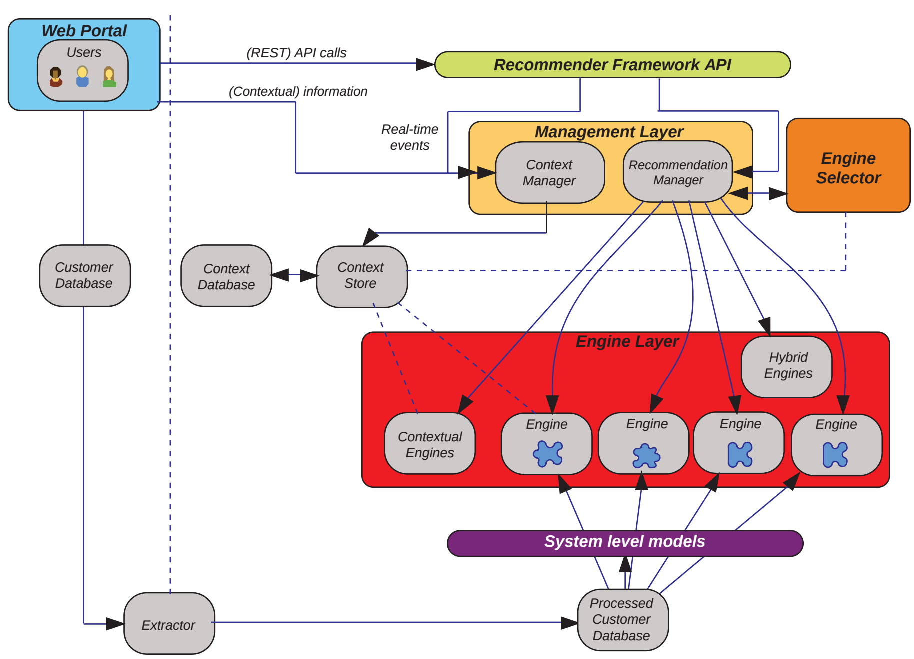
Supporting Framework

Figure 2 shows the architecture of our recommendation framework. The main components are an extractor framework, a management layer, an engine selector and an engine-hosting layer.

The extractor component manages getting incremental changes from the customer's underlying databases and making them available for processing into the recommendation models used by the deployed engines. While there are some cases of reusing shared (or system-level) models, most recommendation engines build their own models and optimize these for rapid recommendation generation. As a result, each engine's configuration will typically register a processing module with the extractor in order to receive and deal with each incremental update to the customer's data. By centralizing the extraction, impacts on the customer database are minimized as is the effort need to provide access to a new type of data source.

There are two components which are instrumented through the page and page-support mechanisms. The first of these manages contextual information. Some contextual information is already gathered elsewhere (e.g. from information

Figure 2. Recommendation framework



already logged), as discussed in Section 3.1, and this is accessed from those sources; for example, the pages activated. Other contextual information is application specific and are sent as contextual events into the system; for example, dynamically presented information has been accessed. This information is timestamped and placed into a central store which system's components, usually the recommendation engines, can use when responding to a request.

The other part instrumented on the page is requesting recommendations be generated. Each page could be displaying multiple different types of recommendations (experts, similar people, pages, portlets, etc). These are requested, normally using an AJAX request, and displayed in whatever manner the user experience prefers. The system provides dojo widgets which can be leveraged for quickly adding recommendations to a page. These include a data widget which manages requesting recommendations, parsing the response into javascript objects and providing easy access to these objects. It also includes both a simple display widget for showing any one recommendation type (e.g. experts) as an HTML list and a more complex widget that displays each recommendation type in a separate tab of a tabbed pane construct.

When the system receives a request for recommendations, it examines the request to determine what recommendation engines should be used to generate the response. The management layer delegates this decision to an engine selector and then invokes the identified engines, passing them both the request parameters and a reference to the active context store. The engine selector remains an area of active research with the simplest variant being a rule engine. It is expected that this research will realize a significant improvement by having the engine selector specify both the engines to use and how their results are to be combined. Each engine will use whatever portion of the supplied information is appropriate to their algorithm in order to generate their response. Section 5.2 discusses some of these engines, including those supplied along with the base configuration. The recommendation manager will then combine the responses from all invoked engines and provide a single, unified response to the client.

The recommendation engines are hosted by a layer that uses Inversion of Control [Wikipedia Foundation, 2008] concepts to define engine instances and the configuration parameters to apply to each instance. Often the referenced components are wrappers that map our internal API and data structures to the ones used by the actual engine. For example, a wrapper for the open source collaborative filtering engine Taste [Apache Foundation, 2008] has been used to incorporate this externally developed engine. Due to the nature of the design, we expect similar wrappers to be easily developed for other external engines. This architecture allows very different engines to be deployed (collaborative filtering, content-based, knowledge-based, etc.) and work independently of each in a manner that will together generate improved results for the user. It also makes it easier to add recommendations to a website as many of the issues of accessing engines are provided by the framework.

EXPLOITING THE MODELS FOR ADAPTATION AND RECOMMENDATION

Now that we have described which information about the Portal resources and users are available to our system, we can explain how this information is used to improve the user experience with the Portal. We propose methods to adapt the content, to recommend content, to offer additional information and to recommend experts. In the following, more details about the approach are given.

Adaptations

Within the context of this project we have come up with different solutions allowing for adaptation and recommendation of the Portal's structure. Most of them focus exemplarily on the adaptation of the navigation.

Manual Adaptation. First, options to manually adapt the navigation have been introduced. Therefore we implemented specialized portlets that allow each single user to generate her own navigation matching her preferences best. The first portlet allows users to generate their own navigation by hiding irrelevant nodes (pages) and by reordering nodes being part of the navigation in order to reach relevant nodes more quickly. The second portlet allows users to record paths (i.e. sequences of pages) traveled often. These recordings can be recalled later and navigated through by just clicking previous and next links. The recordings can even be exchanged with other Portal users which allows experts to record common paths for their colleagues.

Automated Adaptation. Automated adaptation relieves users from generating an optimized navigation manually. We leverage our user models to understand users' needs.We use a structure reordering algorithm to rearrange pages: more important nodes are promoted to better navigational positions, less important ones demoted or even hidden. Continuous adaptation, based on the most current user models available, guarantees that the navigation permanently fits the users needs as best as possible. As soon as users' behaviors change their user model changes, too and hence the navigation provided. We leverage our user models to understand users' needs. We use a structure reordering algorithm to rearrange pages: more important nodes are promoted to better navigational positions, less important ones demoted or even hidden. Continuous adaptation, based on the most current user models available, guarantees that the navigation permanently fits the users needs as best as possible. As soon as a user's behavior changes, their user model is updated and hence the navigation provided takes into account the new behavior.

Automated Recommendation. Especially users that navigate according to the aimed navigation paradigm [Robertson, 1997] will not like automated adaptations because of its aggressiveness. Automatic provisioning of recommendations avoids the permanent restructuring of the navigation while still providing users with shortcuts. We blend-in recommendations into the Portal's theme that provide users with reasonable shortcuts to relevant pages. These shortcuts are dynamically generated depending on the current navigational position. Our recommendation system applies a MinPath algorithm [Anderson et al.,2001]. We try to predict shortcuts to nodes that are far away from the current node but have a high probability to be navigated to. The probability itself is calculated based on Markov chains as described in [Anderson et al., 2002; Smyth and Cotter, 2003].

Context-adaptivity. As mentioned above, users may have several different context profiles. By switching to a different profile, the Portal will be adapted accordingly based on the information contained in that profile.

Recommendations

In our framework, we have developed a set of recommender engines that provide recommendations to background information and related content either by displaying shortcuts to relevant pages or by showing widgets/portlets containing links to relevant information. Currently these engines come in two flavors (though the framework is not restricted to just these flavors): recommender engines based on the user model and collaborative filtering recommenders. Further-more, both types of recommender engines can leverage the context model in order to provide recommendations with respect to the current activity of the user.

Recommenders Based on User Model

The first category of recommender engines provide recommendations to related content based on the information stored in the user model, which reflects static information about the user (like language preferences, age, location, etc.) as well as navigation behavior of individual users, including their favorite pages and routes through the entire navigation topology of the Portal. Recommendations are generated based on certain properties of the user model and the resulting recommendations can be used to suggest forward navigation links or to automatically change the navigation topology in order to reveal the interesting pages and hide irrelevant content.

Collaborative Filtering Recommenders

In contrast to the recommenders that provide recommendations based on the analysis of the history of a particular user, the collaborative filtering (CF) recommenders provide recommendations based on the analysis of the behavior of multiple users. The collaborative filtering engines first try to identify the users that have similar tastes and behaviors to the current user and then retrieve the items that these users liked most and recommend them to the current user. The CF-based recommendations help to discover new and unknown content items that might be of interest for a particular user.

The CF-approach can be applied in numerous ways; for recommending navigation links, the set of similar users might best be chosen based on having a similar set of target hits as the current user. The CF-approach would then recommend those pages for which these users also have high target hits, but which the current user does not. In domains such a catalog systems (e.g. movie sales/rentals), preferrably the ratings of the user vs other users is used to generate the recommendations. In the absence of explicit ratings, previous history and the items viewed within the current interaction session can be used as implicit forms of ratings.

Other Recommenders

While not included in the current set of supplied recommendations engines, the framework takes into account that other recommendation strategies exist (and continue to be invented). At a minimum these include Content-Based recommenders and various different approaches to Knowledge-Based recommenders.

Context-Based Recommendations

So far we have described only the recommendations based on the navigation history of individual users and commonalities of interests among multiple users without respect to the context in which the users are acting. However, the initial recommenders of both categories can also access the context model in order to provide recommendations that could be especially relevant in certain situations. Our recommendation framework allows single users to have several context profiles. These profiles can be recommended to the user automatically by the system or users can switch between them manually.

New profiles can be defined using a profile management portlet which allows specifying the settings of a profile (which theme to use, which skin to use, which navigation topology to display etc.) and to associate it with a set of context attributes which define when it should become active. Whatever people do only influences the models associated to the current active profile. If the currently active profile is business, all the navigation behavior, all the usage behavior of Portal pages etc. does only affect the models associated to this profile, but never influences the modelsassociated to the private profile. For the determination of the best matching profile the system continuously observes a set of defined context attributes. Users always have the option to outvote the system's decision and to manually switch to another profile. A learning mode allows users to let the system learn about their needs and tastes in a specific new context.

Recommending Background Information

As said, in today's Web 2.0 world content is created by entire user communities. Different users use different terms to describe the same things. Some terms might be well-understood by most users, some might not.

Thus looking up terms is needed more frequently and becomes a tedious task. But when reading web sites, users want background information at their fingertips. If they do not understand what an abbreviation or a term stands for, who a certain person actually is, or, where a certain city is actually located, they want to be able to retrieve this information as easily and quickly as possible. They do not want to fire up a search engine to search for another site from which they could probably get the information they want, but rather be provided with that information directly, in-place. We provide an environment which unobtrusively enriches the information pieces to allow for such look-ups.

Figure 3 shows our system in action: it illustrates how a fictious person name (John Doe), a location (Stuttgart), and a currency have been identified within a text fragment residing in a portlet and are visualized to the user. Pop-ups provide the users with background information.

Recommending Related Content

Analyzing occurrences of semantically tagged information pieces also allows us to recommend related content. For instance, if the term WebSphere Portal is identified in a news portlet and hence semantically tagged as a product name our system would provide users with background information about WebSphere Portal probably by linking to the product site. But, within a Portal system, the same term might occur at many other places, e.g. in a wiki portlet where users have posted some best practices, tips and tricks when working with this product, in a blog where users have commented on the product and so forth. We track all occurrences and recommend (an appropriate subset) of them as related content as soon as the user interacts with one single occurrence.

This can even be taken one step further. As mentioned above, we allow users to annotate already semantically tagged information pieces. This way we can recommend related content not only by having identified "exactly matching" occurrences of semantically tagged information pieces, but also by having identified similarly annotated, but differently semantically tagged, information pieces. For example, if Alice, Bob, and Charly have been annotated as persons and a user tagged them with the term project-x to ex-

Figure 3. Recommending background information and related content

press their relationship to this project, this allow us to recommend other users of the community as related ”content” as soon as one user is clicked, just because they all seem to be assigned to the same project. This can be useful, e.g., if a user has a question to something he is reading about and tries to contact the author whose name is given but who is currently unreachable. Recommending related users allows him to easily determine backups that could probably help him, too.

Figure 3 shows how we can recommend related information for the detected information pieces Stuttgart and John Doe (other people probably working in the same team, on the same project etc.).

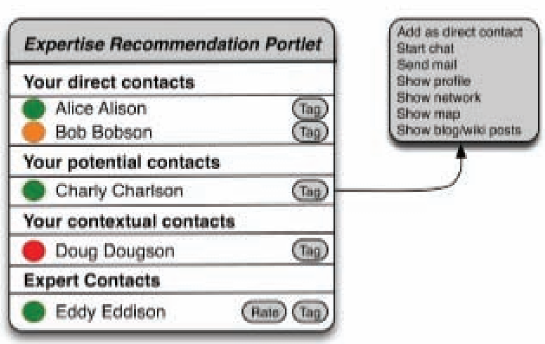
Recommending Expertise

As said, user models also tell us about with which pages and portlets a user is typically working with.

The first assumption is that users working with certain pages and portlets more often have more expertise about how to use them than other users have. The second assumption is that users working with the same pages and portlets more often have a similar behavior and hence interests and preferences.

For example, if users A, B, and C often work with the pages and portlets underneath the page entitled My News we can, on the one hand assume that they have knowledge about how to deal with the pages and portlets provided here, and, on the other hand assume that they have similar interests as they do similar things. A user D accessing the same pages and portlets rarely can then be presented with A, B, and C as experts when dealing with the information and services provided. Therefore we have designed a specialized portlet (cp. Figure 4) that can be accessed from every page (on demand).At the top of the portlet contacts are listed which have explicitly been added as such by the user. Adding contacts explicitly demands sending a request to the contact to be added which, in turn, can accept or reject this request. The second section displays contacts the system has determined to behave similarly which can be derived by comparing user models. These are contacts the user might want to get in touch with, generally to share knowledge. The third section displays contacts currently performing similar actions within the Portal (e.g. viewing the same page or working with the same portlet). The last section at the bottom of the portlet displays contacts the system has determined to be experts with respect to the (content) area currently being visited by the user.

Figure 4. Recommending experts



The list of users displayed in the last three sections dynamically changes as the current user interacts with and navigates through the Portal, while the first section always displays a static list of contacts. The portlet provides several functions to interact with the contacts being listed:

People-awareness ensures that users contact other users being available more likely than the ones being unavailable. To ensure that users do not disturb the latter, e.g. during meetings, an online status is displayed for each single user displayed as part of the portlet. Tagging and rating functions allow contacts to be tagged, expert users even to be rated. This allows for a user-driven categorization of the contacts. Additionally, rating allows assessing how helpful an expert was. Profiles display information about the contact's official job role, his position within the organization's hierarchy, his address, and so forth. Social network visualization functions allow the user to see a visual depiction of how his contacts are related to him and to each other. It presents a graph which nodes are the user's contacts. It allows users to determine which users are part of the same team, who personally knows whom and hence allows e.g. to find backups for persons being currently unavailable easily. It is also possible to see contacts as marks on a GoogleMap.

CONCLUSION AND FUTURE WORK

In this paper we have presented a solution for adapting and recommending content and expertise to satisfy Portal user needs and improve collaboration among them. We have shown means to collect the necessary information and to adapt the navigation structure, to recommend background information, related content and expertise.

We have presented an extensible, pluggable, recommender framework which allows several recommender engines, that can even apply different recommender technologies, to be used transparently.

All the approaches proposed in this paper have been implemented and integrated into IBM's WebSphere Portal.

For evaluation purposes we have set up a demo system and performed some initial surveys. 100% of all participants (all computer scientists, male, 25-50 years old) regarded the system as useful. Of course, we plan to perform more systematic evaluations within the next months.

Future work includes the extension of our recommendation and adaptation techniques.

ACKNOWLEDGMENT

IBM and WebSphere are trademarks of International Business Machines Corporation in the United States, other countries or both. Other company, product and service names may be trademarks or service marks of others.

REFERENCES

R. Agrawal and Sirkant R. Fast Algorithms for Mining Association Rules. In Proc. of the 20th Very Large Data Bases Conf., Santiago, Chile, 1994.

Corin R. Anderson, Pedro Domingos, and Daniel S. Weld. Adaptive web navigation for wireless devices. In IJCAI, pages 879–884, 2001.

Corin R. Anderson, Pedro Domingos,and Daniel S. Weld. Relational markov models and their application to adaptive web navigation. In KDD, pages 143–152, 2002.

Apache Foundation. Mahout (a.k.a. Taste), 2008. http://lucene.apache.org/mahout/.

Peter Brusilovsky. Adaptive Hypermedia. User Modeling and User-Adapted Interaction, 11(1-2):87–110, 2001.

Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, Ramanathan V. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin, and Jason Y. Zien. Semtag and seeker: bootstrapping the semantic web via automated semantic annotation. In WWW, pages 178–186, 2003.

Thorsten Joachims, Dayne Freitag, and Tom M. Mitchell. Web Watcher: A Tour Guide for the World Wide Web. In Proc. of the 15th Intl. Joint Conf. on Artificial Intelligence, pages 770–777, 1997.

Henry A. Kautz, Bart Selman, and Mehul A. Shah. Referral web: Combining social networks and collaborative filtering. Commun. ACM, 40(3):63–65, 1997.

Henry Lieberman. Letizia: An Agent That Assists Web Browsing. In Chris S. Mellish, editor, Proc. of the 14th Intl. Joint Conf. on Artificial Intelligence, pages 924–929, Montreal, Quebec, Canada, 1995. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.

Bing Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications). Springer, 2006.

David W. McDonald and Mark S. Ackerman. Expertise recommender: a flexible recommendation system and architecture. In CSCW, pages 231–240, 2000.

Andreas Nauerz and Martin Welsch. (Context) Adaptive Navigation in Web Portals. In Proc. of the Intl. IADIS WWW/Internet Conference 2007, Vila Real, Portugal, October 2007.

Mike Perkowitz and Oren Etzioni. Adaptive Web Sites: an AI Challenge. In Proc. of the 15th Intl.Joint Conf. on Artificial Intelligence, pages 16–23, 1997.

Mike Perkowitz and Oren Etzioni. Towards adaptive Web sites: conceptual framework and case study. Artif. Intell., 118(1-2):245–275, 2000.

MaryamRamezania,Lawrence Bergman, Rich Thompson, Robin Burkea, and Bamshad Mobashera. Recommending recommenders. In Intelligent User Interfaces, page 439, 2008.

aul Resnick and Hal R. Varian.Recommender systems. Commun. ACM, 40(3):56–58, 1997.

Yehuda Koren Robert M. Bell and Chris Volinsky. The bellkor solution to the netflix prize, 2007.

George G. Robertson. Navigation in Information Spaces, 1997. Submision to CHI 97 Workshop on Navigation in Electronic Worlds, March 23-24th, 1997.

J. Ben Schafer, Joseph Konstan, and John Riedi. Recommender systems in e-commerce. In EC ’99: Proceedings of the 1st ACM conference on Electronic commerce,pages 158–166, New York, NY, USA, 1999. ACM.

B. Smyth and P. Cotter. Intelligent Navigation for Mobile Internet Portals. In Proc. of the 18th Intl. Joint Conf. on Artificial Intelligence, Acapulco, Mexico, 2003.

R. Srikant and R. Agrawal. Mining Sequential Patterns: Generalizations and Performance Improvements. In Advances in Database Technology - EDBT’96,1996.

L.A.StreeterandK.E. Lochbaum. Who knows: A system based on automatic representation of semantic structure. In RIAO, pages 380–388, 1988.

WikipediaFoundation.Inversionofcontrol,2008. http://en.wikipedia.org/wiki/Inversion of control.

Jun Zhang and Mark S. Ackerman. Searching for expertise in social networks: a simulation of potential strategies. In GROUP, pages 71–80, 2005.

This work was previously published in International Journal of Web Portals, Vol. 1, Issue 2, edited by J. Polgar and G. Adamson, pp. 1-17, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 2.10

Context-Aware Applications for the Web: A Model-Driven Development Approach

Florian Daniel
University of Trento, Italy

ABSTRACT

Adaptivity (the runtime adaptation to user profile data) and context-awareness (the runtime adaptation to generic context data) have been gaining momentum in the field of Web engineering over the last years, especially in response to the ever growing demand for highly personalized services and applications coming from end users. Developing context-aware and adaptive Web applications requires addressing a few design concerns that are proper of such kind of applications and independent of the chosen modeling paradigm or programming language. In this chapter we characterize the design of context-aware Web applications, the authors describe a conceptual, model-driven development approach, and they show how the peculiarities of context-awareness require augmenting the expressive power of conceptual models in order to be able to express adaptive application behaviors.

DOI: 10.4018/978-1-60566-290-9.ch003

INTRODUCTION

The evolution of the Information Technology in the last years has seen the World Wide Web transforming from a read-only hypertext media into a full-fledged, multi-channel and multi-service application delivery platform. Current advances in communication and network technologies are changing the way people interact with Web applications. They provide users with different types of mobile devices for accessing – at any time, from anywhere, and with any media – services and contents customized to the users' preferences and usage environments. More and more users themselves ask for services and applications highly tailored to their individual requirements and, especially due to the increasing affordability of new and powerful mobile communication devices, they also begin to appreciate the availability of ubiquitous access. In order to cope with the growing demand for novel, user-centric application features, such as adaptivity and context-awareness, appropriate development methods for Web applications are required.

Adaptivity is increasingly gaining momentum in the context of modern software systems. Runtime adaptivity provides highly flexible and responsive means for the customization of contents and services with respect to the user's identity. Varying device characteristics in mobile and multi-channel computing environments can be adequately taken into account and leveraged by means of adaptive software designs, whose development is facilitated by the availability of standardized communication protocols (e.g. HTTP) and markup languages (e.g. HTML or WML), supported by most of today's mobile devices. Multi-channel deployment does no longer require completely different, parallel design approaches and rather represents a presentation issue on top of unified engineering solutions.

But adaptivity may also enable an application to take into account a wider range of properties describing the interaction between the user and the application, thus paving the way for context-awareness. *Context-awareness* (Dey & Abowd, 2000; Schilit & Theimer, 1994) is often seen as recently emerged research field in information technology and in particular in the domain of the Web. From the perspective of application front-end development it can however be interpreted as natural evolution of personalization and adaptivity, addressing not only the user's identity and preferences, but also his/her usage environment. Personalization has already demonstrated its benefits for both users and content providers and has been commonly recognized as fundamental factor for augmenting the efficacy of the overall communication of contents. Context-awareness goes one step further in the same direction, aiming at enhancing the application's usefulness and efficacy by combining personalization and adaptivity based on an application-specific set of properties (the context) that may affect the execution of the application.

In this chapter, we focus on the development of context-aware applications for the Web and, in particular, we describe a *model-driven* development method that allows developers to approach the problem at a level of abstraction that enables him/her to focus on the real design challenges of such class of applications, leaving low-level implementation concerns to supporting CASE (Computer-Aided Software Engineering) tools. Considering that software systems are continuously getting more complex and difficult to maintain – partly due to the previously described requirements –, we believe that efficient abstraction mechanisms and design processes, such as those provided by visual, model-driven design methods, are becoming crucial. The focus on essential design issues and the ease of reuse in model-driven design methods may significantly accelerate the overall design process. As we will show in this chapter, starting from application models, code generation techniques may then provide for the automatic generation of application code or templates, thus assuring the fast production of consistent and high quality implementations.

MOTIVATING EXAMPLES

Active application features, such as context-aware or adaptive behaviors, may augment the effectiveness of interactions and the efficiency of resource consumption in all those situations where services and contents offered by an application strongly depend on environmental situations, users' abilities or disabilities, or the state or health of a software system. For example, typical applications demanding for active features and adaptivity are:

- *Adaptive personalization*. User profile attributes for personalization purposes may present different levels of variability in time. Profile properties may be static in nature (e.g. the name of a user), slowly changing (e.g. profile data derived from a user's browsing behavior) or even fast changing (e.g. the pulse frequency of a patient). Adaptive personalization mechanisms that take into account such profile peculiarities could allow systems to go beyond the common and static tailoring or services and contents.
- *Interaction-enabling functionalities*. Context could as well consider handicaps or physical disabilities of users, such as vision problems, blindness or paralysis, to adapt the application accordingly and to provide alternative and better suited interaction mechanisms and modalities. Adaptivity could thus provide functionalities enabling handicapped users to properly interact with applications, thus fostering the accessibility of applications.
- *Effective content delivery*. In general, whatever context data may be leveraged to provide appropriate contents and program features at the right time, priority, and emphasis. For example, specifically targeted special offers could be advertised and directed more effectively, presentation properties could be adapted to varying luminosity conditions for better readability, etc. Adaptive or context-aware extensions could thus enhance the overall effectiveness of applications by adapting individual application elements to varying users or usages of the application.
- *Delivery of context as content*. Applications may depend intrinsically and in a structural manner from context data. Location-aware applications, such as city map services or navigation systems, treat position data as core contents of the application and adapt to them, supported by proper localization mechanisms. To such kind of applications, the use of context data represents a functional requirement, rather than an optional feature.
- *Exception handling*. Critical events during the execution of a software system may raise exceptions and require prompt reactions being performed. Process-based or workflow-driven applications, for example, represent a typical class of applications that constantly have to cope with exceptional situations in order to guarantee the consistent termination of a running process. Here, adaptive or context-aware mechanisms could be leveraged to capture respective application events and to enact the pieces of application logic that are necessary to handle the exceptional situation.
- *Production and control systems*. Critical production or control systems may require, for example, highly specific sensing and alerting mechanisms to prevent production losses or product quality degradations. Context-awareness could facilitate the timeliness of reactions and the efficient handling of dangerous situations, but also proactive maintenance approaches, such as those adopted in a steadily growing number of hardware/software systems, may be achieved.
- *Self-healing software systems*. Autonomic or self-healing software systems elevate the idea of proactive maintenance from hardware to software systems and aim at the creation of computing systems that are able to configure, tune, and even repair themselves. Proactive and adaptive capabilities in this context are an essential feature.

REFERENCE SCENARIO

To exemplify the concepts introduced in this chapter and to better convey the underlying ideas, step by step we will show how we developed one of our demonstration prototypes, the PoliTour application. The application runs on a PDA with wireless Internet access and enables visitors to Politecnico di Milano, Italy, to obtain location-aware campus details (i.e. information about roads and buildings) while walking through the campus. If a user is about to leave the WiFi-covered area of the campus, an alert message is shown.

CONTEXT-AWARENESS AND WEB APPLICATIONS

Due to a lack of appropriate technologies and concepts, for a long time context-awareness has not been considered suited to the domain of the Web. Web technologies (both hardware and software) are however continuously evolving and the attitude toward reactive and context-aware behaviors in Web applications is changing. As a matter of fact, support for a multitude of non-functional requirements, whose inadequate coverage prevented the adoption of Web technologies for the implementation of reactive applications, has now been developed. Just to mention a few:

- The *reliability* of data communications has been considerably enhanced along both the software and the hardware dimension. The introduction of reliable messaging techniques (e.g. digital certificates or the WS-Reliability specification) provides for trustworthy communications on top of standard Web protocols, such as HTTP or SOAP. The success of fiber optics – as an example of hardware evolution – has allowed the Ethernet protocol (typically used in the Web) even to enter industrial production environments, where the high electromagnetic interferences that exist in the presence of high-voltage machineries practically prohibited the use of conventional, unreliable network technologies.
- The *pervasiveness* and *availability* of Web applications is continuously growing due to the introduction of novel networking technologies, such as ADSL (Asynchronous Digital Subscriber Line) or fiber optics for home and office users and WiFi and 3rd generation mobile telephony technologies (e.g. UMTS, GPRS, EDGE) for mobile users.
- Web applications have proved a high *scalability* (it suffices to think about certain portal applications that serve millions of users every day), facilitated *maintainability* and high *cost efficiency*.

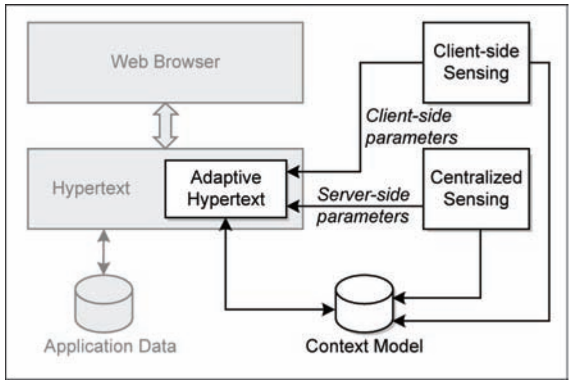
Provided that technological advances enable and facilitate the development of adaptive Web applications, it is important to recognize that context-awareness, rather than being a mere technological concern, represents a true *design* issue. In the following, we will thus focus on the typical design concerns in the development of context-aware Web applications.

Enabling Context-Awareness in the Web

Developing context-aware applications for the Web demands some characteristic architectural components, in order to support adaptations to context. Figure 1 proposes a possible functional architecture that extends the traditional architecture of Web applications with components aimed at supporting the acquisition, storage, and use of context data.

The typical context-aware application's data source includes both the application data (i.e. the business objects that characterize the application domain and the user) and the context model, which offers at any moment an up-to-date

Figure 1. Context data in context-aware Web applications. Gray shaded boxes correspond to conventional, non-adaptive parts, white boxes correspond to extensions required to support context-awareness



representation of the context state. The context model captures all the context-characterizing properties and enables the system to adapt to changes thereof, assuming that such changes may demand for proper reactions by the application. An application typically consists of adaptive (i.e. context-aware) and non-adaptive parts; we call the former adaptive hypertext. The pages of the adaptive hypertext present some form of adaptive behavior, i.e. they are able to react to changes in the context, while pages of the non-adaptive hypertext do not present any adaptive behavior. To decide which adaptation is required – if any –, the adaptive hypertext makes use of context data during the rendering of hypertext pages. Context data needs to be sensed (e.g. by means of suitable instruments, such as GPS positioning systems, thermometers, or similar) and communicated to the Web server that hosts the application, in order to be processed.

The above architecture allows for three main communication mechanisms to pass context data from the sensing devices to the application: (i) as parameters sensed at the *client side* and sent to the application (e.g. GPS position data); (ii) as *server-side* parameters (i.e. HTTP session variables) provided by a centralized sensing infrastructure (e.g. system usage data); and (iii) by means of direct updates of the *context model*. Typically, client-side parameters are generated by client-side sensing solutions, server-side parameters are filled by centralized sensing solutions, and database updates may be performed by both.

Context-awareness in Web applications therefore requires addressing the following issues:

- *Context data modeling.* Context properties that are relevant for the provisioning of the context-aware behaviors of the application must be identified and represented in an application-accessible format. The result of this task is the context model that can be queried for adaptation purposes.
- *Modeling of adaptive application behaviors*. Starting from the context model, adaptation operations need to be defined in order to react to situations demanding for adaptation. That is, detected changes to the context data are translated into visible effects or operations that aim at augmenting the effectiveness and usability of the application.
- *Context model management*. The context model only captures the static aspect of context data, i.e. their structure; in order to also capture the dynamics of context data, and hence to be able to trigger adaptive behaviors, we also need to:

  - *Acquire context data* by means of measures of real-world, physical properties, corresponding to the properties of the context model. The so acquired data are then fed into the context model, so as to keep the context model up to date.
  - *Monitor context data* to detect those variations in context data that trigger adaptivity. Relevant variations are used to enact the adaptation operations in the adaptive hypertext, thus causing an automatic, adaptive behavior of the Web application.

While the definition of the context model and the monitoring of context data can easily be assisted by proper context modeling methods and a proper runtime framework providing basic monitoring functions, it is not as easy to assist designers in the development of suitable context acquisition (i.e. sensing) infrastructures. In fact, the former two activities can be generalized beyond the needs of individual applications, while the design of sensing infrastructures remains tightly coupled with individual application requirements and technological choices. The exact development of sensing infrastructures is thus out of the scope of this chapter.

Context-Aware Behaviors in Web applications

But what exactly does it mean to adapt a Web application or to react to context? Starting from the work by Brusilovsky (1996) on adaptive hypermedia systems, in context-aware Web applications, adaptive behaviors may affect:

- *Contents and services* delivered by the accessed pages: the application may autonomously chose contents or services based on changing context data.
- The *navigation:* the application may perform automatic navigation actions on behalf of the user toward pages that better suit the current context conditions.
- The whole *hypertext structure:* the application may choose to apply coarse-grained adaptations (e.g. to the layout of the application), for example to react to changes of the user's device, role, or activity within a multi-channel, mobile environment.
- *Presentation properties*: the application may apply more fine-grained adjustments to the application's appearance (e.g. to style properties or fonts in use).
- *Generic operations*: the application may decide to enact generic operations in the background, e.g. to log specific application events or to interact with external applications.

In this chapter, we will describe how these behaviors have been realized in the model-driven design method WebML and how the resulting extended version of the method can be leveraged for the development of context-aware applications. Before proceeding with the discussion, it is thus appropriate to shortly introduce the WebML development method, which will serve as reference throughout this chapter.

The Web Modeling Language (WebML)

WebML is a visual language for specifying the content structure of Web applications and the organization and presentation of contents into one or more hypertexts (Ceri et al., 2002).

WebML application design starts with the specification of a *data schema*, expressing the organization of the application contents by means of well established data models, such as the Entity-Relationship model or the UML class diagram. On top of such data schema, WebML design then proceeds with the specification of a so-called *hypertext model*, which describes how contents, previously specified in the data schema, are published into the application hypertext. The overall structure of the hypertext is defined in terms of *site views*, *areas*, *pages*, and *content units*. A *site view* is a hypertext, designed to address a specific set of requirements. Several site views can be defined on top of the same data schema, for serving the needs of different user communities, or for arranging the composition of pages to meet the requirements of different access devices like PDAs, smart phones, and similar appliances. A site view is composed of *areas*, which are the main sections of the hypertext, and comprise recursively other sub-areas or pages. *Pages* are the actual containers of information delivered to the user; they are made of *content units*, which are the elementary pieces of information extracted from the data sources by means of queries, and published within pages. In particular, content units denote alternative ways for displaying one or more entity instances. Unit specification requires the definition of a *source* and a *selector*: the source is the name of the entity from which the unit's content is extracted; the selector is a condition, used for retrieving the actual objects of the source entity that contribute to the unit's content. Content units and pages are interconnected by *links* to constitute site views. Besides representing user navigation, links between units also specify the transportation of parameters that can be used by the destination unit in its selector condition. Some WebML units also support the specification of content management operations. Standard operations are creating, deleting or modifying an instance of an entity or adding or dropping a relationship between two instances; custom units may be defined. Finally, WebML also allows the management of *session parameters*; parameters can be set and consumed through proper units.

In addition to the visual representation, WebML also comes with an XML-based, textual representation, which allows one to specify additional detailed properties, not conveniently expressible in the graphic notation. The availability of the XML specification enables the automatic generation of the application code (Web Models, 2008), comprising rendering formats like HTML (which is the standard choice for deployment) or WML. For a detailed description of WebML, the interested reader is referred to (Ceri et al., 2002).

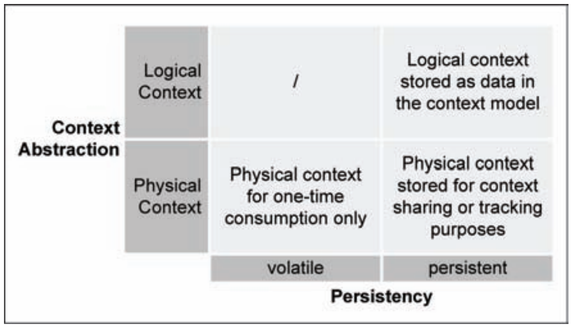
MODELING DATA FOR CONTEXT-AWARE WEB APPLICATIONS

Context data can derive from several sources integrating sensed, user-supplied, and derived information (Henricksen, 2004; Henricksen, 2002). While user-supplied data are generally reliable and tend to be static, sensed data are typically highly dynamic and can be unreliable due to noise or sensor errors. The problem of unreliability has been addressed in literature for example by associating context information with quality data (Lei, 2002). Although we recognize the importance of reliable context data, in this work we rather concentrate on the exploitation of context in the design of Web applications. For simplicity, throughout this chapter we thus consider sensed data as trustworthy.

Characterizing Context Data

The main goal of context modeling is the formalization and abstraction of the context properties that affect the application. In this regard, a first characteristic distinguishing context properties is the distinction between physical and logical context. We call *physical* context those properties that are immediate representations of sensed, physical quantities (e.g. the values of an analog/digital converter), and *logical* context those properties that enrich physical context with semantics and additional abstractions of the raw sensed data (e.g. the city corresponding to physical longitude and latitude values).

Figure 2. Persistence of physical and logical context data



A second characteristic affecting the structure of the context model is the *persistence* of context properties in the system, i.e. the property that expresses whether individual context properties represent persistent data or volatile data. *Persistent* data need to be stored in the application's data source and therefore require proper data entities being modeled as part of the context model, while *volatile* data do not need any storage and can thus be omitted from the context model. The context model therefore only captures persistent context data (indeed, in WebML the context-model is part of the database underlying the application).

Starting from these two characteristics and from the reference architecture introduced in Figure 1, Figure 2 summarizes the resulting characterization of context data:

- *Volatile physical context*. Context data communicated via client-side parameters or via server-side session parameters represent volatile data. They are immediately available during the execution of the application, independently of the underlying context model. Volatile context data are not enclosed in the context model; they might however be used during page computation to adapt the application.
- *Persistent physical context*. Context data sharing (e.g. between members of a same group) or tracking (e.g. to derive differential context properties or to keep a context history) typically require the persistent storage of data. Persistent physical context data are thus included in the context model and updated according to their dynamics.
- *Persistent logical context*. Logical context data is stored as data in the context model, so as to enable the data-driven transformation of physical context into logical context. Logical context data are typically static, as they provide abstractions of physical context; dynamic updates and/or extensions can, however, be supported as well.

Physical and logical context data therefore coexist in the application's data source. This coexistence typically requires a transformation or mapping between raw data and information that can directly be used when specifying hypertext schemas. Consistently with the data-driven approach that characterizes WebML, we propose a formalization of such transformation at the data level by means of suitable associations between data entities representing physical and logical context data, respectively. Although technically legal, we do not expect the use of *volatile logical context*, as volatile context data typically represents sensed raw context data.

It is worth noting that even though there are several properties commonly regarded as *context attributes* (e.g. position, time, or device characteristics), there exists no universal context model that applies to all kinds of applications. For this reason, also in this chapter we do not prescribe any precise, rigid context model for WebML applications; we rather introduce some WebML-specific modeling guidelines that enable the designer to provide context-aware applications with suitable context meta-data.

Example Data Schema for Adaptation in WebML

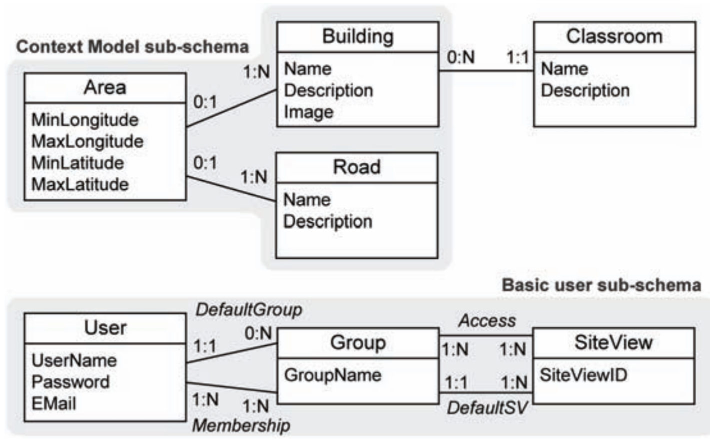
Let's consider the PoliTour application shortly discussed in the introduction. Figure 3 illustrates a possible Entity-Relationship diagram with basic user profile data and context data, grouped in the figure into so-called sub-schemas:

- *User profile sub-schema*. Users, groups, and site views are represented as "first-class citizens" in the application data source, as required by the WebML design process. The entity User provides a basic profile of the application's users, the entity Group associates access rights to users (i.e. a role), and the entity Site View contains the site views that may be accessed by the members of a group. The relationship Membership expresses that users may belong to multiple groups, which in turn cluster multiple users. The relationship DefaultGroup connects a user to his/her default role and, when logging into the application, the relationship DefaultSV allows the application to forward the user to his/her default group's default site view. The relationship Access expresses which site views a specific group is allowed to access; this relationship is required as varying context conditions may require different interaction and navigation structures for a same group. In this way, depending on the context state, the application is able to determine the most appropriate site view and to forward the user accordingly.
- *Context model sub-schema*. The context model of the application is represented by the entities Area, Building, and Road, which all provide logical context data. The actual GPS position data used for delivering the location-aware guide through the Politecnico campus (i.e. longitude and latitude) and the signal strength of the WiFi connection are not part of the context model in the application's data source; in developing the PoliTour application, we will handle such as volatile context data. Starting from the physically sensed data, the entity Area allows the application to identify a geographical area inside the campus; an area is then associated either with a Building or a Road, meaning that starting from the user's position we can identify whether he/she is located close to a building or rather walking through one of the roads in the campus.
- *Application data*. The remaining entity Classroom represents application data that are not part of the context model. This means that from a building it is possible to access the list of classrooms of the building, but there are no adaptive behaviors associated with the entity Classroom.

MODELING CONTEXT-AWARE HYPERTEXTS

While the first step of the WebML design method, i.e. data modeling, does not require any extension of the modeling primitives for capturing context data (the standard Entity-Relationship primitives suffice), WebML hypertext modeling does require a few model extensions to express adaptivity concerns. Next we therefore introduce the new concepts and primitives that have been developed to express adaptive behaviors, and we clarify how different adaptivity policies can be used to enact adaptations.

Figure 3. Adaptation-triggering data in WebML applications, partitioned into basic user sub-schema, personalization sub-schema and context sub-schema

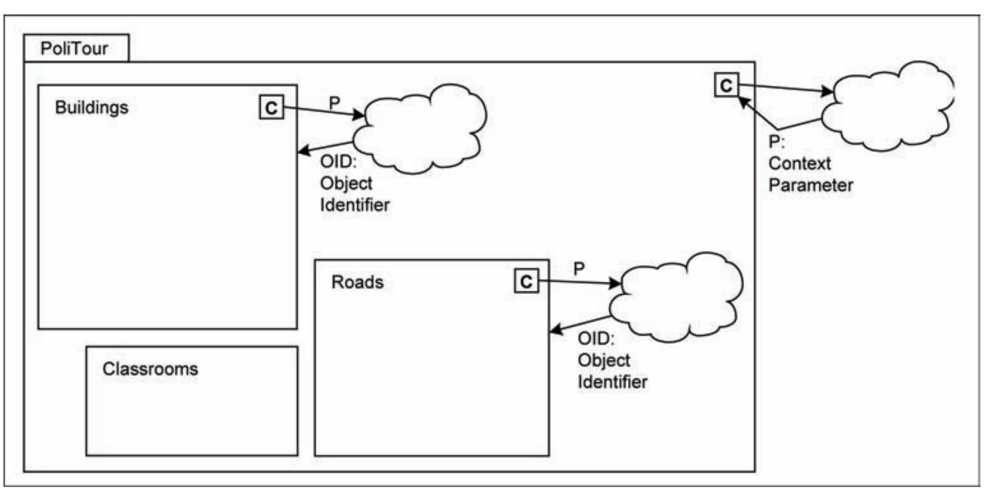


Context-Aware Pages and Containers

Our basic assumption in the modeling of context-aware hypertexts is that context-awareness or adaptivity is a property to be associated only to some *pages* of an application (the *adaptive hypertext*), not necessarily to the application as a whole. Location-aware applications, for example, adapt core contents to the position of a user, and so-called "access pages" (e.g. containing categories or lists) typically are not affected by the context of use.

As can be seen in Figure 4, we tag context-aware pages with a C-label (standing for *context-aware*) to distinguish them from conventional pages. The label indicates that an adaptivity rule (stylized as a cloud) is associated with the page and that during the execution of the application this logic must be taken into account when com-

Figure 4. WebML hypertext schema with one context-aware site view and two context-aware pages. The parameter P exemplifies the propagation of reusable context data by hierarchically passing context parameters from an outer area to an inner page



puting the page. Specifically, Figure 4 states that pages Buildings and Roads are context-aware, while the page Classrooms does not present any adaptive behavior.

There might also be the need for adaptivity rules with effects that spread over multiple pages. For this purpose, we exploit the hierarchical structure of hypertexts; that is, we allow the definition of context-aware *containers* (i.e. *site views* and *areas*, in terms of WebML). This allows the designer to insulate and to specify only once adaptivity rules that are common to multiple C-pages inside a container and thus to reduce the redundancy of the schema. Adaptivity rules associated to containers and pages are evaluated recursively, starting from the outermost container and ending with the actual pages. The site view PoliTour in Figure 4 is context-aware; we will see later on why.

Localized and Sparse Adaptivity Rules

The adaptivity rules attached to the context-aware pages and containers in Figure 4 represent the actual adaptivity logic (i.e. the set of adaptivity actions to be performed). The adaptivity logic is external to the page or container, and the chain of adaptivity actions it clusters is kept separate from the page or container specification. The aim is to highlight the two different logics deriving from the role played by pages/containers and adaptivity operations: while the former act as *providers* of contents and services, the latter act as *modifiers* of such contents and services.

Adaptivity actions attached to a C-page typically present effects that are visible in the page they are attached to. The notion of context-aware page and adaptation logic therefore defines what we call a *localized adaptivity rule*: the scope of a localized adaptivity rule is strictly coupled with a fixed set of hypertext pages, where "scope" refers to those (adaptive) pages to which the page's adaptivity actions are associated.

The notion of context-aware container allows us to define *sparse adaptivity rules*: we talk about *sparse* adaptivity rules in those cases, where adaptivity actions are associated to containers that contain multiple pages; the scope of such actions spans a set of pages, more precisely, all context-aware pages in the container. Coming back to the PoliTour application sketched in Figure 4, we can thus associate the logic to interpret the signal strength of the WiFi connection to the pages Buildings and Roads by applying the logic to the site view as a whole.

Parameter Passing

Adaptivity logic is associated to a page by means of a directed arrow, i.e. a link exiting the C-label. This link ensures the communication between the page logic and the adaptivity logic: it may transport parameters deriving from page contents, which may be used to compute the specified actions; in turn, a link from the adaptivity logic to the page may transport context parameters or generic values that might be required to perform the final adaptation during page computation.

But Figure 4 also illustrates the possibility of *hierarchically* passing parameters from an outer container to an inner one. More precisely, if the evaluation of outer adaptivity logic produces results to be reused at an inner level, as it might happen in the case of context parameters, it passes such values back to the C-label that activated the computation of the logic. Subsequently, such parameters can then be "consumed" by adaptivity logics of the inner levels. As for context-aware pages, parameter passing from a container to its adaptivity logic occurs through the logic-activating link. Links exiting the last evaluated logic, i.e. at the end of the last adaptivity action, might carry parameter values for the computation of units inside a page.

Typical actions to be specified at the container level are the acquisition of fresh context data and the updating of the context model, e.g. if the data are to be shared among multiple users or if a history of context data is to be tracked. Hence, especially if persistent context data are adopted, we propose two levels for adaptivity actions:

- *Actions for context model management*, addressing operations for context data acquisition and context model updating, should be associated with outer containers (site views or areas) and are inherited by inner containers (areas or pages). These adaptivity actions need to be executed prior to the execution of any other action possibly specified in an inner context cloud, as such "internal" actions could depend on context data acquired and stored in the context model through "external" actions.
- *Actions for hypertext adaptivity*, defining the rules for page and navigation adaptation (and possibly depending on persistent context data), should be associated with C-pages.

Specifying Adaptivity Logics

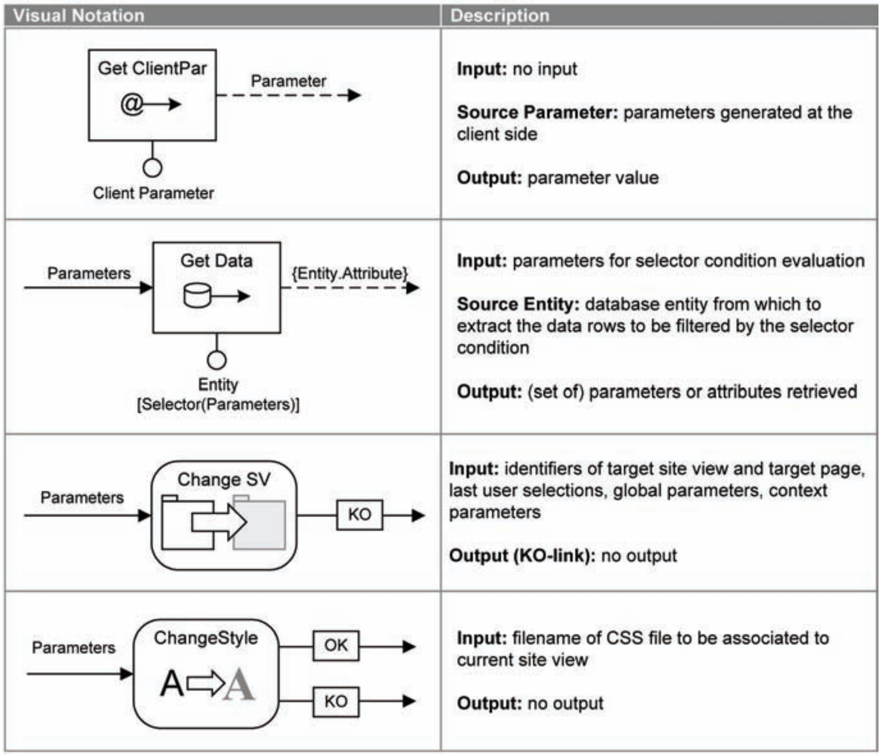
The main novelties for modeling context-aware pages reside in the specification of adaptivity rules by means of WebML constructs. In the following, we introduce the new WebML modeling concepts that ensure full coverage for the specification of context model management and hypertext adaptation logics. The new primitives allow designers to visually specify actions for acquiring and updating context data and to define adaptivity actions.

Managing Context Data

In order to support adaptivity with respect context, the application must be able to acquire and manage context data according to the mechanisms illustrated in Figure 1. For this purpose, some new WebML operations have been defined, which, together with the already available operations, provide the necessary primitives for:

- *Specifying the acquisition of fresh context data through client-side parameters*. A new Get ClientParameter unit (see Figure 5) has been defined to support the retrieval of parameters generated at the client side and communicated back to the application via client-side parameters (e.g. parameter-value pairs attached to the page request query string).
- *Specifying the acquisition of fresh context data through server-side parameters*. Context data directly made available as HTTP session parameters can be accessed by means of conventional WebML Get units (Ceri et al., 2002).
- *Specifying the acquisition of context data from the context model*. The execution of adaptivity actions may require the retrieval and evaluation of context meta-data, for example, in situations where certain data are just needed to evaluate condition expressions. For this purpose, a so-called Get Data unit (see Figure 5) has been introduced, enabling the retrieval of values (both scalars and sets) from the data source according to a selector condition. The semantics of the Get Data unit is similar to the one of content publishing units (Ceri et al., 2002), with the only difference that data retrieved from the data source are not published in hypertexts, but just used as input for units or operations.
- *Updating the context model*. Once fresh context parameters have been retrieved, they can be used to update the context model at data level. This action consists in modifying values previously stored in the data source. In WebML, this is already facilitated by operation units (Ceri et al., 2002) providing support for the most

Figure 5. WebML units that have been defined for the specification of adaptivity actions



common database management operations (e.g., modify, insert, delete).

Evaluating Conditions

The execution of adaptivity actions may be subject to the evaluation of some *conditions*, refining the triggering logic for context clouds. The most recurrent pattern consists in evaluating whether context changes demand for adaptation. The evaluation of conditions is specified by means of two control structures, represented by the If and Switch operation units, which have been introduced for workflow modeling in WebML (Brambilla et al., 2003).

Executing Adaptivity Actions

Once the current context state has been determined, and possible conditions have been evaluated, adaptivity actions can be performed to adapt the page contents, the navigation, the current site view structure, and/or presentation style properties. These actions are specified as follows:

- *Adapting Page Contents*. Page contents are adapted by means of proper data selectors, whose definition is based on context parameters retrieved from the context model or newly computed within the page's context logic. The use of parameterized selectors allows for both *filtering* data items with respect to the current context state and conditionally *including/excluding* (i.e. showing/hiding) individual content units.
- *Adapting Navigation*. In some cases, the effect of condition evaluation within the context cloud can be an automatic, i.e. context-triggered, navigation action, causing the redirection of the user to a different

page. The specification of context-triggered navigations just requires connecting one of the links exiting the adaptivity logic of the page to an arbitrary destination page of the hypertext. Therefore, links exiting the context cloud and directed to other pages than the adaptivity logic's source page represent automatic navigation actions.

- *Adapting the Site View*. In some cases, a context-triggered switch toward a different site view may be required. Changes in the interaction context may in fact ask for a coarse-grained restructuring of the whole hypertext, for example because the user device has changed, or because the user shifted to a different activity. To switch between different site views, we have introduced a Change Site View unit (see Figure 5), which takes in input the identifiers of the target site view and the target page, to be visualized in case a switch toward the specified site view is required. In order to support "contextual" switching, the input link also transports parameters characterizing the current state of interaction, i.e.:
    1. The input parameters of the source page, which represent the last selections operated by the user;
    2. Global parameters, representing session data (e.g. user identifier and group identifier), as well as past user selections that have been used for the computation of the current page;
    3. Client-side and server-side context parameters retrieved during the latest performed data acquisition cycle and characterizing the current context state.
- *Adapting Presentation Style*. Sometimes context changes may require only fine-grained adaptations of presentation properties (e.g. due to varying luminosity conditions), not a complete restructuring of the overall hypertext. We have defined a Change Style unit for dynamically assigning presentation style properties (see Figure 5). Style properties are collected in proper .css (Cascaded Style Sheet) files, and the unit enables the application to change its associated style sheet at runtime.
- *Enacting generic operations*. The context-triggered invocation of generic operations or, for instance, external Web services can easily be specified by placing the respective WebML operation unit into the page's adaptivity logic and by providing the unit with the necessary input parameters.

Triggering Adaptivity Rules

But *when* do we enact an adaptivity rule? In this regard, it is possible to define two different *adaptivity policies* for context-aware pages, assigning different priorities to users and context:

- *Deferred Adaptivity*: the *user* is granted the highest priority. Therefore, after the user has entered the page and the page has been rendered according to the user's selections, the page's adaptivity logic is evaluated at periodic time intervals, enabling the application to possibly adapt the already rendered page. Periodically evaluating the adaptivity logic means periodically refreshing the page visualized in the browser.
- *Immediate Adaptivity*: *context* is granted the highest priority. The page's adaptivity logic is evaluated each time the page is accessed, being the access due to the user or to the periodic refresh of the page. This means that the page is subject to adaptation each time it is rendered, even at the first time the page is accessed by the user.

Consider for example our PoliTour guide that shows contents about the buildings and roads in the Politecnico campus. At a given point, the user might want to get information about a specific building located in a road that is not related to his/her current position; such a preference is typically expressed by selecting a link to that building from a list. With a deferred policy, the requested page shows the building information as requested by the user, without taking into account the user's current location. Only after expiration of the refresh interval, the page becomes subject to adaptivity and the contents are adapted to the user's location. With an immediate policy, context is granted higher priority with respect to the user and, thus, the user's request for the building would be overwritten by the context and the application would show the building or road associated to the user's current location, discarding the user's selection.

Note that in addition to these adaptivity policies, we recognize that there may be situations that demand for an explicit control of the adaptation dynamics through the user. Therefore, should for example a user temporarily not be interested in having the contents adapted to his/her location, he/she can simply disable/enable adaptivity at will. In WebML, the adaptivity policy for context-aware pages and containers is declared by means of the Adaptivity_Policy property of context-aware pages and containers.

Adaptivity policies can also be associated to context-aware containers. When a C-page is requested, also the possible context clouds of its containers are evaluated recursively (from the outermost one to the innermost one), according to the adaptivity policy associated to each container. In general, a container's adaptivity policy is independent of the policy of inner containers and pages (if not, this must be taken into account by designers when associating policies to containers and pages). Therefore, it may happen that the actions in a container's context cloud are evaluated immediately, even if the actions associated to inner containers or pages adopt a deferred evaluation, or vice-versa. If, for example, the adaptivity actions associated to the container serve for tracking a context history, they could require an immediate policy, while inner adaptivity actions keep their deferred policy for front-end adaptations. The hierarchical definition of context clouds may therefore also be considered a facility to achieve different "layers" of adaptivity actions.

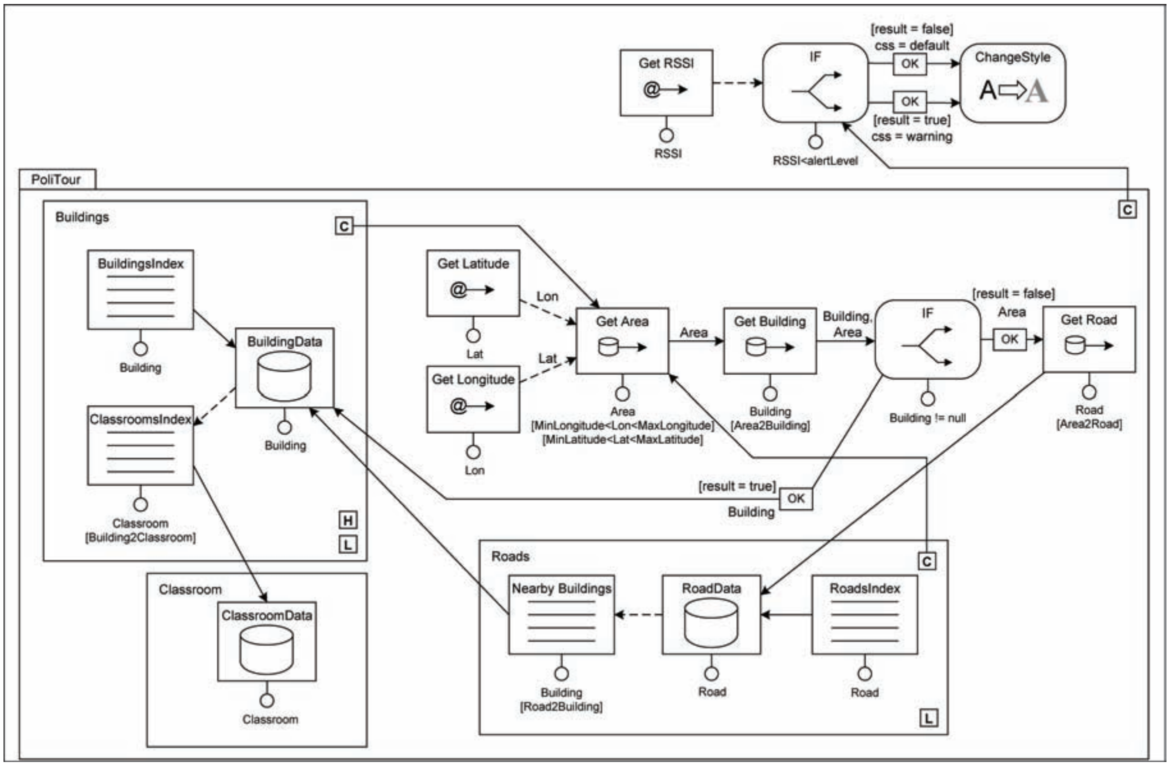
In our approach, we assume *deferred* adaptivity as default policy. This choice aims at minimizing application behaviors that might be perceived as invasive or annoying by users and has been experienced as the most natural for modeling adaptation. However, the *immediate* policy could be needed for handling exceptional situations, as in such cases the timely reaction to context changes could be more important than following the user's indications. We therefore, in general, recommend the selection of the adaptivity policy that is appropriate to the application requirements and that is able to minimize the application behaviors that could be perceived as invasive or annoying by the users. In order to choose the right adaptivity policy for an adaptive page, a developer therefore needs to predict what kind of adaptive behavior a user will expect when accessing that page.

Example Hypertext Model

Figure 6 shows the adaptive WebML hypertext model of the PoliTour application. The figure provides a refinement of the coarse hypertext model introduced in Figure 4 and details the internals of pages and adaptivity logics.

The pages Buildings and Roads share the same adaptivity logic providing location-awareness to the displayed contents. The logic starts with two Get ClientParameter units accessing the user's longitude and latitude, which are then used by the Get Area unit to associate a logical area to the user's position. A further Get Data unit (the Get Building unit) then tries to retrieve a building for the identified area. If a building could be retrieved, the If unit sends the user to the Buildings page, providing updated page parameters. If instead no building could be retrieved (e.g. because the user is located in the center of a road or not close

Figure 6. Hypertext model of the PoliTour application leveraging volatile context data



enough to a building), the If unit forwards the Area identifier to the Get Road unit, which retrieves the road associated to the current position.

Therefore, if the user views the page Buildings while walking around the campus, the application automatically updates the contents published each time a new building can be found. If only the road can be identified, the application performs an automatic navigation action toward the Roads page, where the described adaptive behavior starts again, possibly causing the adaptation of contents or automatic navigation actions. Only if the user navigates to page Classroom, no adaptations are performed, as this page is not tagged as context-aware.

The adaptivity actions associated to the surrounding site view specify how to alert users who are about to leave the WiFi-covered area. The Get RSSI unit accesses the volatile RSSI parameter sensed at the client side, and the If unit compares the retrieved value with a predefined level (alertLevel), below of which the connectivity is considered low. In case of low connectivity, the style sheet warning is adopted; otherwise, the default style sheet is adopted. We therefore model the alert of low connectivity conditions by means of a Change Style unit: under low connectivity conditions the application is rendered with a red background, under normal conditions the application is rendered with a gray background.

We recall that actions associated to containers are evaluated before any action at the page level is started. Hence, in Figure 6 the actions associated to the site view are executed before the actions associated to the pages Buildings and Roads.

RUNTIME CONTEXT MODEL MANAGEMENT

In order to manifest context-aware behaviors, the application must be equipped with the capability to monitor the context state and to trigger adaptivity actions, if required. The standard HTTP protocol underlying most of today's Web applications implements a strict *pull* paradigm, in which computations can only occur in response to client-side generated page requests. Therefore, in the classical Web architecture, lacking proper push mechanisms, context monitoring can occur only when a page is computed, i.e. when a respective page request reaches the Web server. Three main solutions can be adopted to trigger the evaluation of adaptivity rules: (i) context evaluation on user-generated page requests, (ii) periodical, automatic refreshes of viewed pages to enable context evaluation, and (iii) active context evaluation to trigger adaptivity in real time. The first solution is not able to cope with the dynamic nature of context. The periodic refresh of context-aware pages provides a way to ensure the update of the page even in absence of explicit user actions enabling the re-computation of the page. In the following, we will show an active mechanism for triggering adaptivity, which operates independently of the user in the background and comes close to the real-time triggering solution.

In absence of dedicated server-side *push* mechanisms for delivering updated pages, the HTML http-equiv META-option or JavaScript, JavaApplets, and Flash scripts, provide valuable client-side mechanisms to approximate the required active behavior. The approximation is based on periodic HTTP requests toward the application server, which are operated in the background and may serve a twofold purpose:

- On the one hand, they provide the necessary polling mechanism to query the context model and trigger the adaptivity rule attached to the page.
- On the other hand, generating page requests allows the client to transmit client-side sensed data, thus enabling the communication of context data to the application server.

Context-aware pages are therefore also characterized by an individual *refresh interval*, which can be specified as property (Refresh_Interval) of the page in the XML representation of the WebML model. Differently from C-pages, a container does not require the specification of any polling interval, which is instead derived from the interval associated to the currently viewed C-page of the container.

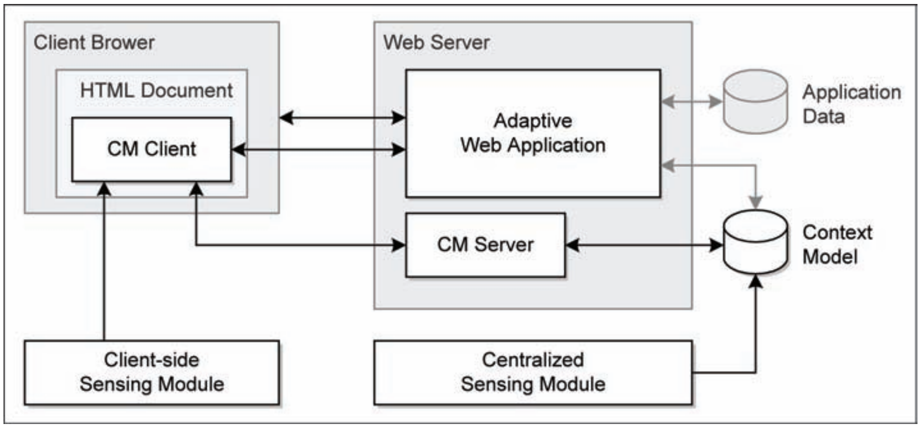
Context Monitoring

Context monitoring in the background (i.e. without the user observing any unwanted rendering activity) enables the application to limit the use of the refresh to those situations that really ask for adaptation and to perform context monitoring without any visual effect for users.

Figure 7 shows a functional architecture for adaptive Web applications that extends the described architecture of WebML applications (see Figure 1) with a new client-server module, called *Context Monitor* (CM), providing the necessary context monitoring logic. As further depicted by the figure, in case of client-side context sensing, the CM module also enables the communication of client-side sensed context parameters, which could be required at the server side to evaluate context changes and/or conditions over context parameters.

The CM consists of two separate modules, one on the client side and one on the server side. The CM Client module is a piece of business logic embedded into the page's HTML code and executed at the client side (e.g. a JavaScript function, a Java applet, or a Flash object), while the CM Server module works in parallel to the Web application on the same Web server. The CM

Figure 7. Functional architecture for background context monitoring



Client is in charge of periodically monitoring the context state and deciding whether possibly occurring context variations demand for the adaptation of the currently viewed page.

In order to be able to take a decision about whether adaptivity actions are to be triggered or not, the CM Client is assisted by the CM Server, which has full access to the context model of the application maintained at the server side. In response to the polling executed by the CM Client, the CM Server queries the context model and provides the CM Client with an updated picture of the effective context state. By comparing the state of the (server-side) context model acquired by the current polling with the one acquired by the last polling (or the state at page computation time), the CM Client knows whether the state has changed. If the state has changed, the CM Client asks the Web application for a refresh of the currently viewed page, i.e. the adaptation; if the state has not changed the CM Client proceeds with the monitoring of the context state.

Page Context

In general, the *state* of the context is expressed by the values of all the persistent parameters stored in the context model and of the volatile parameters sensed at the client or server side. However, an individual page's adaptive behavior is typically influenced by only a subset of the overall context data or, more specifically, by a function expressed over context data. The subset of context data corresponds to a page-specific view over the application's context data, narrowing the focus of the context monitoring activity. This observation leads to the definition of a new concept, i.e. page context, which can be leveraged to enhance the efficiency of the context monitoring activity: the *page context* of a page corresponds to a page-specific view over the application's context data, capturing all (and only) those context characteristics that effectively determine the adaptive behavior of the page.

Instead of monitoring the whole state of the application's context data, the definition of a page context for each adaptive page enables the context monitoring activity to focus its observation of the context state to the only page context. This implies, that during hypertext specification each adaptivity rule can be related to a subset of context parameters to be controlled, so that rule conditions do not need to check the state of the whole context model.

Page Context Parameters

In line with the idea of page context, the CM focuses its attention only to the subset of context data in the context model that really determines the adaptive behavior of the viewed page. This implies explicit knowledge of the pages' page context, which can be achieved by defining proper page context parameters for each context-aware page: p*age context parameters* define the view over the context model that captures all the static and dynamic properties of a page's page context by means of suitable queries over the context model.

This definition implies that each change to a page context parameter effectively corresponds to a need to adapt the page. The granularity of the *values* of page context parameters must thus be chosen in a way that each change of a parameter value translates into the triggering of the page's adaptivity rule. Each C-labeled page in the adaptive hypertext model is thus associated with an individual page context by means of proper page context parameters stored in the textual representation of the WebML schema, as they are not conveniently expressible in a visual manner. Page context parameters are expressed by means of parametric queries over the context data, where the parameters correspond to client- or server-side context parameters.

Context Digest

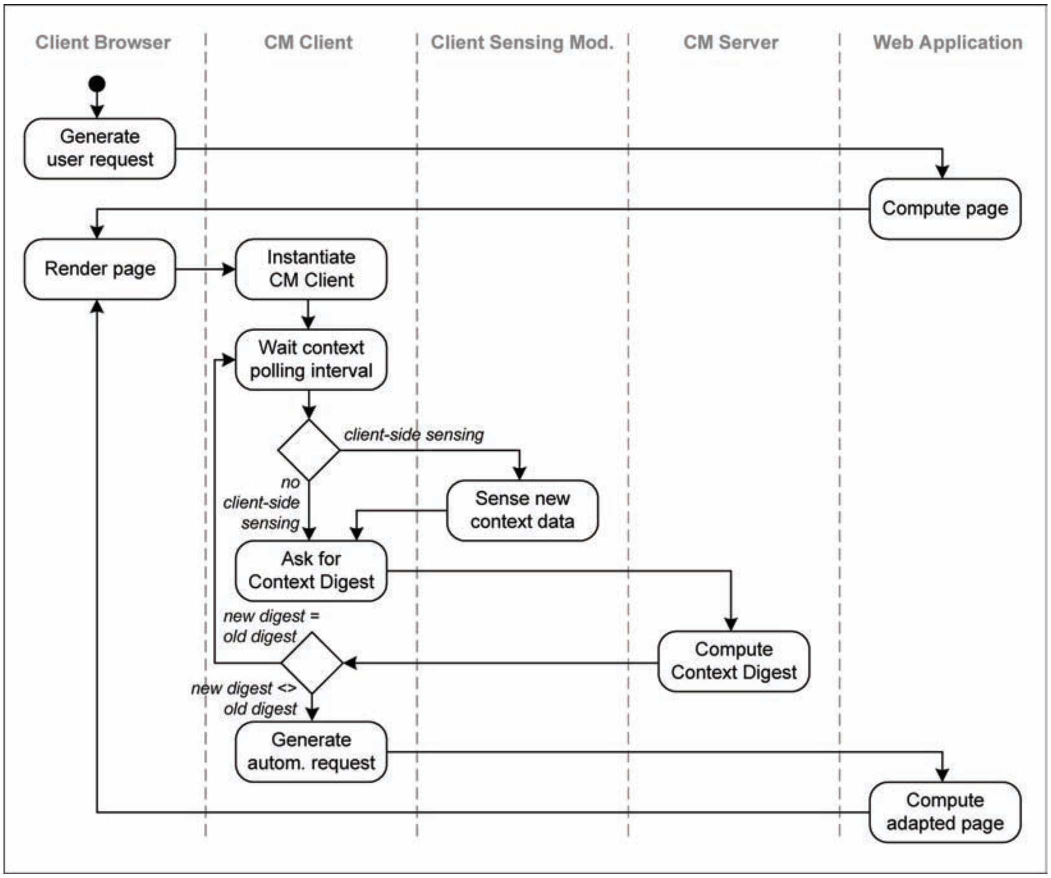
In order for the CM to be able to decide whether adaptivity is required, changes to the page context (i.e. the page context parameters) must be communicated from the CM Server to the CM Client. In order to enhance the efficiency of the overall context monitoring activity, the state of the page context is not communicated from the CM Server to the CM Client in form of the set of page context parameters, but instead it suffices to transmit and compare a numeric digest computed over the respective page context parameters, as each change to the values of the page context parameters also results in a change of the numeric digest. We call such a numeric digest context digest: the *context digest* corresponding to the page context of a page is the numeric checksum computed over the ordered list of page context parameters.

The context digest is the basis for the decisions to be taken by the CM Client: its values identify variations in the page context, which correspond to the need to adapt the page. The decision is based on the comparison of the current context digest with the last context digest; the first context digest, i.e. when the user accesses the page, is initialized with the context digest valid during page computation.

Figure 8 details the resulting flow of activities enabling the active behavior of the application and shows how the single modules cooperate in order to determine whether adaptivity is required or not. The diagram has one start node (Generate user request), which corresponds to the user's navigation to a C-page, and no end node, since the cycle in the lower part of the diagram is only interrupted by an explicit user navigation leading the user to another C-page (which corresponds to starting again from the start node of the diagram and monitoring the Page context of the new page) or to a conventional page (which does not cause any context monitoring activity).

Note that the described mechanism assumes that connectivity is available during the viewing of a C-page in order for the CM client to be able to communicate with the CM server. In case of intermittent connectivity, which is a very frequent situation in mobile environments, the CM client keeps working by periodically polling the CM Server, despite the absence of connectivity. The CM Client is however programmed to manage possible lacks of connectivity and therefore does not generate errors, with the only side effect that adaptivity is suspended until the connectivity is restored.

Figure 8. Background context monitoring for active context-awareness (with client-side context sensing): communicating context data and triggering adaptivity

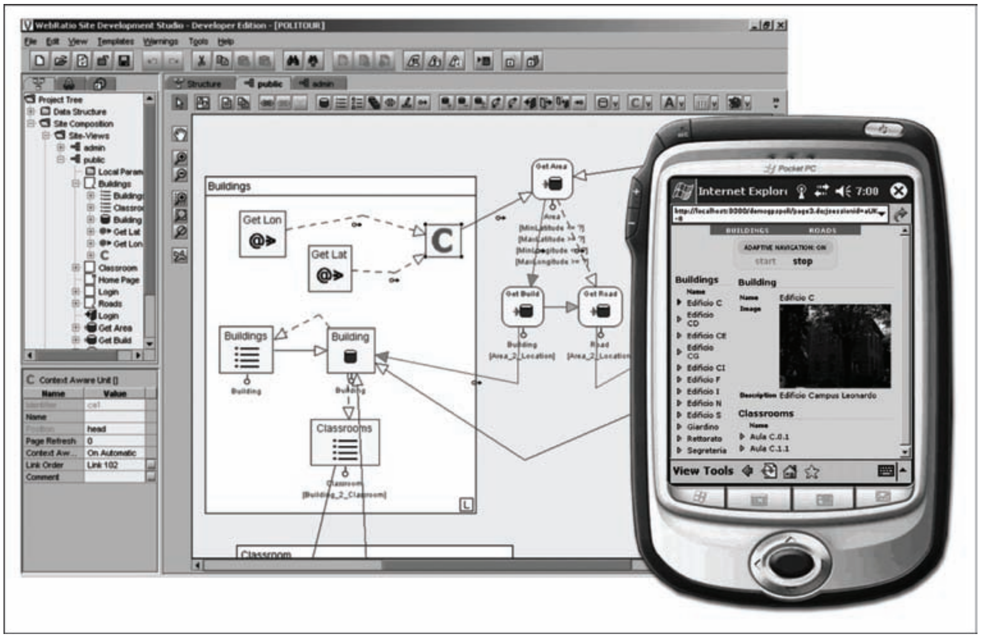


APPLICATION IMPLEMENTATION

The extensions that have been introduced into the WebML development method to cope with the new requirements posed by context-awareness and adaptivity in Web applications have been implemented as prototype extension of the WebRatio CASE tool, the official WebML modeling tool, equipped with a powerful automatic code generator. Due to implementation restrictions imposed by the modeling tool, the implementation of the adaptivity logic slightly differs from the models described in this paper (e.g. it was not possible to implement context-aware containers or to place all the adaptivity operations outside pages). Nevertheless, the described expressive power for the specification of adaptivity rules could be preserved.

Figure 9 shows a screenshot of the WebRatio tool at work. The figure shows the WebML hypertext model of the Buildings page of the PoliTour application, along with its adaptivity logic: two Get ClientParameter units access the GPS coordinates and pass them to the C-label, which forwards them to the outer adaptivity logic (cf. Figure 6). Starting from the shown hypertext model, the PoliTour application has been automatically generated on top of a J2EE platform. The configuration of the Context Monitor has been performed manually. To access GPS position data, we leverage a client-side Bluetooth GPS device, interfaced via the Chaeron GPS Library (http://www.chaeron.

Figure 9. The WebRatio CASE tool showing the hypertext model of the buildings page with respective adaptivity actions and the generated PoliTour application running on a PDA



com/gps.html) and wrapped by means of Flash (to exchange position data between the CM Client and the GPS library). The WiFi RSSI indicator is acquired in the PDA using Place Lab (http://www.placelab.org).

RELATED WORKS

Several other well-established, conceptual design methods have been so far extended to deal with Web application adaptations. Frasincar & Houben (2002), for example, extend the Hera methodology with two kinds of adaptation: adaptability with respect to the user device and adaptivity based on user profile data. Adaptation rules (and the Hera schemas) are expressed in RDF(S) (Resource Description Framework/RDF Schema), attached to slices and executed by the AHA engine (De Bra et al., 2003). The UWA Consortium proposes WUML (Kappel et al., 2001) for conceptual hypertext design. Adaptation requirements are expressed by means of OCL-based customization rules, referring to UML class or package elements. Casteleyn et al. (2003) present an extension of WSDM (De Troyer & Leune, 1998) to cover the specification of adaptive behaviors. In particular, an event-based Adaptive Specification Language (ASL) is defined, which allows designers to express adaptations on the structure and the navigation of the Web site. Such adaptations consist in transformations of the navigation model that can be applied to nodes (deleting/adding nodes), information chunks (connecting/disconnecting chunks to/from a node), and links (adding/deleting links). Baumeister et al. (2005) explore Aspect-Oriented Programming techniques to model adaptivity in the context of the UML-based Web engineering method UWE. Recently, WebML (Ceri et al., 2002) has been extended to cover adaptivity and context-awareness (Ceri et al., 2007). New visual primitives cover the specification of adaptivity rules to evaluate conditions and to trigger some actions for adapting page contents, navigation, hypertext structure, and presentation. Also, the data model has been enriched to represent metadata supporting adaptivity.

Recently, active rules, based on the ECA (Event-Condition-Action) paradigm, have been proposed as a way to solve the previous problem. Initially exploited especially in fields such as content evolution and reactive Web (Alferes et al., 2005; Bailey et al., 2002; Bonifati et al., 2002), ECA rules have been adopted to support adaptivity in Web applications. In particular, the specification of decoupled adaptivity rules provides a way to design adaptive behaviors along an orthogonal dimension. Among the most recent and notable proposals, the work described in (Garrigos et al., 2005a) enriches the OO-H model with personalization rules for profile groups: rules are defined in PRML (Personalization Rule Modeling Language) and are attached to links in the OO-H Navigation Access Diagram. The use of a PRML rule engine is envisioned in (Garrigos et al., 2005b), but its real potential for adaptivity management also at runtime remains unexplored.

The previous works benefit from the adoption of conceptual models, which provide designers with powerful means to reason at a high-level of abstraction, independently of implementation details. There are however also co-called transcoding solutions, which adopt active rules for adapting Web pages. Most of them focus on the presentation layer and provide mechanisms to transform HTML pages according to (possibly limited) device capabilities (Hori et al., 2000) or users' visual disabilities (Yesilada et al., 2004). Moreover, they typically support only adaptability and modify Web pages in relation to a static set of user or device parameters. Fiala and Houben (2005) adopt the transcoding paradigm for the development of the Generic Adaptation Component (GAC). GAC provides a broad range of adaptation behaviors, especially supporting run time adaptivity. An RDF-based rule language is used for specifying both content adaptation and context data update rules. A collection of operations implementing these rules is provided. A notable feature, promoting portability, is that GAC can be integrated as a stand-alone module into any Web site architecture.

CONCLUSION AND FUTURE TRENDS

In this chapter, we have proposed a model-driven approach to the development of context-aware Web applications, an increasingly relevant kind of applications on the Web. We have shown that context-awareness is a first-class design concern that can considerably be aided by model-driven development techniques. But we have also shown that properly dealing with context-awareness and adaptivity at the conceptual level requires extending the expressive power of the adopted conceptual application model, so as to provide developers with suitable modeling constructs and implementation abstractions, proper of such new class of application features. In this chapter, such extensions have been introduced into the already well-established WebML modeling language, but in a similar way we could have also opted for another modeling language, as the ideas and concepts introduced in this chapter are general enough in nature to be applied to other conceptual models as well.

For the future, we believe that a *decoupled runtime management* of adaptivity features will represent a next step in the area of adaptive Web applications. The development of Web applications is more and more based on fast and incremental deployments with multiple development cycles. The same consideration also holds for context-aware and adaptive Web applications and their adaptivity requirements. In (Daniel et al., 2008) we describe our first results obtained with a decoupled environment for the execution and the administration of adaptivity rules. The described approach allows us to abstract adaptive behaviors, to extract them from the main application logic, and to provide a decoupled management support, finally enhancing the maintainability and evolvability of the overall application.

In line with the current hype of so-called Web 2.0 applications, we are also working on the *mash-up* of context-aware Web applications, in the context of our component-based development method for Web applications called Mixup (Yu et al., 2007). The final goal of the work is to enable even end users to mash up their own context-aware applications, starting from a set of so-called context components and other components equipped with own user interface (which is used to build up the user interface of the mash-up application). Mash-up development is assisted by an easy-to-use and intuitive graphical development environment that supports a drag-and-drop development and by a light-weight runtime environment that is able to interpret and run the mashup, both fully running in the client browser and based on AJAX technology.

REFERENCES

Alferes, J. J., Amador, R., & May, W. (2005). A General Language for Evolution and Reactivity in the Semantic Web. In *PPSWR'05* (pp. 101–115). Dagstuhl, Germany: Springer.

Bailey, J., Poulovassilis, A., & Wood, P. T. (2002). An Event-condition-action Language for XML. In *WWW'02* (pp. 486–495). Honolulu, Hawaii: ACM.

Baumeister, H., Knapp, A., Koch, N., & Zhang, G. (2005). Modelling Adaptivity with Aspects. In *ICWE'05* (pp. 406-416). Sydney, Australia: Springer.

Bonifati, A., Braga, D., Campi, A., & Ceri, S. (2002). Active XQuery. In *ICDE'02* (pp. 403-412). San Jose, California: IEEE.

Brambilla, M., Ceri, S., Comai, S., Fraternali, P., & Manolescu, I. (2003). Specification and Design of Workflow-Driven Hypertexts. *Journal of Web Engineering*, *1*(2), 163–182.

Brusilovsky, P. (1996). Methods and Techniques of Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, *6*(2-3), 87–129. doi:10.1007/BF00143964

Casteleyn, S., De Troyer, O., & Brockmans, S. (2003). Design Time Support for Adaptive Behavior in Web Sites. In *SAC'03* (pp. 1222-1228). Melbourne, Florida: ACM.

Ceri, S., Fraternali, P., Bongio, A., Brambilla, M., Comai, S., & Matera, M. (2002). *Designing Data-Intensive Web Applications*. San Francisco, CA: Morgan Kauffmann.

Ceri. S., Daniel, F., Matera, M., & Facca, F. M. (2007). Model-driven Development of Context-aware Web Applications. *ACM Transactions on Internet Technologies, 7*(1), article no. 2.

Daniel, F., Matera, M., & Pozzi, G. (2008). Managing Runtime Adaptivity through Active Rules: the Bellerofonte Framework. *Journal of Web Engineering*, *7*(3), 179–199.

De Bra, P., Aerts, A. T. M., Berden, B., de Lange, B., Rousseau, B., Santic, T., et al. (2003). AHA! The Adaptive Hypermedia Architecture. In *Hypertext'03* (pp 81-84). Nottingham, UK: ACM.

De Troyer, O., & Leune, C. J. (1998). WSDM: A User Centered Design Method for Web Sites. *Computer Networks, 30*(1-7), 85-94.

Dey, A. K., & Abowd, G. D. (2000). Towards a Better Understanding of Context and Context-Awareness. In *CHI'00 Workshop Proceedings*, The Hague, The Netherlands.

Fiala, Z., & Houben, G.-J. (2005). A generic transcoding tool for making web applications adaptive. In *CAiSE'05 Short Paper Proceedings*, volume 161 of CEUR Workshop Proceedings. CEUR-WS.org.

Frasincar, F., & Houben, G.-J. (2002). Hypermedia Presentation Adaptation on the Semantic Web. In *AH'02* (pp. 133-142). Málaga, Spain: Springer.

Garrigós, I., Casteleyn, S., & Gómez, J. (2005a). A Structured Approach to Personalize Websites Using the OO-H Personalization Framework. In *APWeb'05* (pp. 695-706). Shanghai, China: Springer.

Garrigós, I., Gómez, J., Barna, P., & Houben, G.-J. (2005b). A Reusable Personalization Model in Web Application Design. In WISM'05 (pp. 40-49). Sydney, Australia.

Henricksen, K., & Indulska, J. (2004). Modelling and Using Imperfect Context Information. In *PERCOMW'04* (pp. 33-37). Washington, United States: IEEE.

Henricksen, K., Indulska, J., & Rakotonirainy, A. (2002). Modeling Context Information in Pervasive Computing Systems. In *Pervasive'02* (pp. 167-180). London, UK: Springer.

Hori, M., Kondoh, G., Ono, K., Hirose, S., & Singhal, S. K. (2000). Annotation based Web Content Transcoding. *Computer Networks, 33*(1-6), 197-211.

Kappel, G., Pröll, B., Retschitzegger, W., & Schwinger, W. (2001). Modelling Ubiquitous Web Applications - TheWUML Approach. In *ER'01 Workshops* (pp. 183-197). Yokohama, Japan: Springer.

Lei, H., Sow, D. M., Davis, J. S. II, Banavar, G., & Ebling, M. R. (2002). The design and applications of a context service. *SIGMOBILE Mobile Computing and Communications Review*, *6*(4), 45–55. doi:10.1145/643550.643554

Schilit, B. N., & Theimer, M. M. (1994). Disseminating Active Map Information to Mobile Hosts. *IEEE Network, 8*(5), 22–32. doi:10.1109/65.313011

Web Models, l. (2008). WebRatio Site Development Studio. Retrieved January, 2008, from http://www.webratio.com.

Yesilada, Y., & Harper, S. Goble. C. A., & Stevens, R. (2004). Screen readers cannot see: Ontology based semantic annotation for visually impaired web travellers. In ICWE'04 (pp. 445-458). Munich, Germany: Springer.

Yu, J., Benatallah, B., Saint-Paul, R., Casati, F., Daniel, F., & Matera, M. (2007). A Framework for Rapid Integration of Presentation Components. In *WWW'07* (pp. 923-932). Banff, Canada: ACM.

This work was previously published in Context-Aware Mobile and Ubiquitous Computing for Enhanced Usability: Adaptive Technologies and Applications, edited by D. Stojanovic, pp. 59-82, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 2.11

Different Web Strategies for Different E-Marketplaces

L. Geppert
Catholic University of Milan, Italy

ABSTRACT

This chapter presents two possible models of electronic marketplaces put in place at the beginning of this century, which, after their introduction, the first incoming wave of connected economy-based paradigms was ended. Both the two models show a particular use of Web-based information technology in order to exploit their mission and represent meaningful cases of application of well-defined Web strategies. Even though, at the moment, the initially built Web sites supporting those related business have been closed and merged with other Web sites, they may introduce examples of a Web strategy approach having a relevant historical meaning that may be still redefined in practical implementations once revised and adequately updated. The mentioned cases described in this chapter are usteel.com and up2gold.com, two examples of Web-based business in two well-defined supply chains, that is, the "steel" chain and the "gold and silver" chain.

DOI: 10.4018/978-1-60566-024-0.ch006

INTRODUCTION

After the first incoming wave of the connected economy ended, investors, business managers, consultants, and specialists began to review their plans, actions, and methodologies. Many e-business models, both business-to-business (B2B) and business-to-consumer (B2C), are currently under consideration in order to assess and verify their real effectiveness (Kalakota & Whinston, 1996). The models applied so far have generally shown that, with some exceptions related to e-procurement implementations of large enterprises (Timmers, 1999), vertical marketplaces owners must look for different business paradigms in order to economically survive and to be actually able to create value to market industries chains they address (McDonald, 1997). In the following pages, two real cases are described: one trading marketplace addressing the steel industry as an example of “first wave” e-marketplace and one supply chain-like Web portal built for the Arezzo’s (Italy) gold manufacturing supply chain. The first case relates to a “traditional Web strategy”-based e-marketplace with a standard approach and contents (Shapiro & Varian, 1998) and the second shows an innovative strategy and a new business model (Porter, 1985). Even though, at the moment, both of the abovementioned e-marketplaces have merged their activities and brands in other larger e-commerce initiatives, at the time they were presented (2001), they represented important and innovative examples of an e-marketplace-based business model.

BACKGROUND

As some analyst observes, “the first high-visibility Internet businesses conceived to attract significant investor attention were, at the beginning, in the business-to-consumer (B2C) relationships, providing facilities for consumers to choose and purchase goods online. It seemed logical to think, as a related consequence, a development also of B2B, based on the “buying and selling” part of the relationship between businesses.

Thus, starting in 1999, and accelerating rapidly through 2000, the focus of B2B attention was very firmly on those initiatives that seemed to be capable of exploiting information and communication technology to alter the balance of power in trading relationships. These initiatives, which covered a wide spectrum of business and technical services, gained the general title of “marketplaces,” and, throughout 2000, the focus of B2B activities was on developing marketplace interaction strategies and examining procurement processes.

The focus on the transactional elements of B2B relationships will come to be seen as

an aberrant commencement to the development of the connected economy that will emerge in the 21st century. Businesses interact with other businesses in hundreds of ways. It is true that the interactions are generally intended to result in a commercial transaction. However, for many business relationships, the crucial performance-enhancing interactions take place before and after the transactions, and B2B will expand to occupy these broader horizons. (Kyte, 2001)

Business managers’ and investors’ expectations were very high at the end of the last century. They understood the strategic importance of B2B electronic model in particular and invested their money and their commitment, demanding rapid deployment of innovative business systems, which the IT community was increasingly hard-pressed to deliver. Many of these systems were new breed of applications, designed to change the external relations in the B2B world.

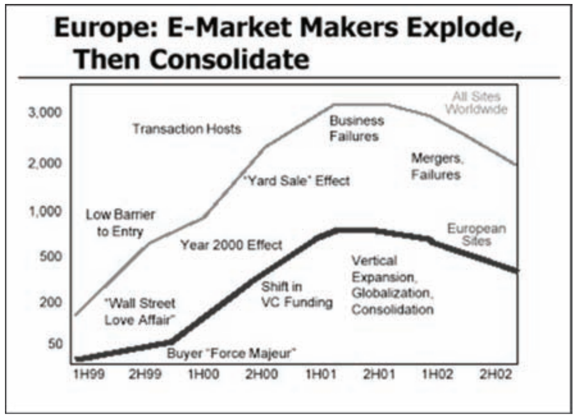
These systems work fine on the drawing board. However, actually developing externally focused applications as robust, reliable, secure systems with high availability and very high adaptability was and still currently is a challenge that only few IT specialists know how to address.

Furthermore, models of marketplaces so far implemented put the stress and focus of their business model on the transactional aspect. This model, which can act and succeed—or, better, survive—in all those case where the liquidity is reached in a sensible spell of time, does not take into account a natural evolution of electronic ecosystems where the complexity of relationships is becoming more and more vital.

However, the connected economy and B2B marketplace models with it have just started their evolution and growth, which will not be stopped; they just need some adaptation mechanism to consolidate the new paradigm.

Thus, after that, a first pioneer period went by, where marketplace models have attracted business managers and, consequently, finance

Figure 1.



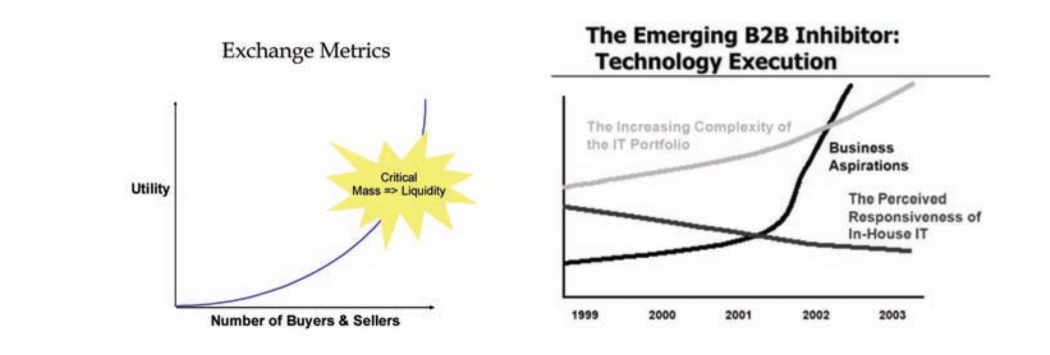
investors, showing the possible strategic value of Web technology, that is, the ability to extend the enterprise business model and to improve process efficiencies; now that paybacks have not been respected and investors begin to retire their support, the marketplace model must be readapted and reengineered (Bollier, 1996).

In order to better respect investors' and business managers' constraints and expectations, that is, to present a more realistic behaviour of the promised return of investment, it is now crucial for marketplace business models to take into account the following aspects:

- Membership and brand is more important than transaction fee (Dibb, Simkin, Pride, & Ferrel, 1991).
- Quality and quantity of services can aggregate other partners, which can bring additional financial resources and customers, in order to give more value to the industry they address.
- Processes begin before transactions and continue after it; support must be given to a wider range of customer tasks.

In the next pages two significant real implementation cases have been considered. The first concerns the possible evolution of a typical vertical marketplace positioned in the steel supply chain. The second shows how a service center for a geographical industrial district populated by many fragmented small enterprises can give value to the related supply chain and presents

Figure 2.



a valid paradigm to overcome possible "digital divide"-based approaches.

MAIN THRUST OF THE CHAPTER

Case Study 1: u-steel.com[1]

General Aspects

U-Steel is the first next generation e-marketplace, neutral, independent, and completely online, to be dedicated to the commerce of steel and related services. Initially launched in September 2000 by the strong sponsorship of a group of European partners such as COFACE (F), Falck Group (I), Mastr@ (I), Iperbusiness (I), Sopaf (I) and Mittel Generale Investimenti (I), U-Steel's aim is to bring the most qualified players of the steel industry online, increase the consumers purchasing power, and expand the markets.

From an operational perspective, U-Steel is an e-marketplace integrated with a large number of support services. Strategically, however, it is not just a virtual channel between buyer and seller but also an important tool to allow all supply chain participants to become a part, easily and safely, of the steel's industry future dynamics.

All effective members of the marketplace who purchase steel through U-Steel are certified and assigned a credit rating by COFACE, a worldwide leader in credit certification. Sellers who require total guarantee of their credits towards other U-Steel members can take out a credit insurance policy offered by COFACE exclusively to the portal's subscribers at extremely favourable terms.

When in 1999, as a group of friends, the founders came up with the idea of starting a company with the purpose of building an e-marketplace for steel, they immediately understood the huge importance of the project. Shifting to "virtual" steel meant entering a dimension completely new to the steel industry, traditionally tied to the business rules of the so-called "*old economy*."

In order to be successful, this new initiative required three elements: high level stockholders, capable of ensuring neutrality and independence to the marketplace; managerial staff knowledgeable in both the steel and the information technology industries; and finally, the most advanced technology available (Ansoff 1987; Hammer, 1996).

The idea was presented to a number of potential partners, all of whom accepted the project with great enthusiasm, starting from Iperbusiness of the TC Sistema Group, who became the technology partner, taking responsibility for the coordination of the project and the system integration requirements.

U-Steel has chosen a business model mainly based on the concept of *membership*, despite other initiatives based solely on the transaction mechanism. In an industry like steel-working, which only in Europe claims a turnover of approximately 285 million Euros, U-Steel's further objectives is to bring together partners who have a technologically innovative offering, placing particular emphasis on the technological brand.

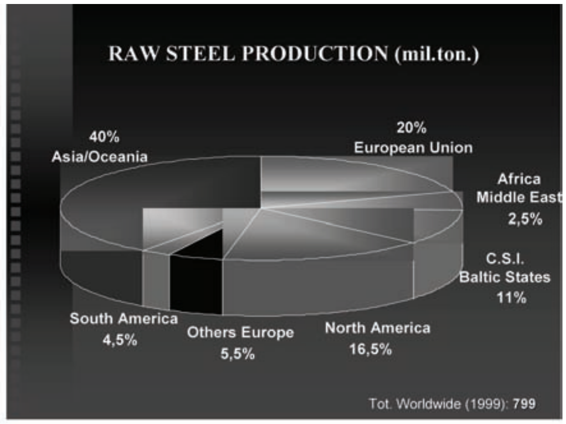
The marketplace starts off as a platform on which a number of additional services can be included; the objective is to increase the offering based on the efficiency of the cycle with other services ranging from insurance, credit, and logistics, all the way to integrating the partner's IT infrastructures with a technology across platforms.

U-Steel's transactional system is globally among the most advanced, with a variety of exchangeable products including plate, coil, rod, wire, pipe, and special irons and steels. Additionally, there is the possibility of marketing excess lots or scrap lots for the plants.

When a purchaser wishes to find a potential seller, they choose the product to buy, select the supplier, and can contact them through the marketplace to establish prices and terms by means of an interactive chat.

This way U-Steel becomes the point of contact between purchaser and seller, the location where the negotiation is conducted and finalized, regard-

Figure 3.



less of where in the supply chain the negotiation takes place. The model allows considerable savings in terms of time, cost, and back-office infrastructure.

But the portal is not only about buying and selling; this is just the initial step of an online process to which it will soon be possible to add additional services such as invoicing, logistics, warehousing, management and handling, accounting, and credit management.

Overall, a system capable of handling all processes not strictly connected to the steel's core business.

Who are the players interested in accessing the marketplace? The system potentially involves all components of the cycle, from the manufacturers to the retailers, from service centers who purchase large amounts and resell smaller lots of semifinished product, to the retailer or storekeeper who distribute the product to end users.

U-Steel does not promote disintermediation. On the contrary, it encourages the repositioning of sales figures who previously benefited from the inefficiencies of the supply chain and for whom it is now possible to become value added consultants. A trading tool only adds another level of competition within the procurement process.

A seller approaching U-Steel is given the opportunity of publishing the seller's catalogue directly on the marketplace, complete with prices and current available stock, or with the option of showing only the products without any further

Figure 4.

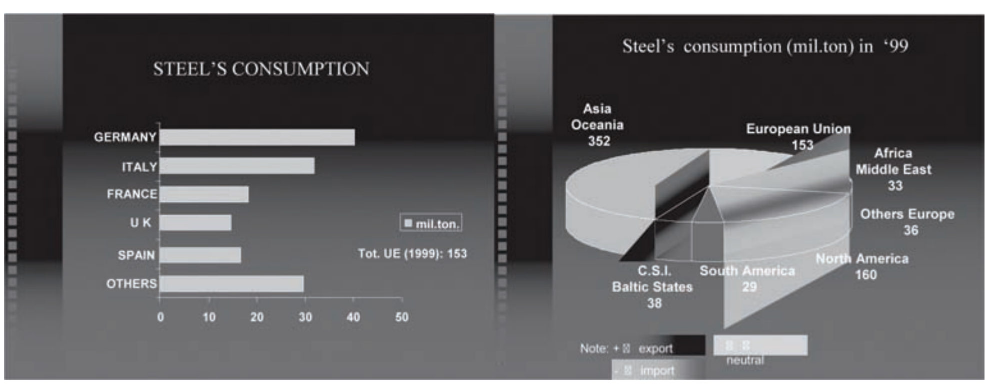
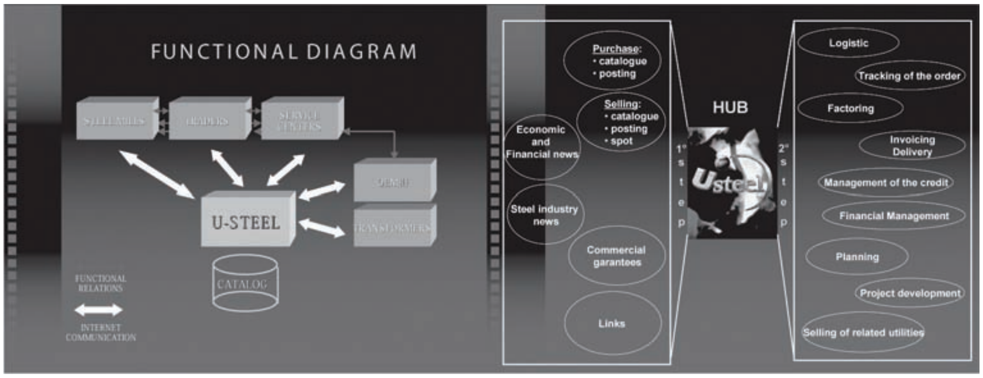


Figure 5.



information. In both cases, the subscriber is given the possibility of integrating the U-Steel catalogue with the seller's own information system so that each transaction results in a warehouse operation, a production cycle advancement, or a purchase order.

Alternatively, the manufacturer may simply replicate a portion of their product catalogue, put it online, and choose when to synchronize the two systems.

The technological components was chosen, developed, and implemented by Iperbusiness on the BroadVision application. Hewlett Packard was chosen as outsourcer of all the systems management services, including the environment infrastructure (i.e., premises, power supplies and surge suppressors/generators, security systems, fire prevention, local area networks, etc.) and the required hardware platform housed in the HP Data Centre located in Bergamo (I).

The system, designed and commissioned in just 4 months time, is based on a UNIX HP UX operating system, is configured on two HP9000 A400 clustered Web servers, two HP9000 A500 to run the BroadVision applications, and an HP9000 A500 box acting as DB server with the Oracle database. The information is stored on two HP SureStore ESC10 clustered devices with an HP SureStore E 2/20 backup library.

The network is based on two Cisco 2900 routers, a Cisco Local Director 416 firewall, and a Cisco PIX 525 in failover. The entire system was designed to provide the highest level of availability, and with doubled security and load balancing systems, it was built to handle an initial volume of 1.200 users in its first months.

Furthermore, HP specialists handle all the system resource administration activities ensuring the users' seamless and transparent functionality.

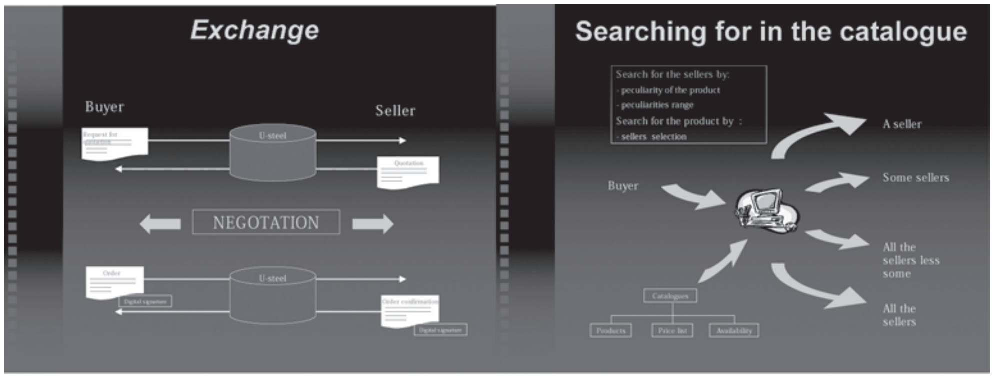
The E-Business Model

The business model of *u-steel.com* is based the on typical characteristics of the European steel supply chain and on the model of a B2B marketplace.

The European steel supply chain is formed by a very limited number of large steel producers and brokers, which are positioned at the top links of the chain, and by a great deal of other smaller players, grouped in dealers, steel service centers (SSC), and final companies which transform and use steel components in their own production. Even though this supply chain includes the presence of large producers, brokers, and even buyers (e.g., large automotive industry enterprises), the typical power play is well balanced, in that, medium and small companies can usually buy steel also oversee, that is, in America and Asia.

Thus, u-steel.com, which includes among its partners ex-business managers of the steel chain

Figure 6.



who know very well the market and the industry behaviour, is positioned as a vertical, *neutral*, *independent*, *aggregating*, and *collaborative* electronic market maker for steel trading exchanges and other services, in order to satisfy the needs of all the players and to support all the competitive forces of the steel supply chain.

Thus, first of all, u-steel.com is a vertical marketplace, which implements typical processes of a well identified market and industry where trading activities are accomplished in a particular context.

It is *neutral*, in that it is neither “seller” or “buyer” advocated. It offers to both the groups of players an electronic trading place where it enables secure commercial transactions and other services. Buyers and sellers can access the site and operate it using the typical trading approaches: catalogue searching and sourcing, online and even interactive exchange, and offer-posting processes.

The expected macro-benefits produced through these approaches can be grouped in the following main classes:

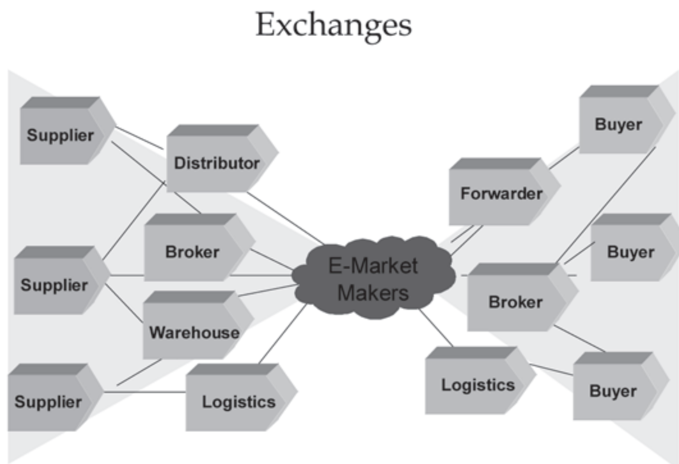
- Improving efficiency of intercompany operational processes by reducing the typical time usually needed to manually accomplish the same tasks.
- Enhancing commercial visibility and exploitation of resources, particularly for Steel Service Centres and small dealers, which can reach a larger audience of final transformers.
- Reducing buying cost, in term of final price, for small and medium transformers that can increase their ability to compete.
- Reducing stocking and work-in-progress costs of the whole supply chain through a better synchronization of production, distribution, and delivery.

Furthermore, taking into account the peculiarity of the European steel supply chain, U-Steel will offer in the immediate future a very sophisticated auction-based system, which will give the opportunity to large producers (i.e., steel mills a brokers) to launch forward actions on the European market and which, at the same time, will provide to small dealers and transformers with a mechanism of reverse actions to buy at better conditions also on oversee marketplaces.

U-steel is also an *independent* electronic market maker, in that it lies in a well-balanced power play position. Its claimed mission is “acting as informediary on the steel market,” with no particular preference for any side.

The historical role of shareholder partners ensures both the knowledge of the steel supply chain, through the Falck Group currently involved in the utilities industry, and the absolute transparency

Figure 7.



and independence, being the others well known players of other industries.

But U-Steel is an *aggregating* and *collaborative* market maker as well. These issues represent the innovative aspects of the U-Steel business model.

In fact, after the "hype cycle" of the connected economy has shown failures and constraints of an expected "no limits" growth, U-Steel proposes a further possible direction in a marketplace model evolution. U-steel offers other services by aggregating other business partners and delivering new functionalities which create new value for the steel business community.

There are two types of services that U-Steel delivers: business services and technological services.

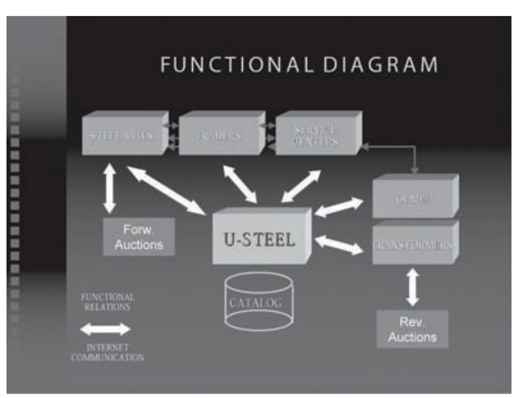
Business services include

- Credit insurance (already delivered at the moment)
- Logistics delivery

Technological services include

- Process integration
- Platform outsourcing

Figure 8.



Credit insurance gives the opportunity to sellers to insure their credit, following a credit rating mechanism classified and tuned by Coface (F), one of the worldwide insurance market leader and one of the U-Steel shareholder. Through these functionality, the seller can decide to insuring the seller's credit and thus electronically activate the operation with an automatic interaction with the Coface credit system, in order to assess and possibly accept economic conditions, as proposed by Coface. The seller pays a further transaction fee for that.

Logistics delivery, which is not currently available, but it will be in the future, will give the possibility to the buyer to choose a possible logistics and transportation player to be electronically involved in the process in order to deliver goods in a specified spell of time.

Process integration means the availability of integrating, when needed, the information system of both seller and buyer, in order to automatically accomplish several collaborating processes, such as procurement, fulfilment, billing, and invoicing through an XML interface. It means also the possibility to be integrated with other marketplaces supporting the same feature in order to extend the catalogue interaction to other "ecosystems."

The most innovative *collaborative* aspect of U-Steel is the availability of a technological platform, which can be rented by supply chain players. The technological platform on which U-Steel is based is a very flexible architecture provided by BroadVision and Oracle "best of breed" components. It ensures the possibility to build separate and independent environments where different players can implement their corners. In this way, a steel mill company or a broker, for example, can also deliver their own extranet, outsourced to U-Steel, avoiding the investment of a huge amount of money in a capital expenditure which would ask them for large financial resources and would force them to track a control on the return of investment.

In this model, on the contrary, U-Steel has the possibility to reduce this payback time and thus to shrink its ROI, while other players receive the value to have their own environment, classified as a variable cost in their profit and loss account. In other words, U-Steel can reduce the initial trade-off between revenues and costs and the other players using the feature, use the platform as a general expenditure, and do not have to directly involve themselves in technology, which is not their core business.

In this sense, the U-Steel model appears as a vertical model for trading and commercial operations, and, at the same time, as a horizontal model for distributing technological skills and services; a sort of a hybrid, new model, where U-Steel can act as a business service provider as well. A model that, taking into account the current evolution of other vertical marketplaces in other industries, can be considered a valid evolution and a possible successful paradigm.

The Revenue Lines

The U-Steel model is built on the following economic revenue lines:

- Membership fee, which is an annual fee to be paid by sellers and buyer in order to operate on the marketplace.
- Transaction fees on goods, as a percentage of the total amount.
- Transaction fees on services delivered by other partners (credit and logistics).
- Annual fee for ERP integration, following defined technological standards.
- Annual BSP fee, which is a yearly based fee for using the platform hosted by u-steel and adequately customized on customer needs.
- Advertising.

U-Steel will involve more then 6,000 European users after 3 years and will trade more then 5 millions of steel tons in the same period (3% of the whole European market). Expected revenues will be over 40 Millions Euro at the third year.

The www.u-steel.com site, which is currently available for test purposes to a limited community of "family and friends" players of the steel supply chain, will be launched and delivered worldwide on June 20, 2001.

Case Study 2: up2gold.com[2]

General Aspects

Up2Gold.com is a service center conceived to deliver consulting and technological services to the supply chain of gold and silver manufacturing located in Arezzo (I). Arezzo is one of the most important industrial sites for gold and silver manufacturing, a typical Italian industrial district which geographically groups many SMEs belonging to the same supply chain.

At the starting point of the gold (and silver) supply chain there are the raw material suppliers, that is, larger companies which are in a limited number, while production plants are divided into a plethora of small and very small companies linked together by complex multidirectional channels.

The distribution layer is, again, a "long chain" where various intermediaries and dealers are involved before reaching the end user.

These large companies control the supply links and schedule main production flows, while large dealers mainly govern the sales channel and, in particular, the export market.

As mentioned before, the productive part of the supply chain is very fragmented and powdered in a myriad of SME, involved in production subphases mainly as third parties. As it is usual in these cases, they do not participate to the research and development process; they just execute and manage orders received from other supply chain links.

These enterprises are often familiar companies, involving even less than 10 persons and unable to base their activities on some information system, which would be useless yet.

The final production output is finished goods like gold chains, rings, and ear-rings. The product is considered to be medium quality goods, usually competing with Indian and Far East equivalent productions.

Such a supply chain, despite its pervasive presence in many other sectors, is obviously inefficient, for the following reasons:

- The main obstacle to a more efficient chain is, as said before, the dimension of the company involved, the number of companies, and the related fragmentation.
- Moreover, the efficiency of the chain seems to be influenced by the flow of information regarding stock or availability of goods or raw material. In fact, while the commercial side of the chain (the end) does not know the availability, prices, quality, and other information, the producers do not know the real needs of their markets in order to be able to adequate their production volumes, stiles, and so forth.
- The third type of inefficiency is linked to the sales forecast or production schedule, which are extremely difficult to govern, because of a very fragmented catalogue of products; is very common the case of a catalogue with more than 10,000 articles, and practically every company has a catalogue of thousands of articles. Moreover, many articles are subject to season waves, other are linked to styles and stylists.

The result of these inefficiencies are, of course, poor service at the end user level (e.g., unavailability of products in certain period of the year) or, which is worst due to the cost of raw gold, obsolescence of the entire stock with the consequent inefficiency at the producer level. This also causes lack of competitive presence in those markets where the newcomers are getting more and more aggressive due to a lower production cost and a more flexible supply chain.

In order to better understand the subject, the

Table 1.

Site	**1-9 workers**	**10-49 workers**	**50-99 workers**	**100-999 workers**	**Total**
Arezzo *Companies* *Workers*	991 3366	269 4547	7 423	4 1279	1271 9615

Arezzo's gold and silver supply chain is described in Table 1.

Arezzo's supply chain shares the 20% of that global export, that is, more than 1.500 billion ITL. The total Arezzo's gold and silver production is more than 6.000 billion ITL.

The aim of up2gold.com is to offer a class of services to overcome the inefficiencies described above through the deployment of the a service center able to deliver:

- Consultancy services to help managing personnel, legal, and work trends issues;
- ERP services, with particular stress on an accounting and payroll application.
- A B2B supply chain marketplace to exchange raw material, semifinished goods, and human resources work availability and to interconnect along the chain goods' availability and stocks.
- A B2C marketplace, sell side-oriented, in order to increase sales of medium size companies.

The shareholders of the Up2Gold are:

- Arezzo's and National artisans associations (CAN).
- TC Sistema SpA and Datamat SpA, two famous IT players of the Italian market;
- FlyNet, a local ISP.
- Banca dell'Etruria e del Lazio, which is a local bank.

General concepts and strategic issues of the center have been designed in the following terms:

- The marketplace model appears innovative and suitable to fit chain needs and SMEs' expectations.
- It can be considered as a new paradigm for supply chain local systems and exported and disseminated to other SMEs districts; a real B2B best practice.

Table 2. The total Italian production export of gold and silver goods is shown.

Region	**Production export 1989 (billions ITL)**	**Production export 1999 (billions ITL)**
Europe	1173 (28.15%)	2323 (28.76%)
North America	1684 (40.41%)	2694 (33.36%)
Latin America	175 (4.20%)	634 (7.85%)
Middle East	468 (11.23%)	747 (9.24%)
Far East	379 (9.10%)	860 (10.65%)
Others	288 (6.91%)	818 (10.13%)
Total	4167 (100%)	8076 (100%)

The E-Business Model

When considering the business model of up2gold.com, it is necessary to underline that the company is a service center which delivers a set of services supporting a supply chain. It means that the center offers traditional B2B services along with "connected economy"-like functionalities.

Yet, even though the business model of the center is very articulated and all of its components should be considered and analyzed in a broader perspective, for the purposes of this chapter we would like to draw the attention to the marketplace part, which is an innovative B2B model and includes elements of as the supply chain management approach.

When addressing this marketplace, it is necessary to take into account typicality and behaviours of the supply chain to which it belongs. As already mentioned before, the Arezzo's gold chain is a very fragmented network of small suppliers all working for a few larger enterprises, which act as raw material suppliers and distribution channels (Fariselli, Oughton, Picory, & Sugden, 1997). These companies control the chain and play a strong power role: they make the research and development, that is, design gold and silver final products, schedule large scale production (distribution and sale channels, in particular), and control the raw material procurement process (raw material suppliers). They usually involve the other links of the chain as third parties, requesting them to produce finished or semifinished goods by a well defined delivery date (make-to-order).

This links, on their turn, involve other links of the supply chain, ordering subcomponents and other semifinished goods, scheduling delivery dates, and so on along the whole supply chain. All the subproducers links must generate by and large the following tasks:

- Procuring raw material, buying it in general from large raw material suppliers.
- Procuring human resources or other smaller production companies.
- Procuring semifinished goods, if any, from other chain players.
- Producing components as ordered.
- Delivering goods.

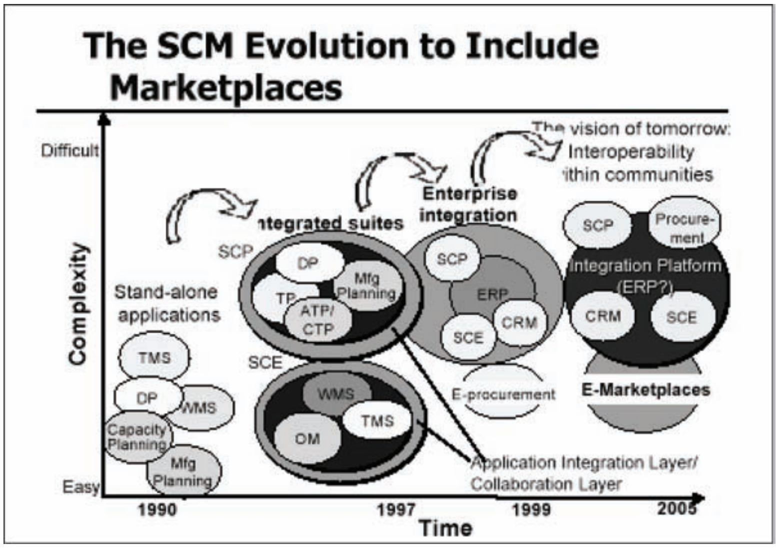
The value of information in accomplishing these activities is very high; if existing, it would reflect in a larger economy of scale, and in a better stock and delivery efficiency. Sometimes it happens that sale channels must refuse oversee orders, in that they are not able to declare on time a right delivery date and final prices to customers, for a lack of updated information!

These activities are not supported by an integrated information system and this fact, as already mentioned above, generates inefficiency to the whole network chain.

Up2gold.com is building a marketplace which offers two important services to the companies:

- A typical exchange marketplace in order to accomplish the first three activities as described above: procuring raw material, semifinished goods, and human resources. This marketplace gives raw material suppliers the ability to exchange gold and silver and give other enterprises the possibility to buy possible semifinished goods from others' stocks and/or human working resources availability. An additional auctions and reverse auctions system is included in order to sell stocks or buy components at a more convenient price.
- A collaborative system to support a supply chain management marketplace. Through this remote system, accessed via Web, any company, even small enterprises, can give their availability-to-promise, by claiming their actual stock availability, production capacity and, of course, prices. This system can give the sale channel an immediate

Figure 9.



picture of both the order progress status and the possible make-to-order availability of the whole supply chain.

The model that up2gold.com implements is not exactly that of a supply chain management marketplace, in that production orders scheduling is not taken into account. It is a sort of warehouse and stock availability integration, useful to catch up-to-date information before and after the order is assigned. "Before the order" means that the sale channel can assess and verify possible online availability along the whole supply chain and possible delivery dates to communicate to customers. "After the order" means that the sale channel and other links of the chain can assess online the real progress status of orders, controlling the actual delivery date.

This model is very simple in that no information system is required for SMEs; they are only requested to keep their order profile updated by entering simple information at the up2gold.com site.

Up2gold.com conceived this marketplace model in order to match actual requirements of SMEs-based supply chains, that is, the typical Italian industrial chains. Up2gold.com named this model *partner chain management* (PCM).

The Revenue Lines

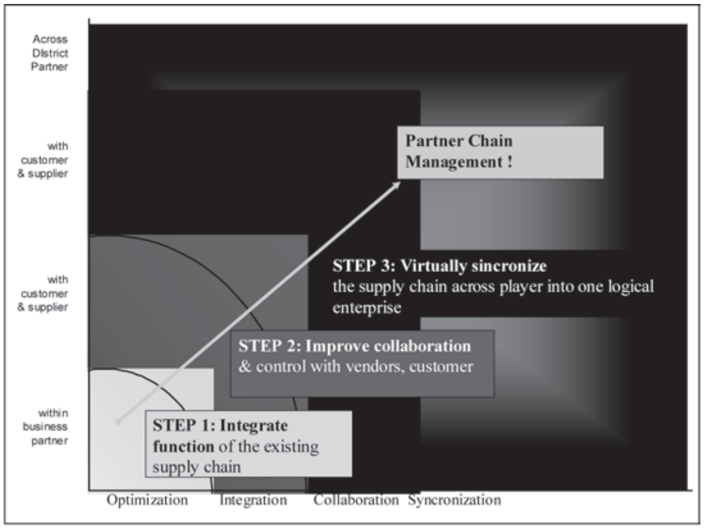
The Up2Gold model is built on the following economic revenue lines:

- Membership fee, which is an annual fee to be paid by any company in order to access the center services.
- Transaction fees on goods, as a percentage of the total amount.
- Annual fee for consulting and ERP function access.
- Advertising.

Up2Gold will involve more then 1.200 district companies after 3 years; the value of transaction involved will be more then 1.000 billion ITL.

The current site is only for presentation and advertising purposes. The service center will release its services progressively, starting at the beginning of 2002. The marketplace in particular will be launched and delivered by the first quarter of 2002.

Figure 10.



FUTURE TRENDS AND CONCLUSION

The B2B e-marketplaces will represent in the future the new efficient paradigm that will be used to accompany all trading process concerning goods, commodities in particular. They will be the "low cost" arenas where large companies will invite their provider to submit proposals at the best prices. But they will also be the "low cost" arenas where SMEs will present their products and their brands in order to be known and contacted by a larger and larger audience. E-marketplaces will more likely act per industry, instead of being "cross-industry, general purpose," and will show the two natures as previously mentioned, that is:

- Organized and driven by a large enterprise that will use the e-marketplace in order to attract a great number of suppliers, globally widespread, and able to enhance their good value proposition. This will ensure a strong competition and low supply prices.
- Organized and driven by communities (associations) particularly composed by small enterprises that will group their efforts in order to make their brand and their products more globally known as possible. Through this mechanism, SMEs will also implement collaboration tools and instruments in a sort of "partners' chain network," where different facilities, such as marketing services, information systems, administrative, and financial services, can be shared at lower and more competitive internal prices by creating at the same time both efficiency and new dimensions of effectiveness.

REFERENCES

Ansoff, H. I. (1987). *Corporate strategy* (2nd ed.). Harmondsworth: Penguin.

Bollier, D. (1996). *The future of electronic commerce*. Queenstown, MD: The Aspen Institute.

Dibb, S., Simkin, L., Pride, W., & Ferrel, O. (1991). *Marketing, concepts and strategies*. Boston: Houghton Mifflin.

Fariselli, P., Oughton, C., Picory, C., & Sugden, R. (1997). *Electronic commerce and the future for SMEs in a global marketplace: Opportunities and public policy requirements*. Brussels: CEC.

Geppert, L. (2001, December). *E-Marketplaces: New models.* Paper presented at the First Policy Group Meeting of DEEDS Project, Brussels

Hammer, M. (1996). *Beyond reengineering*. London: Harper Collins Business.

Kalakota, R., & Whinston, A. (1996). *Frontiers of electronic commerce.* Reading, MA: Addison-Wesley.

Kyte, A. (2001, February). Paper presented at the iEB Europe Gartner Conference 2001, London.

McDonald, M. H. B. (1997). *Marketing plans: How to prepare them, how to use them.* Oxford: Butterwoth-Heinemann.

Porter, M. (1985). *Competitive advantage.* New York: Free Press.

Shapiro, C., & Varian, H. R. (1998). *Information rules: A strategic guide to the network economy*. Boston: Harvard Business School Press.

Timmers, P. (1999). *Electronic commerce: Strategies and models for B2b trading*. New York: Wiley.

ENDNOTES

1. At the moment, u.steel is not longer directly operating.
2. At the moment, Up2gold.com is not longer directly operating. It has been included in another larger initiative.

This work was previously published in Business Web Strategy: Design, Alignment, and Application, edited by L. Al-Hakim and M. Memmola, pp. 118-132, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 2.12
Developing Rule-Based Applications for the Web: Methodologies and Tools

Vassilis Papataxiarhis
National and Kapodistrian University of Athens, Greece

Vassileios Tsetsos
National and Kapodistrian University of Athens, Greece

Isambo Karali
National and Kapodistrian University of Athens, Greece

Panagiotis Stamatopoulos
National and Kapodistrian University of Athens, Greece

Stathes Hadjiefthymiades
National and Kapodistrian University of Athens, Greece

ABSTRACT

Embedding rules into Web applications, and distributed applications in general, seems to constitute a significant task in order to accommodate desired expressivity features in such environments. Various methodologies and reasoning modules have been proposed to manage rules and knowledge on the Web. The main objective of the chapter is to survey related work in this area and discuss relevant theories, methodologies and tools that can be used to develop rule-based applications for the Web. The chapter deals with both ways that have been formally defined for modeling a domain of interest: the first based on standard logics while the second one stemmed from the logic programming perspective. Furthermore, a comparative study that evaluates the reasoning engines and the various knowledge representation methodologies, focusing on rules, is presented.

DOI: 10.4018/978-1-60566-402-6.ch016

1. INTRODUCTION AND MOTIVATION

Nowadays, with the evolution of traditional web of documents to a more complex web of services, an increasing demand for embedding intelligence to Web applications arises. In this context, the efficient management of knowledge seems to play a key role in order to achieve smart behavior of Web applications and to overcome several issues of such environment (e.g., information integration). Ontologies, mainly written with Semantic Web technologies, constitute a well-established paradigm for representing knowledge on the Web. Though, current efforts are focused on extending ontologies with more expressive forms of knowledge like rules. In fact, given the state-of-the-art in the realization of the Semantic Web vision, rules constitute the next prominent challenge. Since the ontology layer of the Semantic Web architecture stack has reached a sufficient degree of maturity through Web Ontology Language (OWL) (Dean et al., 2004), the next step of progress involves the integration of rules with ontologies, most of them based on subsets of First Order Logic (FOL).

Rules are capable of extending the expressiveness provided by ontology languages through the definition of more complex relationships between individuals. Additionally, as a modular form of knowledge, they fit well in domains like personalization, policies and business-to-business (B2B) interaction. However, it has been shown that extending ontologies even with simple forms of rules can lead to undecidability of key inference problems.

On the other hand, many business-logic applications have extensively taken advantage of existing rule management systems or solvers (Jess, 2008; ILOG, 2008; Drools, 2008), aiming at facilitating the knowledge management process. As a result, the success of rules in non-Web applications moved Web researchers to use traditional rule engines on the Web.

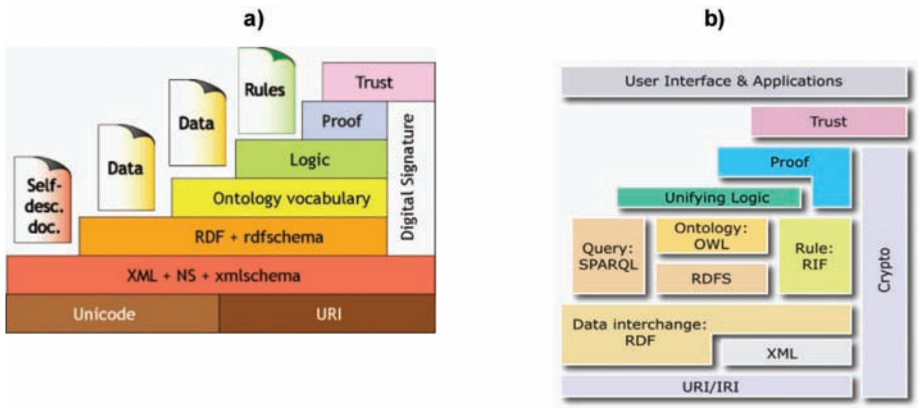
However, the aforementioned stable rule systems have not been originally created for open and heterogeneous environments like the Web. Such platforms have adopted different knowledge representation formalisms, mainly based on principles of logic programming, instead of classical logic. As a consequence, they differ from recent Semantic Web technologies in many aspects, including representational features and reasoning functionality, as well. Hence, building a rule-based application for the Web with existing rule technologies is not a straightforward task.

In the rest of this chapter, we provide foundational knowledge on this topic together with implementation issues, techniques and design patterns. Section 2 briefly describes how the things have gone so far in the area of Web knowledge representation formalisms. In Section 3, various knowledge representation methodologies and tools are discussed. Specifically, Section 3.1 demonstrates the different languages and formalisms, derived from both classical logic and logic programming view, while Section 3.2 focuses on various engines able to reason over such knowledge bases. Section 4.1 gives the main requirements for rule-based web applications. The evaluation presented includes both a qualitative comparison (Section 4.2) of the existing approaches and a performance analysis (Section 4.3) of current ontology reasoners and rule engines. Finally, several future trends and open issues are identified in Section 5. Hence, this chapter aims at becoming a helpful guide for applying rules to Web applications.

2. THE STORY SO FAR

The knowledge representation languages proposed (see Section 3.1) for representing knowledge on the Web are based either on the Classical Logic (CL) perspective or on Logic Programming[1] (LP). As a result, a debate was started between the Database community and AI researchers, respectively, in order to determine the more suitable of the two approaches in the formalization of Web knowledge. Additionally, different languages of the same perspective, providing various degrees of expressivity, have been proposed. Hence, the integration of knowledge with Web applications was more complicated. Recently, with the evolu-

Figure 1. a) Tim Berners-Lee Semantic Web layer "cake" and b) the latest form of Semantic Web stack diagram (adapted from W3C Semantic Web Activity, 2008)

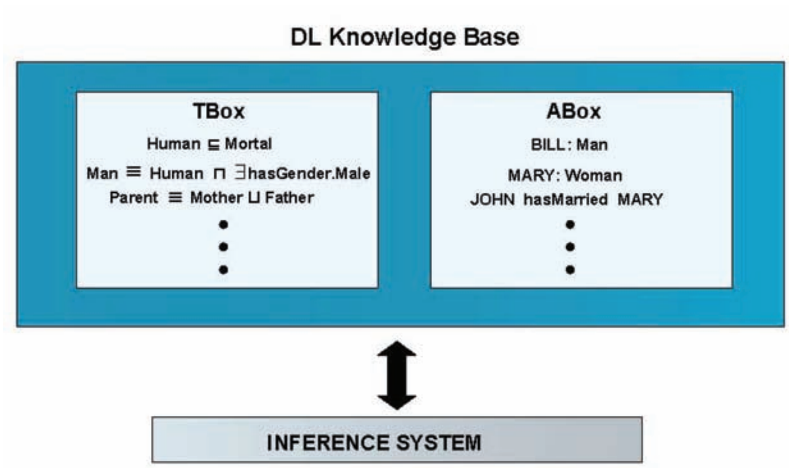


tion of Semantic Web technologies, these modeling paradigms have been extensively discussed by researchers (Motik, Horrocks, Rosati, & Sattler, 2006; Patel-Schneider & Horrocks, 2006; Eiter, Ianni, Polleres, Schindlauer, & Tompits, 2006a; Donini, Lenzerini, Nardi, & Schaerf, 1998; Antoniou, et al., 2005; Boley, Kifer, Patranjan, & Polleres, 2007).

Clearly, this confrontation also affected rules. Although, significant efforts have been devoted in order to develop appropriate rule languages for the Web (see Section 3.1), a debate between proponents of the different perspectives has appeared. The main argument involves the degree that rules would be combined with ontologies. Initially, as formulated by Tim Berners-Lee, it had been commonly accepted that Semantic Web should be structured in a hierarchy of layers that seamlessly interoperate (Antoniou & van Harmelen, 2004). However, some proposals (mainly stemmed from the database community) preferred to keep the rules and the ontology layer separated, in order to preserve expressiveness and decidability of reasoning process, as mentioned in (Eiter et al., 2006a). Specifically, in (Kifer, de Bruijn, Boley, & Fensel, 2005) the authors argue that the rules layer of Semantic Web architecture should be placed next to the ontology layer of the Semantic Web stack in order to take advantage of important rule languages based on logic programming. However, in (Horrocks, Parsia, Patel-Schneider, & Hendler, 2005) the authors criticize the aforementioned approach and denote that such a distinction would lead to two Semantic Webs based on different semantics. These discussions along with the Semantic Web Activity workgroup have led to several modifications of the stack. Figure 1 presents the initial and the current form of the Semantic Web stack diagram specified by the W3C Semantic Web Activity Group (W3C Semantic Web Activity, 2008).

The aforementioned facts gave rise to confusion inside the Web community. If the experts (i.e., the logicians stemming from both AI and databases) could not decide and recommend a unified framework for formalizing knowledge, how Web developers and users should choose the best language according to their needs? The current solutions are either to study the capabilities provided by all these languages (i.e., available expressiveness, support by existing reasoning engines, tractability etc.) or to exclude knowledge from their applications.

Figure 2. Architecture of a Description Logics knowledge-based system



3. KNOWLEDGE REPRESENTATION METHODOLOGIES AND TOOLS

This section surveys different knowledge representation methodologies along with reasoning modules that have been proposed for managing knowledge on the Web. After an overview of Description Logics that comprise the main formalism for representing ontological knowledge in Web applications, we focus on rules.

3.1 Knowledge Representation Formalisms

Description Logics

Description Logics (DLs) (Baader, Calvanese, McGuiness, Nardi, & Patel-Schneider, 2003) are subsets of first-order-logic, originating from semantic networks and frame-based systems. They constitute a family of knowledge representation languages that aim at providing well-understood mechanisms in order to formalize knowledge that describes a domain. This way, DLs are equipped with formal, logic-based semantics, emphasizing on reasoning process. Typical reasoning tasks include consistency checking of the knowledge base, concept satisfiability, instance checking etc.

A DL-based knowledge base is composed of two components: TBox and ABox. TBox contains the vocabulary of the application domain, called *terminology,* as well as axioms based on that vocabulary. Practically, such vocabulary consists of *concepts* and *roles*. Concepts are generic descriptions of sets of individuals, while roles constitute binary predicates for defining properties of the individuals. On the other hand, ABox includes assertions of individuals that may refer to either concepts or roles. For example, a statement declaring that a specific individual is instance of a concept resides in ABox, while a statement denoting that "every human is mortal" belongs to TBox. Figure 2 shows the generic architecture of a DL knowledge representation system.

Each description logic language is determined by a set of constructs, enabling the use of atomic concepts and atomic roles in order to define complex ones. These constructs directly affect the expressive power of the language and, thus, the complexity of inference tasks. As a result, the selection of the appropriate description logic language in order to describe a specific domain includes the examination of the imposed requirements for representation expressiveness.

DLs set up the base for the definition of Resource Description Framework (RDF) (Klyne &

Caroll, 2004), RDF Schema (Brickley & Guha, 2004) and Web Ontology Language (OWL) (Dean et al., 2004) that constitute W3C standards for representing knowledge on the Web. Specifically, RDF is a simple data model for representing information on the Web. RDF statements are expressed in the form of (*subject, predicate, object*) triples. A set of such statements can be viewed as a graph where the subjects and the objects of these statements constitute the graph nodes while the predicates correspond to the graph edges. RDFS is an extension of RDF for expressing simple taxonomies through the definition of class/property hierarchies and domain/range of properties. OWL is totally based on DLs and it comes in three species: OWL-Lite, OWL-DL and OWL-Full. OWL-Full is the most expressive OWL species, since it takes advantage of all the OWL language primitives. OWL-DL limits OWL-Full to a subset of OWL primitives so as to achieve efficiency of reasoning. Finally, OWL-Lite is a sublanguage of OWL-DL by restricting more its expressiveness.

Description Logic Programs

The expressiveness of Description Logic Programs (DLP) approach (Grosof et al., 2003) corresponds to a fragment of OWL defined by the expressive intersection of Description Logics and logic programming. This approach intends to define a mapping from DL to logic programming (specifically, Horn programs in which no function symbols, negation and disjunction are allowed) and vice versa. An instance of such a mapping in the case of conjunction follows:

$$C_1 \sqcap C_2 \sqsubseteq D \equiv D(x) \leftarrow C_1(x) \wedge C_2(x) \quad (1)$$

However, for the sake of decidability, DLP offers limited expressiveness, since the aforementioned mapping covers only a few DL-constructs (in particular, conjunction, disjunction and quantification restrictions). For example, Description Logic Programs do not cover negation in class descriptions nor fully support cardinality restrictions.

Semantic Web Rule Language

Semantic Web Rule Language (SWRL) (Horrocks et al., 2004b) is probably the most popular formalism in Web community for expressing knowledge in the form of rules. Specifically, SWRL is based on a combination of Web Ontology Language (OWL) (Dean, 2004) and Rule Markup Language (RuleML) (RuleML, 2008) and has been proposed as a W3C candidate standard for formalizing the expression of rules in Web context. Contrary to DLP, SWRL extends OWL-DL with a specific form of Horn-like rules.

The proposed rules are in the form of an implication between the body and the head of the rule. A typical SWRL rule can be of the following form:

$$a_1 \wedge a_2 \wedge \ldots \wedge a_n \rightarrow b_1 \wedge b_2 \wedge \ldots \wedge b_m \quad (2)$$

where a_i and b_i are OWL atoms of the following forms:

- Concepts, e.g., C(x), where C is an OWL description, in general, and x is either a variable, an OWL individual or a data value.
- Object properties, e.g., P(x,y), where P is an OWL property and x, y are either variables, individuals or data values.
- Datatype properties, e.g., P(x,y), where P is an OWL property, x is variable or individual, while y is a data value.
- B(x1,x2,…), where B is a built-in relation and x1, x2, … are either variables, individuals or data values.

- sameAs(x,y) or differentFrom(x,y) where x, y are either variables, individuals or data values.

The main advantage of SWRL is the simplicity it offers, while extending the expressiveness of OWL. Another benefit of SWRL is its compatibility with OWL syntax and semantics, since they are both combined in the same logical language. On the other hand, extending OWL-DL with SWRL rules leads to undecidability of simple inference problems. A possible solution of this problem is presented in (Motik, Sattler, & Studer, 2005) which introduces the notion of *DL-safe* rules. Specifically, that approach restricts the application of SWRL rules only to individuals of the ABox part of the DL knowledge base. Moreover, SWRL does not support negation (neither classical nor negation as failure - NAF) and disjunctions. Finally, there is no efficient support of first-order provers to execute reasoning over SWRL. Usually, the SWRL rules are translated to existing rule systems (e.g., Jess (O'Connor, Knublauch, Samson, & Musen, 2005)) that handle the reasoning tasks partially, since they are not aimed to manage knowledge expressed in terms of first-order logic or subsets.

In the case of a Web service composition paradigm, some SWRL example rules could be the following:

$$\text{profile:hasOutput(?S1,?out)} \wedge \text{profile:hasInput(?S2,?in)} \wedge \text{process:parameterType(?in, ?cin)} \wedge \text{process:parameterType(?out, ?cout)} \wedge (\text{rdfs:subClassOf(?cout, ?cin)} \vee \text{rdfs:subClassOf(?cin, ?cout)}) \rightarrow \text{composableWith(?S1, ?S2)} \quad (3)$$

The abovementioned rule captures the knowledge that a web service S1 is composable with a service S2, if an input of S1 is either subclass or superclass of at least one output of S2. This rule uses specific namespaces of OWL-S ontology like "profile" and "process".

Answer Set Programming

Answer Set Programming (ASP) (Gelfond & Lifschitz, 1991) is a paradigm for knowledge representation and declarative programming. It has several advantages compared to other logic programming paradigms (e.g., Prolog) such as:

- *Full declarativity*: the order of rules in a program is not important.
- *ASP programs are in general decidable*
- *Non-monotonic inference*: both negation as failure (NAF) and strong negation are supported, thus enabling default reasoning and reasoning under the Closed World Assumption (CWA).
- *Availability of efficient solvers*: there are several ASP solvers that are scalable enough to deal with large knowledge bases.

A general ASP rule is of the following form:

$$a_1 \vee a_2 \vee \ldots \vee a_n \leftarrow b_1 \wedge \ldots \wedge b_k \wedge \text{not } b_{k+1} \wedge \ldots \wedge \text{not } b_m \quad (4)$$

where a_i and b_j are literals (atoms or strong negations of atoms) and *not* denotes NAF.

A set of such rules is an ASP program. What is interesting is the fact that these rules can have disjunctions in their head. This is a very important feature of ASP, since it introduces non-determinism in the inference process (i.e., an ASP program may have several models which are called *answer sets*).

In order to be able to use ASP on the Web, the ASP rules should be combined with Web knowledge. Since ontologies is the most common way to represent knowledge on the Web, an interaction between ASP programs and ontologies is deemed necessary. A solution to this integration problem is description-logic programs (or dl-programs). These consist of ASP rules that may contain queries to DL knowledge bases. For example, the following rule "brings" into the ASP program all instances of the class MovieTitle in the DL knowledge base:

$$\text{movie}(X) \leftarrow \text{DL}[\text{“Movie”}](X). \tag{5}$$

Several extensions to dl-programs were proposed (and implemented) so that they become more "suitable" for open environments like the Web, where information may be expressed in many diverse ways (e.g., multiple different ontologies). The most well known extension is HEX-programs (Eiter, Ianni, Schindlauer, & Tompits, 2005), which enable handling knowledge expressed in various formalisms, even with potentially different semantics (e.g., RDF(S) and OWL). HEX-programs contain several features (e.g., higher-order logic features) that enable more flexible integration with external knowledge bases. These extensions result in new syntax elements. For example, the atom *&rdf[u](s,p,o)* evaluates to true if <s p o> is an RDF triple asserted at URI u. Finally, another advantage of HEX-programs is that they allow using external data processing services that logic programming cannot handle (e.g., string processing). An interesting engine for HEX-programs is dlvhex[2] which is described in Section 3.2.

There have been proposed ASP-based several applications in the context of (Semantic) Web. One of the most promising is Web service composition (Rainer, 2005). The authors apply ASP techniques to "build" service compositions from available services that match a certain service request.

Web Service Modeling Language (WSML)

WSML (de Bruijn et al., 2005) is a language of representation languages for the Semantic Web. These languages are based on several different formalisms such as Description Logics, Logic Programming and First-Order Logic. Some of the basic variants of WSML are the following:

- **WSML-Core:** A subset of a Description Logic which falls inside the Horn logic fragment of FOL. It supports subsumption reasoning and query answering.
- **WSML-Flight:** An extension of WSML-Core which also supports full Datalog rules, default negation and integrity constraints. It can provide query answering in the context of Logic Programming.
- **WSML-Rule:** An extension of WSML-Flight with support for function symbols and unsafe rules.

Reasoning for these WSML variants can be implemented by several Logic Programming engines. For some features of these languages, DL reasoners can also be used or First Order theorem provers. These languages have been extensively used in several European projects, mainly in the application domain of semantic web services. Hence, several APIs, tools and other facilities are available for building WSML-enabled applications.

Defeasible Rules

Defeasible logic (Nute, 1994) is a rule-based, non-monotonic approach able to deal with incomplete knowledge and inconsistencies. These features have been widely remarked in the context of realizing Semantic Web vision, mostly in information integration areas (e.g., ontology merging). As a result, some efforts in research community (Antoniou, Billington, Governatori, & Maher, 2001) were devoted to carry the advantages of defeasible logic in the area of Semantic Web technologies.

The main idea behind defeasible logic reasoning systems is the ability to handle a number of additional features with regard to classical rules like priorities of rules, default inheritance, exceptions, etc. There are three different types of rules in a defeasible logic reasoning system: a) classical rules (called *strict rules*), b) *defeasible rules* that can be contradicted by other rules and c) *defeaters* used to specify exceptions of defeasible rules. This way, an important aspect achieved by such reasoning modules is their capability of resolving the possible conflicts that arise among defeasible rules.

Classification of Approaches Integrating Ontologies and Rules

Although several approaches have been discussed for combining rules with Semantic Web ontologies (Horrocks et al., 2004b; Grosof, Horrocks, Volz, & Decker, 2003; Eiter, Lukasiewicz, Schindlauer, & Tompits, 2004; Bassiliades, Antoniou, & Vlahavas, 2006; Rosati, 2006a; Rosati, 2006b), there is no totally accepted solution in the field. The main topic of argumentation is the degree of integration between the ontology layer and the rules layer. In this section we intend to provide a brief classification of the proposed approaches.

Two main categories of integration approaches have been distinguished in this context:

a. **Homogeneous approaches.** These approaches suppose a tight semantic integration of the two layers. Specifically, both ontologies and rules are embedded in a common logical language, permitting predicate sharing in a coherent way. In such approaches, ontology concepts and properties may be defined through rules. The most typical homogeneous paradigm is the combination of SWRL rules with OWL ontologies. This is also the most In addition, Description Logic Programs (DLP) (Grosof et al., 2003) constitutes another similar approach.
b. **Hybrid approaches.** These approaches correspond to a strict semantic separation between ontologies and rules. In particular, this strict separation concerns the rule predicates and the ontology elements. Hence, the vocabulary (concepts and properties) offered by the ontologies is used as a conceptualization of the domain and rules cannot directly define ontology classes or properties. Many integration approaches adhere to this category, including Answer Set Programming (Gelfond & Lifschitz, 1991), *dl-programs* (Eiter et al., 2004) and *DL+log* (Rosati, 2006c).

The user may find more details on this topic in (Eiter et al., 2006a).

Table 1 summarizes the aforementioned knowledge representation languages and their basic features.

3.2 Reasoning Engines

In this section we intend to provide the reader with a comparative feature analysis of existing reasoning modules, including description logic reasoners and rule engines, as well.

Jena2 (McBribe, 2002) is the second generation of Jena Semantic Web programming toolkit, which is a Java framework for developing applications based on Semantic Web technologies. Specifically, Jena provides an Application Programming Interface (API) for creating, storing, managing and querying RDF graphs as well as RDFS, OWL ontologies in various formats (RDF/XML, N3 and N-triples). The RDFS reasoner included in Jena framework does not support datatypes and blank node entailments. The built-in OWL reasoner is very limited, since it is a rule-based implementation of OWL-Lite. However, Jena is supplied with an interface which facilitates the connection and interoperation of the framework with any external reasoner that supports the DIG (DL-Implementation Group) standard (Bechhofer, Moller, & Crowther, 2003). Hence, the API provided by Jena could be integrated with most of the existing description logics reasoners. Furthermore,

Table 1. Basic features of various knowledge representation languages

Feature Language	Logical Foundation	Decidability*	Serialization formats
OWL	Classical Logic (FOL subset)	OWL-Lite: decidable OWL-DL: decidable OWL-Full: undecidable	XML/ N-triples (textual)
OWL + SWRL	Classical Logic (FOL subset)	undecidable	XML
DLP	DL and LP intersection	decidable	textual serialization (in terms of rules)
ASP	Extension of LP (disjunction in rule heads, DL queries)	decidable	textual serialization (in terms of rules)
DL+log	DLs + Datalog rules (disjunctive, non-monotonic)	decidable	textual serialization (in terms of rules)

Note: *regarding key inference problems (e.g., consistency of the knowledge base)

Jena provides a query engine in order to execute SPARQL (Prud'hommeaux & Seaborne, 2005) queries over RDF graphs.

RacerPro system (RacerPro, 2008) is the commercial[3] extension of Racer (Haarslev & Moller, 2001), which is probably the most popular reasoning engine for OWL ontologies to practitioners of Semantic Web technologies. RacerPro can be seen as a knowledge-based repository that can handle and it is a system for managing OWL ontologies, in particular. It implements the DIG interface and it offers an optimized tableau calculus for the description logic SHIQ(D). Additionally, it supports qualified cardinality restrictions as well as some extensions of OWL (e.g., OWL-E (Pan & Horrocks, 2004) except user defined XML datatype expressions). Moreover, the latest version of RacerPro includes a first implementation of an SWRL rule engine.

Pellet (Sirin, Parsia, Grau, Kalyanpur, & Katz, 2007) is a Java-based, open source reasoner capable of handling expressive OWL ontologies. It implements an optimized tableau algorithm, augmented with a number of additional features (e.g., support for Unique Name Assumption - UNA, closed world reasoning, SPARQL query answering). Pellet also provides an explanation service in order to facilitate the debugging of the ontology engineering (Parsia, Sirin, & Kalyanpur, 2005). Although typical reasoners detect the inconsistencies between concepts of the knowledge base (KB), Pellet can explain *why* a concept description led to unsatisfiability. This way, the reasoner provides user with additional knowledge (e.g., relevant axioms or restrictions) sufficient to understand the problem and reform the KB properly. Finally, Pellet allows ontologies to use XML-Schema built-in and user-defined datatypes that extend numeric and date/time types.

Bossam (Jang & Sohn, 2004) is a RETE-based, forward-chaining reasoning module for reasoning and querying over RDF(S) and OWL documents, while it also supports rules execution (SWRL rules are included). It is based on Logic Programming (LP), augmented with some expressiveness features stemming from First-Order-Logic (FOL). For example, a number of additional to LP features are provided by Bossam, including support for both classical negation and NAF and disjunctions in the body of rules. Hence, it does not support complete reasoning over OWL Ontologies. Furthermore, it facilitates the integration of rules with Java by supporting a procedural attachments mechanism for SWRL rules. Finally, Bossam provides an API for managing the engine, loading ontologies and rules, querying RDF(S)/OWL documents and giving explanations about derived facts. Currently, Bossam does not support SPARQL query answering, while the serialization of the knowledge base to a persistent store (e.g., file system) is another missing feature.

FACT++ system (Tsarkov & Horrocks, 2006) is the descendant of FACT (Horrocks, 1998). Contrary to the lisp-based FACT system, FACT++ is an open source reasoner for SHOIQ(D) implemented in C++. It is based on tableaux algorithms in order to provide both TBox and ABox reasoning tasks and it can be accessed through the DIG interface. It also supports a number of additional features like handling enumerated classes (a.k.a. nominals) and it is a very efficient TBox reasoning engine. However, the main disadvantage of FACT++ is its inefficiency to support complete ABox reasoning. Hence, FACT++ is unsuitable to applications that call for instance classification and retrieval.

KAON2 (KAON2, 2008) is the successor of KAON Project (KAON, 2008) and unlike pure description logic reasoners, KAON2 does not implement a tableaux algorithm. In fact, KAON2 is a hybrid reasoning module able to handle both ontologies expressed in description logics terms and Disjunctive Datalog programs. It implements algorithms that reduce description logic SHIQ(D) to Disjunctive Datalog (Hustadt, Motik, & Sattler, 2004), taking advantage and applying well-known practices stemming from deductive databases (e.g., magic sets) to DL reasoning. It can also handle SWRL, F-Logic ontologies and SPARQL query answering, as well. Moreover, it provides a Java API in order to accommodate the management and the integration of different knowledge formalisms (e.g., OWL ontologies with SWRL rules). However, KAON2 does not support reasoning about nominals and cannot handle a large number of cardinality restrictions. Hence, it cannot deal with the full OWL-DL expressiveness.

dlvhex (Eiter, Ianni, Schindlauer, & Tompits, 2006b) is the name of a prototype application for computing the models of so-called HEX-programs, which are an extension of Answer Set Programs towards integration of external computation sources. dlvhex can communicate with OWL and RDF knowledge bases and can also return the results in RuleML syntax. A strong point of dlvhex is that it enables developers to write and embed plug-ins to the core engine. In fact, support for RDF and OWL is implemented through plug-ins too. Recently, another plug-in for querying HEX models through the SPARQL language (Polleres & Schindlauer, 2007). Through this plug-in dlvhex can be used as a query engine by providing a rewriter from SPARQL to rules. The source code and binaries of dlvhex are publicly available.

DR-DEVICE (Bassiliades, et al., 2006) is a defeasible logic reasoner for the Web based on the CLIPS expert system shell (CLIPS, 2008) that intends to integrate Semantic Web standards (RDF metadata, XML-syntax of rules) with non-monotonicity (e.g., strong negation). Specifically, it provides reasoning services over RDF metadata by taking advantage of rules defined by defeasible logic (strict rules, defeasible rules and defeaters). It also claims for reasoning efficiency compared to other systems based on logic programming.

Jess (Jess, 2008) is a Java framework for editing and applying rules, since it contains a scripting environment and a rule engine, as well. It supports a CLIPS-like language suitable for developing applications based on declarative rules (a.k.a. expert systems). Jess also uses an optimized version of Rete algorithm (Forgy, 1982) tailored for Java, comprising a very efficient rule engine. Recently, the evolution of rule technologies on the Web has led Jess to rebound its practical value in the community of Web developers. Moreover, the fact that Jess is a Java-based system facilitates its integration with a number of Web programming paradigms like Java servlets or applets. Finally, it supports backward-chaining and some additional features like procedural attachments.

Table 2 presents the types of inference support provided by the aforementioned reasoning

Table 2. Types of inference support by various reasoning engines

Inference Support Modules	TBox Reasoning	Abox Reasoning	Rules Reasoning
Jena2	Limited (incomplete RDFS/OWL reasoning)	Limited (incomplete RDFS/OWL reasoning)	JenaRules (forward/tabled backward chaining)
RacerPro	√	√	nRQL rules, first implementation of SWRL (forward chaining)
Pellet	√	√	DL-safe rules (SWRL subset)
Bossam	sound, incomplete	sound, incomplete	SWRL, Buchingae (forward chaining)
FACT++	√	sound, incomplete	-
KAON2	√ (except nominals)	√ (except nominals)	DL-safe rules (SWRL subset)
dlvhex	Limited, through interface with external reasoner	Limited, through interface with external reasoner	DL-Rules (non-monotonic logic program rules with queries to DL KB)
DR-DEVICE	-	-	Defeasible rules
Jess	Limited*	Limited*	SWRL*, Jess rules (forward/backward chaining)

* through appropriate transformations (Mei, Bontas & Lin, 2005; O'Connor et al., 2005)

modules. In (Cardoso, 2007), an analysis of the current trends and the adoption of available reasoning engines by the Semantic Web community are presented in detail.

4. EVALUATION OF EXISTING APPROACHES

In this section, a comparative study across the aforementioned methodologies for introducing rules in Web applications is presented. Firstly, we give some fundamental requirements that have to be satisfied by such formalisms. Afterwards, we compare various aspects from the perspectives of classical logic and logic programming, giving indicative examples. Finally, some experimental results that examine the efficiency of current efforts in the combination of ontologies with rules are presented.

4.1 Rule-Based Web Applications are Still "Web Applications"

Before evaluating the existing approaches to introducing rules in Web applications, we should identify the main requirements for such approaches. These requirements mainly stem from the nature of the Web itself and its current status. Firstly, WWW is a ubiquitous and massive multi-user distributed environment. In fact, this massive characteristic is its strong point, and is a direct consequence of its architectural and technological simplicity. Web technologies such as HTTP, HTML and XML are very simple to learn and use, even for plain users who are not IT experts. Such simplicity should be taken as granted for any Web-related technology, and this applies to rule systems as well. In this context, rules for the Web should be written and used even by users/developers not familiar with advanced knowledge engineering concepts. For example, negation is a rather advanced topic in logic-based systems (with Horn rules). Hence, the semantics of rules should be such that can be easily understood by naïve users. However, this would affect their expressiveness and would limit the inference power of rule-based systems that may be required by more demanding applications (i.e., Web service discovery engines). It is worthy to mention that most of the existing rule-based applications for the Web have adopted SWRL approach (see Section 3.1) in order to express rules. SWRL is neither a highly expressive language (e.g., no negation is available) nor a decidable one, but it remains simple.

Another requirement is that the rule-based systems for the Web are compatible with existing (Semantic) Web standards, such as XML, RDF, OWL and SAWSDL (Farrell & Lausen, 2007). This means that direct support for URIs or XML syntax should be available. On the other hand, writing rules in XML (e.g., like the SWRL and RuleML approaches propose) is a cumbersome task. This is indeed a difficult decision that has to be taken by the Web architects.

Furthermore, the addition of rules in Web applications requires their compatibility with existing forms of Web knowledge that have already attained a certain maturity level. Most of the knowledge bases developed in the context of Web are expressed in the form of ontologies (mainly written in OWL or RDFS). As a result, rules have to be integrated with ontologies properly.

4.2 Qualitative Comparison of Classical Logic and Logic Programming Approaches

The aforementioned knowledge representation methodologies (see Section 3.1) are either based on classical logic (e.g., DLs and SWRL) or logic programming (e.g., answer-set programming and defeasible rules). Across the literature, several differences between classical logic and logic programming paradigms have already been identified. In this section, we intend to survey and discuss these efforts by presenting the main incompatibilities that impede the reconciliation of the two approaches.

Monotonicity vs. Non-Monotonicity

Classical logic is based on standard model theoretic semantics and adheres to monotonicity of entailment. Informally, monotonicity means that the addition of new information to the knowledge base cannot invalidate conclusions inferred by current knowledge. As a result, classical logic is able to deal with incomplete information by nature.

Instead of classical logic, logic programming has non-monotonic features. It assumes complete knowledge and there is a unique model describing the state of the world. This way, addition of knowledge may reduce the inferences.

Unique Name Assumption

Another difference between the two paradigms is that logic programming approaches typically deploy the *Unique Name Assumption* (UNA). This assumption supposes that different names represent different objects of the world. This fact imposes severe limitations on the Web context, since several distinct URIs[4] may refer to the same content or data. On the contrary, in classical logic there is no one-to-one correspondence between the names and the objects of the domain. Hence, equality between individuals represented by different names can be derived. Although it imposes a huge computational cost, most of the current description logic reasoners support reasoning with equality.

Negation

Classical logic and logic programming face the aspect of negation from different perspectives. With respect to its monotonic nature, negation in classical logic allows inferring new information only if the truth or the falsehood of a statement is explicitly declared. This fact is related with the *Open World Assumption* (OWA) of classical logic theory where incompleteness of the knowledge base is considered.

On the other hand, *Negation-As-Failure* (NAF) adheres to the *Closed World Assumption* (CWA). Specifically, if a description is not known to be true, then the truth of the negated description is drawn. In that sense, absence of knowledge draws to negated knowledge and, thus, NAF has a non-monotonic behavior. Modeling the world according to CWA seems to be somewhat inappropriate for the Semantic Web. Since knowledge on the Web is not always available (e.g., web servers breakdowns), such an assumption could lead to incorrect inferences. However, the usefulness of different types of negation in rule-based Web applications is demonstrated in (Wagner, 2003; Alferes, Damasio, & Pereira, 2003).

Constraints and Restrictions

Another aspect where the classical paradigm differs from logic programming is on the treatment of *constraints* and *restrictions*. Regardless of the apparent similarity between constraints and restrictions, there are important differences between them.

Restrictions are used in the classical logic paradigm and they constitute part of the logical theory. By adding restrictions to a knowledge base, the knowledge statements are also increased and, thus, the inference of additional knowledge is permitted.

Restrictions may be further classified to the following main categories:

- *Value restrictions* or *range restrictions*. They are used to restrict the values of a property and possibly infer new information according to their type (a.k.a. range).
- *Cardinality restrictions.* They are used for restricting the number of values that an individual may have for a specific property. They can possibly infer the existence of new instances or equality between known individuals.

In terms of Description Logics, an example of cardinality restriction in the web service paradigm could be the following:

$$\text{Profile} \sqsubseteq \geq 1 \text{ hasInput} \qquad (6)$$

Such rule denotes that every service profile should have at least one input.

On the other hand, constraints (also called *integrity constraints*) are mainly used in deductive databases and logic programming in order to check if the knowledge base is consistent with a number of specified conditions. As a result, constraints cannot draw inference of new information, but they may lead to inconsistencies in the case that some conditions specified by them are violated inside the knowledge base. Usually, constraints are rules without head (i.e., they have a "false" value in their head), denoting that the conditions stated in the body should not be satisfied concurrently by the knowledge base.

Similarly to restrictions, constraints may be divided to:

- *Value constraints* or *range constraints*. Their importance consists in checking the type of a property value.
- *Cardinality constraints*. They are used for checking the number of values that an individual may have for a specific property.

Similarly to (6), a cardinality constraint could be written in logic programming as:

$$\leftarrow \text{Profile(s)} \wedge \neg\text{hasInput(s,i)} \qquad (7)$$

More details about integrity constraints and restrictions may be found in (de Bruijn, Polleres, Lara, & Fensel, 2005), where an alternative ontology language based on the logic programming subset of OWL is presented.

Other Differences

A number of additional differences between the described approaches have also been discussed in the literature. (Eiter et al., 2006a) raises the point of non-ground entailment in the logic programming approach, declaring that, since the semantics of logic programming is defined in terms of sets of ground facts, the inference of non-factual knowledge is not allowed. Contrary to the logic programming, classical logic permits the entailment of non-factual knowledge. Furthermore, decidability issues are explored in the combination of the two perspectives. In addition, (Pattel-Schneider, 2006) describes the treatment of datatypes by classical logic approaches (e.g., DLs) and logic programming. Moreover, it examines the role of tools that support each methodology and how they can facilitate the modeling of knowledge.

The aforementioned modeling paradigms have been proposed for formalizing rules on the Web. Concerning the advantages and the disadvantages of each approach, we believe that both perspectives are useful in the context of Semantic Web.

The open nature of the classical logic seems to be more suitable for formalizing such a distributed environment. By this mean, Semantic Web technologies based on classical logic accomplish the integration of information stemming from different sources. Knowledge becomes shareable in the sense that it can be accessed without any obligation to adopt a specific model or schema. Hence, data are available to everyone through WWW so as to be retrieved and used in applications appropriately.

On the other hand, logic programming makes several assumptions in order to simplify tasks like reasoning or modeling of the knowledge base. There are several useful features of logic programming like high expressiveness, negation-as-failure and decidability of reasoning. Moreover, many of the reasoning engines targeting at managing knowledge expressed in classical logic formalisms have been recently extended to support such features (e.g., Pellet, RacerPro, Bossam). Additionally, both users and developers are very familiar with the world of databases and logic programming languages like Prolog. It is worthy to mention that even SPARQL engines return “no” in ASK query patterns with no solutions. However, “don’t know” would fit better in an open environment where multiple solutions may exist outside the current knowledge base. This way, SPARQL takes advantage of the users’ familiarity with common databases, deciding not to complicate the query answering process.

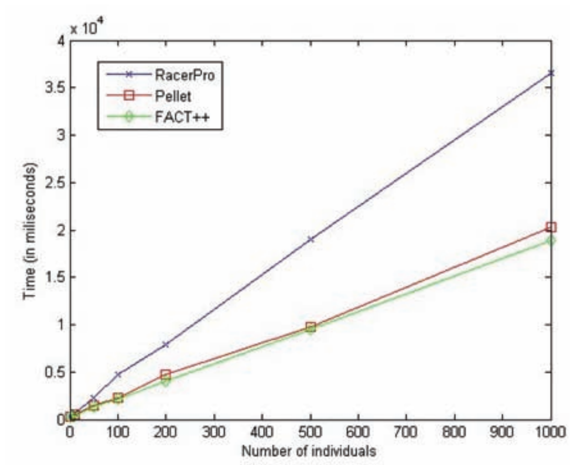
To summarize, last decade has shown to the developers of Web applications that different types of Semantic Web enabled applications requires different styles of modeling. As a result, the development of a framework that satisfy most of these needs seems to become unavoidable.

4.3 Quantitative Evaluation of Reasoning Modules

This section intends to quantify the total time required by popular reasoning modules in order to perform common reasoning tasks. Instead of (Pan, 2005; Gardiner, Horrocks, & Tsarkov, 2006), where the evaluation of description logic reasoners focuses on TBox reasoning tasks, here we examine both TBox and ABox reasoning as well as rules application. In particular, Figures 3 to 6 show the results of some indicative performance tests. These tests aim at demonstrating the efficiency of description logic reasoners and mainstream rule engines.

This quantitative evaluation involved the last stable versions of RacerPro (v1.9.0), Pellet (v1.5.1) and FACT++ (v1.1.11) description logic reasoners that were available for ontology reasoning as well as the Jess rule engine for SWRL rules application. Bossam was also tested as both ontol-

Figure 3. Evaluation of consistency check (TBox)



ogy reasoner and rule engine in this evaluation. However, Bossam's results are not displayed in the figures, since such a comparison would be improper. This stems from the fact that although Bossam constitutes a sound reasoning module it is not a complete reasoner for OWL. However, we should mention that Bossam performed well (i.e., it accomplished smaller execution times than the other OWL reasoners) handling small-size ontologies, while it threw out-of-memory errors when provided with large-scale ontologies (i.e., containing 500 and 1000 instances). Moreover, all reasoners were managed through Protégé OWL API in Java except Bossam which provides its own API. Finally, Jess was accessed via SWRL-Jess bridge (O'Connor et al., 2005).

The knowledge base used in these tests stems from the Web service description paradigm and it consists of a domain ontology and OWL-S service descriptions. The domain ontology, which represents the domain of football (i.e., player/team details and statistics), includes concepts described through necessary and sufficient conditions. The elements of this ontology are used for the

Figure 4. Evaluation of classification (TBox)

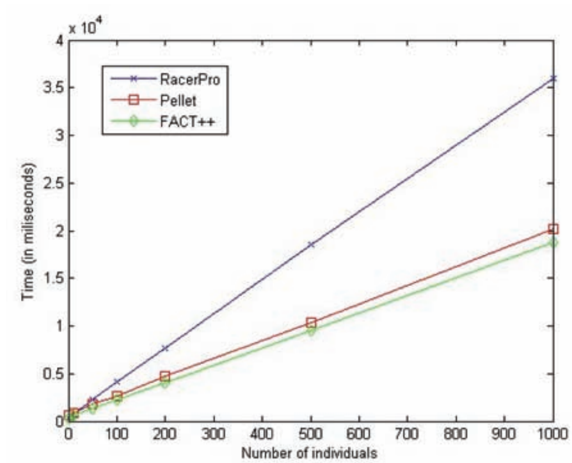
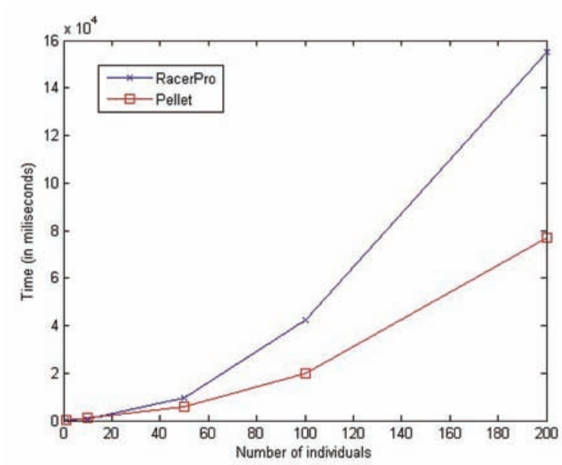


Figure 5. ABox reasoning evaluation (computation of inferred individual types)

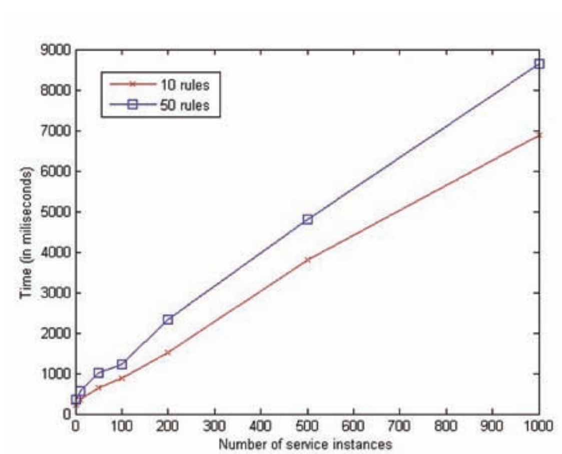


annotation of some Semantic Web services that provide search functionality to some underlying football statistics databases. Specifically, the service inputs and outputs are annotated through concepts of this domain ontology. The extract of the domain ontology used contains 37 concepts, 8 object properties and 3 datatype properties. Each service description involves 19 concept instances and 11 property instances.

Figure 3 depicts the performance of three description logic reasoners while checking the consistency of all classes over the aforementioned knowledge base. The execution times presented in this figure do not involve the time needed to load the models. The figure shows that these times increase as the number of instances contained in the knowledge base also increases. FACT++ and Pellet attained the best performance by achieving similar execution times. On the other hand, RacerPro performed worse than the others.

In Figure 4, the times required for classifying the hierarchy of the ontology are presented. Again, the times shown in the figure do not include the loading time of each model. The resulting perfor-

Figure 6. Jess performance executing SWRL rules



mances are very similar to those of the consistency checking process. FACT++ has slightly better performance than Pellet, while RacerPro has the worst performance again.

Figure 5 shows the efficiency of RacerPro and Pellet reasoners in realizing reasoning over ABoxes. In this figure we do not display FACT++, since it provides incomplete ABox reasoning. In particular, the experiment concerns the identification of the types (i.e., classes) that each individual belongs to. The figure shows that even the reasoning over an ontology containing 100 instances using Pellet (which seems to be the most efficient reasoner in this reasoning task) requires about 20 seconds. Moreover, both reasoners failed to compute inferred individual types for ontologies containing 500 and 1000 individuals.

It is worthy to mention that version 1.9.2 Beta of RacerPro was also available during the experiments. This version seemed to remain slower than FACT++ and Pellet in TBox reasoning tasks (consistency check and classification of hierarchy) while it performed better in ABox reasoning. However, the results are not displayed in the figures above, since this is not a stable version of the reasoner.

Figure 6 shows the time required by Jess rule engine to apply SWRL rules over ontologies containing various numbers of instances. These rules represent service composition constraints that should apply to the inputs and outputs of the services. We applied two different sets of rules to the previously described knowledge bases: the first set consisted of 10 rules while the second of 50 rules. In both cases, the execution time increases proportionally to the number of instances contained in the ontology. As expected, the delay imposed by the execution of rules was affected by the number of rules. However, the number of instances seems to affect execution times more than the number of rules. Hence, the inference process is complicated when real-time demands come up in distributed environments where applications have to support multiple individual instances.

To summarize, the time required to handle a large KB and execute reasoning in terms of Semantic Web technologies like DLs constitutes a limiting factor, even for the most efficient reasoners. This is a fundamental problem on the Web, since applications have to deal with a large number of instances representing data. Especially, when considering real time applications that call for small response times, the aforementioned results become unmanageable. This is the main reason why many researchers have recently focused on developing efficient and scalable reasoning applications over large individual sets. Instance Store (Horrocks, Li, Turi, & Bechhofer, 2004a) is such an approach that permits working in conjunction with any ontology reasoner that implements the DIG interface.

5. FUTURE TRENDS AND OPEN ISSUES

Several open issues that deserve further research efforts in the future have been identified throughout the chapter. Firstly, a Web standard for editing and embedding rules into Web applications should be specified. In our opinion, such a language should focus on its simplicity, instead of providing major expressiveness capabilities that will lead to a clumsy formalism. Rule Interchange Format (RIF) Working Group (RIF, 2008) works in this direction in order to produce W3C recommendations for enabling interchange of rules. Specifically, RIF does not intend to provide explicit mappings between various rule languages, but specify a mechanism capable of defining the meaning of the formulas of a rule language. This way, rules could be automatically translated across different formalisms.

Additionally, more tractable algorithms for Web-rules, TBox and ABox inference problems should be developed. These algorithms should be able to handle complex and heavy knowledge bases that combine rules and ontologies with a large number of concepts and relationships. Such knowledge bases are substantial, especially in the context of information integration where multiple domain vocabularies interact with upper-level ontologies. Furthermore, the Web is a large scale environment, where a huge set of resources are added every day. As a result, future knowledge bases will have to capture and describe more and more individuals. These facts should be taken into account by the developers that will target at designing new reasoning algorithms.

Finally, a more practical issue that should be investigated is the development of a unified reasoning framework capable of managing both ontologies and rules. Today, there is no efficient and easy-to-use integrated reasoning module that can reason over both formalisms. As a consequence, the developer should use at least two different reasoning modules to handle such an integrated knowledge base. This can result to unexpected situations. For example, the restrictions defined by the ontology can be violated by the application of rules, since the rule engines do not take into account these restrictions (e.g., disjointness of concepts). Similarly, the application of rules could produce knowledge that would be useful for further description logic inferences.

6. CONCLUSION

In this chapter we have discussed the application of rules to Web applications in order to achieve intelligent application behavior and efficient management of knowledge. We have described the main methodologies and technologies for integrating rules with ontologies, since the latter constitutes a mature knowledge technology on the Web. The chapter has also examined aspects of the different approaches. We also showed through several experiments that current reasoning modules are not efficient enough to manage knowledge stemming from large-scale environments like the Web, especially in the context of real-time applications which impose severe constraints in response times. Finally, more improvements should be made in the standardization of rules formalism on the Web and in the development of reasoning modules that can handle and reason over both ontologies and rules as integral knowledge.

7. REFERENCES

W3C Semantic Web Activity (2008). Retrieved April 22, 2008, from http://www.w3.org/2001/sw/.

Alferes, J., Damasio, C., & Pereira, L. (2003). Semantic Web logic programming tools. (. *LNCS*, *2901*, 16–32.

Antoniou, G., Billington, D., Governatori, G., & Maher, M., J. (2001). Representation results for defeasible logic. *ACM Transactions on Computational Logic*, *2*(2), 255–287. doi:10.1145/371316.371517

Antoniou, G., & Damasio, C. V., Grosof, B., Horrocks, I., Kifer, M., Maluszynski, J., et al. (2005). *Combining rules and ontologies: A survey*. Deliverables I3-D3, REWERSE. Retrieved from, http://rewerse.net/deliverables/m12/i3-d3.pdf.

Antoniou, G., & van Harmelen, F. (2004). *A Semantic Web primer*. MIT Press.

Baader, F., Calvanese, D., McGuiness, D., Nardi, D., & Patel-Schneider, P. (2003). *The description logic handbook: Theory, implementation, and applications*. Cambridge: Cambridge University Press.

Bassiliades, N., Antoniou, G., & Vlahavas, I. (2006). A defeasible logic reasoner for the Semantic Web . *International Journal on Semantic Web and Information Systems*, *2*(1), 1–4.

Bechhofer, S., Moller, R., & Crowther, P. (2003). The DIG description logic interface. In D. Calvanese, G. de Giacomo, & F. Franconi (Eds.), *Proc. of the 2003 Int. Workshop on Description Logics (DL 2003)* (Vol. 81 of CEUR Workshop Proceedings), Rome, Italy.

Boley, H., Kifer, M., & Patranjan, P. L., & Polleres, A. (2007). Rule interchange on the Web. In *Reasoning Web 2007* (No. 4636, pp. 269-309). Springer.

Brickley, D., & Guha, R. V. (2004). *RDF vocabulary description language 1.0: RDF schema*. Recommendation 10 February 2004, W3C. Retrieved April 22, 2008, from http://www.w3.org/TR/rdf-schema/.

Cardoso, J. (2007). The Semantic Web vision: Where are we? *Intelligent Systems*, *22*(5), 84–88. doi:10.1109/MIS.2007.4338499

CLIPS. *A tool for building expert systems*. (2008). Retrieved April 22, 2008, from http://clipsrules.sourceforge.net/.

de Bruijn, J., Lausen, H., Krummenacher, R., Polleres, A., Predoiu, L., Kifer, M., et al. (2005). The Web service modeling language WSML (Tech. Rep. WSML. WSML Final Draft D16.1v0.21). Retrieved from http://www.wsmo.org/TR/d16/d16.1/v0.21/.

de Bruijn, J., Polleres, A., Lara, R., & Fensel, D. (2005). OWL DL vs. OWL flight: Conceptual modeling and reasoning on the Semantic Web. In *Proc. WWW2005*, (pp. 623-632).

Dean, M., Schreiber, G., Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., et al. (2004). *OWL Web ontology language reference*. W3C Recommendation 10 February 2004. Retrieved April 22, 2008, from http://www.w3.org/TR/owl-ref/.

Donini, M. F., Lenzerini, M., Nardi, D., & Schaerf, A. (1998). AL-log: Integrating datalog and description logics. *Journal of Intelligent Information Systems*, *10*(3), 227–252. doi:10.1023/A:1008687430626

Drools (2008). Retrieved April 22, 2008, from http://www.jboss.org/drools/.

Eiter, T., & Ianni, G. B., Schindlauer, R., & Tompits, H. (2005). A uniform integration of higher-order reasoning and external evaluations in answer set programming. In *Proc. IJCAI 2005*. Morgan Kaufmann.

Eiter, T., & Ianni, G. B., Polleres, A., Schindlauer, R., & Tompits, H. (2006a). Reasoning with rules and ontologies. In P. Barahona, F. Bry, E. Franconi, U. Sattler, & N. Henze (Eds.), *Reasoning Web, Second International Summer School 2005, Tutorial Lectures* (LNCS, pp. 93-127).

Eiter, T., & Ianni, G. B., Schindlauer, R., & Tompits, H. (2006b). dlvhex: A system for integrating multiple semantics in an answer-set programming framework. In M. Fink, H. Tompits, & S. Woltran (Eds.), *Proceedings 20th Workshop on Logic Programming and Constraint Systems (WLP 06)*, (pp. 206-210).

Eiter, T., Lukasiewicz, T., Schindlauer, R., & Tompits, H. (2004). Combining answer set programming with description logics for the Semantic Web. In *Proc. of the International Conference of Knowledge Representation and Reasoning* (KR04).

Farrell, J., & Lausen, H. (2007). *Semantic annotations for WSDL and XML schema*. W3C Recommendation, W3C, August 2007. Retrieved April 22, 2008, from http://www.w3.org/TR/sawsdl/.

Forgy, C. (1982). Rete: A fast algorithm for the many pattern/many object pattern match problem. *Artificial Intelligence*, *19*, 17–37. doi:10.1016/0004-3702(82)90020-0

Gardiner, T., Horrocks, I., & Tsarkov, D. (2006). Automated benchmarking of description logic reasoners. In *Proc. of DL*.

Gelfond, M., & Lifschitz, V. (1991). Classical negation in logic programs and disjunctive databases. *New Generation Computing*, *9*(3-4), 365–386. doi:10.1007/BF03037169

Grosof, B. N., Horrocks, I., Volz, R., & Decker, S. (2003). Description logic programs: Combining logic programs with description logic. In *Proc. of the Twelfth International World Wide Web Conference (WWW 2003),* (pp. 48-57). ACM.

Haarslev, V., & Moller, R. (2001). RACER system description. In R. Gor'e, A. Leitsch, and T. Nipkow (Eds.), *International Joint Conference on Automated Reasoning, IJCAR'2001, June 18-23, Siena, Italy,* (pp. 701–705). Springer-Verlag.

Horrocks, I. (1998). Using an expressive description logic: FaCT or fiction? In *Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning (KR'98), Trento, Italy, June 2–5, 1998*, (pp. 636–647).

Horrocks, I., Li, L., Turi, D., & Bechhofer, S. (2004a). The instance store: DL reasoning with large numbers of individuals. In *Proc. of the 2004 Description Logic Workshop (DL~2004),* (pp. 31-40).

Horrocks, I., Parsia, B., & Patel-Schneider, P. F., & Hendler, J. (2005). Semantic Web architecture: Stack or two towers? In *Proc. PPSWR 2005*, (pp. 37-41), Dagstuhl Castle, Germany.

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004b). *SWRL: A Semantic Web rule language combining owl and RuleML*. W3C Member Submission, 21 May 2004. Retrieved April 22, 2008, from http://www.w3.org/Submission/SWRL/.

Hustadt, U., Motik, B., & Sattler, U. (2004). Reducing SHIQ^{-} description logic to disjunctive datalog programs. In *Proc. of the 9th International Conference on Knowledge Representation and Reasoning (KR2004),* (pp. 152-162).

ILOG. *Business Rule Management Systems, Optimization Tools and Engines, Visualization Software Components, Supply Chain Applications* (2008). Retrieved April 22, 2008, from http://www.ilog.com/.

Jang, M., & Sohn, J.-C. (2004). Bossam: An extended rule engine for OWL inferencing. In *Workshop on Rules and Rule Markup Languages for the Semantic Web at the 3rd International Semantic Web Conference* (LNCS 3323, pp. 128-138).

Jess, The Rule Engine For the Java Platform (2008). Retrieved April 22, 2008, from http://www.jessrules.com/jess/index.shtml

KAON2 – Ontology Management for the Semantic Web (2008). Retrieved April 22, 2008, from http://kaon2.semanticweb.org.

KAON – The KArlsruhe ONtology and Semantic Web tool suite (2008). Retrieved April 22, 2008, from http://kaon.semanticweb.org.

Kifer, M., de Bruijn, J., Boley, H., & Fensel, D. (2005). A realistic architecture for the Semantic Web. In *Proc. RuleML 2005* (pp. 17-29), Galway, Ireland.

Klyne, G., & Carroll, J. J. (2004). *Resource description framework (RDF): Concepts and abstract syntax*. Recommendation 10 February 2004, W3C. Retrieved April 22, 2008, from http://www.w3.org/TR/rdf-concepts/.

McBride, B. (2002). Jena: A Semantic Web toolkit . *IEEE Internet Computing*, *6*(6), 55–59. doi:10.1109/MIC.2002.1067737

Mei, J., Bontas, E. P., & Lin, Z. (2005). OWL2Jess: A transformational implementation of the OWL semantics. In *Proceedings of International Workshops on ISPA* (LNCS 3759, pp. 599-608).

Motik, B., Horrocks, I., Rosati, R., & Sattler, U. (2006). Can OWL and logic programming live together happily ever after? In *Proceedings ISWC-2006* (LNCS 4273, pp. 501–514).

Motik, B., Sattler, U., & Studer, R. (2005). Query answering for OWL-dl with rules. *Journal of Web Semantics: Science . Services and Agents on the World Wide Web*, *3*, 41–60. doi:10.1016/j.websem.2005.05.001

Nute, D. (1994). Defeasible logic. In *Handbook of Logic in Artificial Intelligence and Logic Programming, Nonmonotonic Reasoning and Uncertain Reasoning* (Vol. 3). Oxford University Press.

O'Connor, M., Knublauch, H., Samson, T., & Musen, M. (2005). Writing rules for the Semantic Web using SWRL and Jess. *Protégé with Rules WS*. Madrid.

Pan, J. (2005). Benchmarking DL reasoners using realistic ontologies. In *Proc. of the OWL: Experiences and Directions Workshop*.

Pan, J. Z., & Horrocks, I. (2004). *OWL-E: Extending owl with expressive datatype expressions* (Technical report, IMG/2004/KR-SW-01/v1.0). Victoria University of Manchester.

Parsia, B., Sirin, E., & Kalyanpur, A. (2005). Debugging OWL Ontologies. In *Proceedings of the 14th International World Wide Web Conference (WWW2005), Chiba, Japan, May 2005*. Retrieved from http://www.mindswap.org/papers/debuggingOWL.pdf.

Patel-Schneider, P. F., & Horrocks, I. (2006). Position paper: a comparison of two modelling paradigms in the Semantic Web. In *Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26, 2006). WWW '06, New York, NY* (pp. 3-12). ACM Press.

Polleres, A., & Schindlauer, R. *(2007)*. DLVHEX-SPARQL: A SPARQL Compliant Query engine based on DLVHEX. In Proceedings of 2nd International Workshop on Applications of Logic Programming to the Web, Semantic Web and Semantic Web Services (ALPSWS2007), Porto, Portugal.

Prud'hommeaux, E., & Seaborne, A. (2005). *SPARQL query language for RDF.* http://www.w3.org/TR/2005/WD-rdf-sparql-query-20050217/, 2005.

RacerPro (2008). Retrieved April 22, 2008, from http://www.racer-systems.com/products/racerpro/index.phtml.

Rainer. A. (2005). Web service composition under answer set programming. *KI-Workshop "Planen, Scheduling und Konfigurieren, Entwerfenl" (PuK).*

RIF - Rule Interchange Format Working Group (2008). Retrieved April 22, 2008, from http://www.w3.org/2005/rules/wiki/RIF_Working_Group

Rosati, R. (2006a). DL+log: Tight integration of description logics and disjunctive datalog. In P. Doherty, J. Mylopoulos, & C.A. Welty (Eds.), *Proceedings of the 10th International Conference on Principles of Knowledge Representation and Reasoning (KR 2006), Lake District, UK* (pp. 68-78). AAAI Press.

Rosati, R. (2006b). Integrating ontologies and rules: Semantic and computational issues. In P. Barahona, F. Bry, E. Franconi, N. Henze, & U. Sattler (Eds.), *Reasoning Web, Second International Summer School 2006, Lissabon, Portugal, September 4_8, 2006, Tutorial Lectures*. (LNCS 4126, pp. 128-151).

Rosati, R. (2006c). DL+log: Tight integration of description logics and disjunctive datalog. In *Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2006)* (pp. 68-78). AAAI Press.

Rule, M. L. *Rule Markup Initiative* (2008). Retrieved April 22, 2008, from http://www.ruleml.org/.

Sirin, E., Parsia, B., & Grau, B., C., Kalyanpur A., & Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. *Journal of Web Semantics*, *5*(2). doi:10.1016/j.websem.2007.03.004

Tsarkov, D., & Horrocks, I. (2006). FaCT++ description logic reasoner: System description. In *Proc. of the International Joint Conference on Automated Reasoning (IJCAR 2006).*

Wagner, G. (2003). Web rules need two kinds of negation (LNCS 2901, pp. 33-50).

KEY TERMS AND DEFINITIONS

Logic Programming: It is a declarative paradigm of programming that is mainly based on first-order logic. Prolog is the most common logic programming language.

Knowledge-Based System: A system that exploits knowledge representation and reasoning techniques in order to achieve an intelligent behavior.

Knowledge Representation: It is the process of exploiting formal methodologies and languages (e.g., rules formalisms, ontologies) in order to capture and describe knowledge.

Ontology: A formal representation of a domain of discourse that describes concepts and relationships among them. This way, it provides a common vocabulary and allows for the inference of new knowledge.

Reasoning: The systematic process of inferring new knowledge by applying formal implication rules to a given knowledge base.

Rule-Based Application: An application that takes advantage of rules formalisms (e.g., Horn-clauses) in order to represent and infer new knowledge.

Semantic Web: It is an extension of the current Web where resources are described through formal syntax and semantics in order to be human- and machine-readable.

ENDNOTES

1. Also called as Datalog view in the literature
2. http://con.fusion.at/dlvhex/
3. It offers free trials while educational and research licenses are available
4. Universal Resource Identifiers

This work was previously published in Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches, edited by A. Giurca, D. Gasevic, and K. Taveter, pp. 371-392, copyright 2009 by Information Science References (an imprint of IGI Global).

Chapter 2.13
Modeling of Web Services using Reaction Rules

Marko Ribarić
Mihailo Pupin Institute, Serbia

Shahin Sheidaei
Simon Fraser University, Canada

Milan Milanović
University of Belgrade, Serbia

Dragan Gašević
Athabasca University, Canada

Adrian Giurca
Brandenburgische Technische Universität Cottbus, Deutschland

Sergey Lukichev
Brandenburg University of Technology at Cottbus, Germany

Gerd Wagner
Brandenburg University of Technology, Germany

ABSTRACT

The development process of Web services needs to focus on the modeling of business processes rather than on low-level implementation details of Web services, and yet it also needs to incorporate the support for frequent business changes. This chapter presents the UML-based Rule Language (URML) and REWERSE Rule Markup Language (R2ML), which use reaction rules (also known as Event-Condition-Action rules) for modeling Web services in terms of message exchange patterns. Web services that are being modeled in this way can easily be integrated in the wider context of modeling orchestration and choreography. In order to achieve proposed solution, we have developed a plug-in for the Fujaba UML tool (so called Strelka) and a number of model transformations for round-trip engineering between Web services and reaction rules. Also, the paper presents mappings of models of Web services with reaction rules into the Drools rule language, thus enabling the run time execution semantics for our rule-based models.

DOI: 10.4018/978-1-60566-402-6.ch018

1. INTRODUCTION

Web services provide a communication interface and a workflow management protocol for business systems. Their goal is set to change the Web from a static collection of information, to a dynamic place where different software components (business logics) can be easily integrated (Arroyo et al., 2004). Some of the factors that constrain web services from achieving that goal can be named as follow: when developers try to enrich the existing Web services with some new functionality, they have to implement necessary changes manually, which may lead to potential execution errors; furthermore, business rules, due to their declarative nature, are used for reflecting dynamic changes representing desired alteration in the business process. When business rules are used in a networked business, they are often large and complex, which makes them difficult to manage and change. There is a lack of automatic mechanism for updating Web services based on the business rules changes (Ribaric et al., 2008).

The need for a more straightforward way to express, and manage business rules as a separate part of Web services' composition has been recognized by researchers for the numerous benefits that it provides (Charfi & Mezini, 2004). However, integrating rule-based systems in a service-oriented environment is a complex task, due to the fact that both worlds have their own paradigms (Rosenberg & Dustdar, 2005). Many different factors involve in this integration. However, the most important one of these factors is the architecture of service oriented systems. In this architecture, layers play a great role. The layering architecture is needed to be represented and deployed in the world of rules which become challenging and could not be simple done. This integration approach should be loosely coupled, meaning that it is reasonable to expose business rules as services. The way of modeling Web services presented in this paper facilitates this integration process, and those services can be more easily integrated in the wider context of modeling orchestration and choreography.

A promising way to solve these problems is to use a high-level modeling approach combined with the use of rules. A modeling approach will allow developers to focus on a problem domain rather than on an implementation technology. This is why we propose using an approach based on Model Driving Engineering (MDE). Although, there have been several attempts to leverage MDE principles to model Web services, they are still very low level oriented, as they again focus on technical details covered either by WSDL (Bezivin et al., 2004) (Vara et al., 2005) or OWL-S (Timm & Gannod, 2005) (Gronmo et al., 2005). The use of rules implying that they are defined in a declarative way, they can dynamically reflect business logic changes at run-time without the need to redesign the whole system. Since Web services are used for integration of business processes of various stakeholders, it is important for them to reflect changes in business logic, or policies, as good as possible.

The solution that we propose is to use rules to represent and model business processes from the perspective of message exchange. That is, our modeling approach enables one to model Web services from the perspective of the underlying business logic regulating *how* Web services are used regardless of the context *where* they are used. To do so, our proposal is to leverage message-exchange patterns (MEPs) as an underlying perspective integrated into a Web service modeling language. This perspective has already been recognized by Web service standards (Chinnici et al., 2007), and our contribution is to raise this to the level of models.

As the leading Rule Modeling language we benefit from UML-based Rule Modeling Language (URML) (Lukichev & Wagner, 2006a) for our approach. URML having strong ties with R2ML (REWERSE Rule Markup Language) empowers us to model rules in various languages.

URML supports modeling of domain vocabularies in addition to integrity, derivation, production and reaction rules. .Business rules, being similar to natural language rules, are proposed to be modeled by using reaction rules (also known as Event-Condition-Action, ECA, rules) in this paper. Reaction rules will perform a task (Action) under some circumstances (Conditions) if some events (Event) take place and can best describe a business rule.

We also have developed a plugin (called Strelka) for the well known UML tool Fujaba, which has a support for URML notation. Strelka also has the support transformation of URML models into Web Service Description Language (WSDL) and different types of rule languages such as Drools, Jess, and SWRL. In this paper, we discuss those transformations that are implemented by using a model transformation language entitled the ATLAS Transformation Language (ATL) (ATL, ver. 0.7).

The paper is structured as follows: in the next section we give a bit of the background, including, Web services, R2ML and URML languages with the emphasis on reaction rules. In section 3 we introduce our approach for modeling business rules and services following by section 4 in which we introduce Strelka (Lukichev & Wagner, 2006b), a tool we developed that is capable of presenting URML diagrams, and performing transformations. Section 5 describes mapping process between URML and WSDL that is a part of Strelka, and also covers a deployment of URML diagrams to the Drools rule engine. Section 6 summarized the related work done so far and in Section 7 we conclude this work by summarizing the major points.

2. BACKGROUND

In this section, we give a brief overview of the technologies and techniques relevant to the problem under study. This includes a short description of Web services, Drools rule engine, and rule language definitions based on MDE principles used in our modeling approach. A complete introduction into the ATLAS Transformation Language (ATL) can be found in the chapter entitled “Sharing ontologies and rules using model transformations”.

Web Services

A Web service is a loosely coupled component that exposes functionality to a client over the Internet (or an intranet) by using web standards such as HTTP, XML, SOAP, WSDL, and UDDI (Timm & Gannod, 2005).

SOAP is an XML-based protocol for exchanging information in a decentralized, distributed environment. SOAP builds on XML and common Web protocols (HTTP, FTP, and SMTP) (Coyle, 2002). A SOAP message is the basic unit of communication between SOAP nodes. The “envelope” element represents the root of a SOAP message structure. It contains a mandatory body construct and optional header construct (Gudgin et al., 2007). The header construct is where meta-information can be hosted. In a large number of service-oriented architectures, the header is an important part of the overall architecture, and although it is optional it is rarely omitted.

Web Service Description Language (WSDL) is a language for describing both the abstract functionality of a service and the concrete details of a Web service (Chinnici et al., 2007). At an abstract level, WSDL describes a Web service in terms of interfaces and the operations supported by the interfaces. An operation is an interaction with the service consisting of a set of (input, output, infault and outfault) messages exchanged between the service and other parties involved in the interaction (Chinnici et al., 2007). The messages are described independently of a specific wire format by using a type system, typically XML Schema. WSDL also describes the point of contact for a service provider, known as the endpoint – it provides a formal definition of the endpoint interface and also establishes the physical location (address) of the service.

Potential requestors need a way to discover Web services descriptors. It is necessary that these descriptors are collected and stored in a central registry. The key part of the Universal Description Discovery and Integration (UDDI) specification (Clement et al., 2004) presents standardizing information inside such a registry as well as specifying the way the information can be searched and updated.

Regardless of how complex tasks performed by a Web service are, almost all of them require the exchange of multiple messages (Erl, 2005). It is important to coordinate these messages in a particular sequence, so that the individual actions performed by the message are executed properly. Message exchange patterns (MEPs) are a set of templates that provide a group of already mapped out sequences for the exchange of messages (Erl, 2005). This basically means that MEPs define how services should be used, as they can coordinate input and output messages related to a certain operation.

The WSDL 2.0 specification defines three MEPs:

- in-only pattern – supports a standard fire-and-forget pattern (i.e., only one message is exchanged);
- robust in-only pattern – presents a variation of the in-only pattern that provides an option of sending a fault massage, as a result of possible errors generated while transmitting, or processing data;
- in-out pattern – presents a request-response pattern where two messages (input and output) must be exchanged.

However, WSDL 2.0 specification offers the possible use of five more patterns (Lewis, 2007):

- out-in pattern – also consists of exactly two messages (like in-out pattern), but now a service provider is the one that initiates the exchange of messages;
- out-only – consists of exactly one message (like in-only pattern), and is most often used for message notification;
- robust out-only - presents a variation of the out-only pattern that provides an option of sending a fault massage;
- in-optional-out – is similar to pattern in-out with one exception: sending an message that represents a response is optional, and because of this a requester that has started a communication should not expect this message;
- out-optional-in – is opposite to the previous pattern, where the input message is optional.

Drools

Drools is a business rule management system (BRMS) with a forward chaining inference based rule engine, more correctly known as a production rule system, using an enhanced implementation of the Rete algorithm. BRMS is a production rule system with a focus on knowledge representation to express propositional and first order logic in a concise, non ambiguous and declarative manner. The core of a production rules system is an inference engine that is able to scale to a large number of rules and facts. The inference engine matches facts and data, against production rules, also called productions, to infer conclusions which result in actions. A production rule is a two-part structure using first order logic for knowledge representation (Proctor et al). Figure 1 presents a schema of a production rule. It executes the actions while the conditions are being satisfied. Drools representation contains the same structure, LHS (Left Hand Side) is referring to conditions while RHS (Right Hand Side) is referring to actions. Name of the rule which is presented by “name”

Figure 1. First order logic representation (left) - Drools representation (right)

```
when
    <conditions>
then
    <actions>
```

```
rule "name"
        attributes
        when
                LHS
        then
                RHS
end
```

is just for better understanding of rules and have no functional value. Attributes, marked with attributes in the rule are used for the rule internal functionality itself.

The process of matching new or existing facts against production rules is called *pattern matching*, which is performed by the inference engine (part of Drools). Drools implements and extends the Rete algorithm The Drools Rete implementation is called ReteOO, signifying that Drools has an enhanced and optimized implementation of the Rete algorithm for object oriented systems (Schneier). For better understanding of rules and their structure please refer to the chapter "Object Oriented Rule Languages".

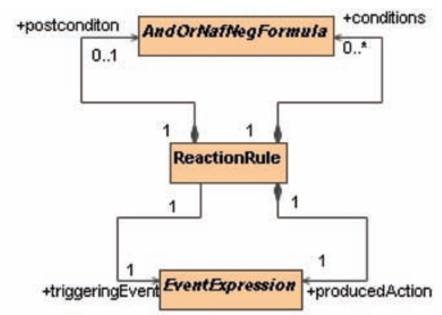
R2ML

Existing Web rule languages (RuleML (Hirtle et al., 2006) and SWRL (Horrocks et al., 2004)) and rule modeling approaches (SBVR (SBVR, 2006) and PRR (Charfi & Mezini, 2004)) very nicely serve the purpose of rule interchange and rule representation independent of any specific platform, respectively. However, none of these languages offers a suitable modeling foundation that can be used for modeling Web services from a more abstract perspective of business processes and business rules in which they are used. That is, the language should be able to model reaction rules. Yet, such a modeling language should be closely related to the already existing software modeling languages such as UML and preferably defined by following MDE principles. Finally, the language should be closely related to the Web standards for defining ontologies (OWL) and rules (RIF).

We have decided to use REWERSE I1 Rule Markup Language (R2ML), as it fully satisfies the above requirements. The language is defined by a MOF-based metamodel which is refined by OCL constraints that precisely define relations between the language's constructs in addition to those defined by the metamodel; it has an XML schema defined concrete syntax; it has a UML-based graphical concrete syntax, so called UML-based Rule Modeling Language (URML), which we will introduce it in the section below; and it has a number of transformations with other rule languages (e.g., JBoss' Drools and OCL) allowing us to translate Web service models to the rule-based languages that can then regulate the use of Web services.

R2ML is a rule language that addresses all the requests defined by the W3C working group for the standard rule interchange format (Ginsberg et al., 2006). The R2ML language can represent different types of rule constructs, that is, it can represent different types of rules (Wagner et al., 2006) including: integrity rules, derivation rules, production rule, and reaction rules. Integrity rules in R2ML, also known as (integrity) constraints, consist of a constraint assertion, which is a sentence in a logical language such as first-order predicate logic or OCL. Derivation rules in R2ML are used to derive new knowledge (conclusion) if a condition holds. Production rules in R2ML produce actions if the conditions hold, while post-conditions must also hold after the execution of actions. A reaction rule is a statement of programming logic (Giurca et al., 2006) that specifies the execution of one or more actions in the case of a triggering event occurrence and if rule conditions are satisfied. Optionally, after the execution of the action(s), post-conditions may be made true. R2ML also allows one to define vocabularies by using the

Figure 2. Definition of reaction rules in the R2ML metamodel

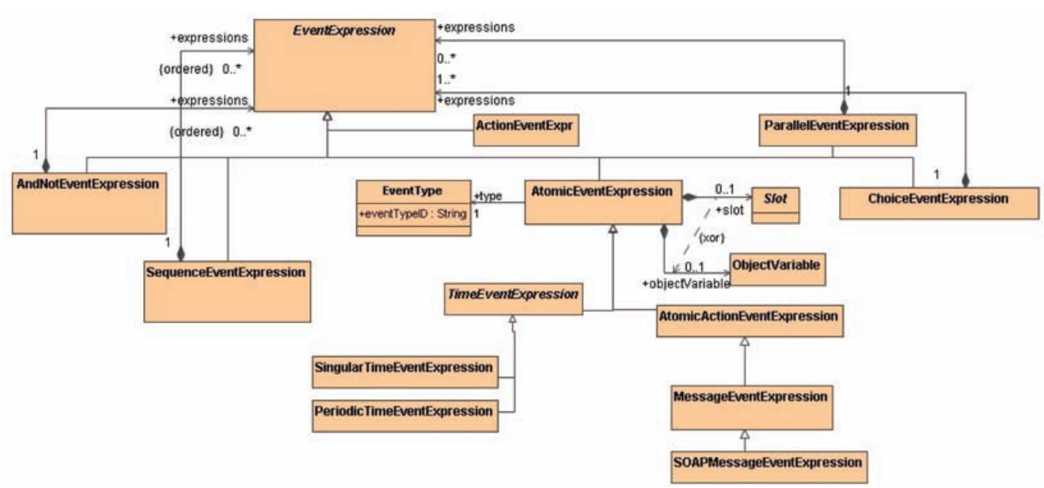


following constructs: basic content vocabulary, functional content vocabulary, and relational content vocabulary. Here we give short description of vocabulary constructs that we use in this paper. Vocabulary is a concept (class) that can have one or more VocabularyEntry concepts. VocabularyEntry is abstract concept (class) that is used for representing other concepts by its specialization. For example, one of the VocabularyEntry-s is an R2ML Class concept which represents the class element similar to the notion of the UML Class. An R2ML Class can have attributes (class Attribute), reference properties (class ReferenceProperty) and operations (class Operation).

Due to the importance for our Web service modeling approach, here we only describe the details of R2ML reaction rules. Reaction rules represent a flexible way for specifying control flows, as well as for integrating events/actions from a real life (Giurca et al., 2006). Reaction rules are represented in the R2ML metamodel as it is shown in Figure 2: triggeringEvent is an R2ML EventExpression (Fig. 4); conditions are represented as a collection of quantifier free logical formulas; producedAction is an R2ML EventExpression and represents a system state change; and (optional) postcondition must hold when the system state changes.

The R2ML event metamodel defines basic concepts that are needed for dynamic rule behavior (Figure 3). There are two types of events: atomic event (AtomicEventExpression) and composite events (AndNotEventExpression, SequenceEventExpression, ParallelEventExpression or ChoiceEventExpression). Characteristic of an atomic event is that it has no duration (duration = 0). For the sake of modeling Web services, we are using MessageEventExpression for both triggering events and produced actions. MessageEventExpression is used for modeling messages that are part of the definition of Web services, including, input, output, in-fault, and out-fault messages. Each MessageEventExpression has its

Figure 3. Event expressions in the R2ML metamodel



own type – EventType. In terms of WSDL, message types are defined by XML Schema complex types, while in R2ML, EventType is a subclass of Class (already defined as a part of the R2ML Vocabulary). This means that each EventType has its own attributes, associations, and all other features of R2ML classes.

URML

UML-Based Rule Modeling Language (URML) is a graphical concrete syntax of R2ML. URML is developed as an extension of the UML metamodel to be used for rule modeling. In URML, modeling vocabularies is done by using UML class models. Rules are defined on top of such models. The URML reaction rules metamodel, which we use for modeling services, is shown in Figure 4a. The figure shows components of a reaction rule: Condition, Postcondition, RuleAction and EventCondition. The figure also shows that reaction rules are contained inside the UML package which represents Web services operation. This means, that such packages have a stereotype <<operation>> in UML diagrams. An instance of the EventCondition class is represented on the URML diagram as incoming arrow (e.g., see Figure 5), from a UML class that represents either an input message or an input fault message of the Web service operations, to the circle that represents the reaction rule. The UML class that represents the input message (input-Message in Figure 4b of the Web service operation is MessageEventType (a subclass of EventType from Figure 3) and it is represented by using the <<message event type>> stereotype on UML classes. The UML class that represents the input fault message (inFault in Figure 4b) of the Web service operation is FaultMessageEventType in the URML metamodel. In URML diagrams, FaultMessageEventType is represented by the <<fault message event type>> stereotype on UML classes. EventCondition contains an object variable (ObjectVariable in Figure 4c), which is a placeholder for an instance of the MessageEventType class. The object variable has a name that corresponds to the arrow annotation, which represents EventConditon.

An instance of the RuleAction class is represented as an outgoing arrow on the URML diagram, from the circle that represents the reaction rule to the class that represents either an output message or an output fault message of the Web service operation. The UML class that represents the output message (outputMessage in Figure 4c) of the Web service operation is MessageEventType and it is represented with the <<message event type>> stereotype on UML classes. The UML class that represents the output fault message (outFault) of the Web service operation is FaultMessageEventType in the URML metamodel and it is represented with the <<fault message event type>> stereotype on UML classes. RuleAction contains an object

Figure 4. a) Extension of the URML metamodel for reaction rules; b) Part of the URML meta-model for EventCondition; c) Extension of the URML metamodel for actions

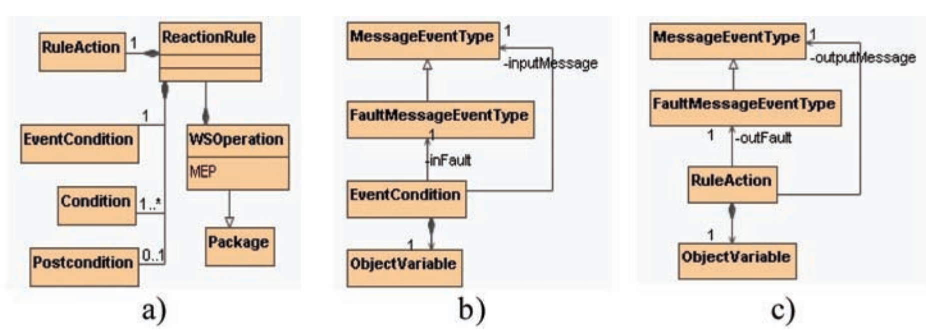
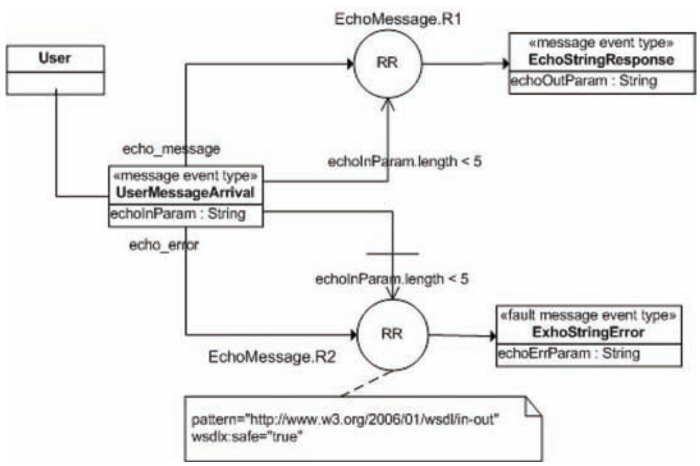


Figure 5. PIM level of the in-out message exchange pattern with out-fault, presented in URML



variable (ObjectVariable), which represents an instance of the MessageEventType class.

3. MODELING

In this section the idea behind modeling of rules and business processes will be exposed. Modeling will be performed with the help of URML and R2ML. In section 5 with the use of modeling, the transformation to the web services and automatic generation of codes will be described.

Rule Modeling

As stated, in this paper we are concerned with reaction rules. Reaction rule formalizes event-condition-action behavioral model, where the action is executed on event with a condition satisfied. Let us analyze the following example: *On customer book item request, if the item is available, then approve order and decrease amount of items in stock.* In this rule *on customer book item request* is a triggering rule event, *if the item is available* is a rule condition, *approve order* is a rule action and *decrease amount of items in stock* is a postcondition, which can be expressed as a logical constraint, requiring less amount of items in stock, than before the rule execution.

There are several general advantages of using reaction rules for specifying web services:

- Requirements are often captured in the form of rules in a natural language, formulated by business people;
- Reaction rules are easier to maintain and integrate with other kinds of rules, used in business applications (integrity rules, which specify constraints the data must fulfill, derivation rules, which explain how a model element can be derived);
- The topic of rules validation and verification is well studied;
- Reaction rules emphasis on events gives a flexible way to specify control flow and web service interaction patterns.

Our approach is on the edge of the MDE, Web Services, and rules. It brings several large communities together and helps to facilitate the use of rules and MDE technologies in the growing area of Web Services.

Business Rule Modeling

In this section we describe the concept of our approach talking about modeling of Business rules with the help of web services. In our approach, we look from the perspective of the potential patterns of the use of services. That is, we model services from the perspective of MEPs. It is important to point out that we **first start from the definition of a business rule that corresponds to a MEP under study**, but without considering the Web services that might be developed to support that rule. In this way, unlike other approaches to modeling of Web services, we are focused on the business rules describing how particular services are used, but without explicitly stating that we are talking about Web services. This approach enables us, not only to focus on the problems under study and the underlying business logic regulating the use of Web services, but also now we are able to translate such Web service modeling to both Web service languages and rule-based languages that can regulate how services are used at run-time.

Here, we describe how the in-out MEP is modeled in URML. The *in-out* MEP consists of exactly two messages: when a service receives an input message, it has to reply with an output message. This means that a fault message must be delivered to the same target node as the message it replaces, unless otherwise specified by an extension or binding extension. To show how this variation of the in-out MEP can be modeled by using reaction rules, we start our discussion from an example of the in-out MEP with an out-fault message. Let us consider the following business rule: *When a user sends a message, if the length of the message is less than five characters long return the message to the user, otherwise return an error message.*

The business rule is an in-out MEP and can be represented by two reaction rules, represented in semi-formal pseudo rule syntax:

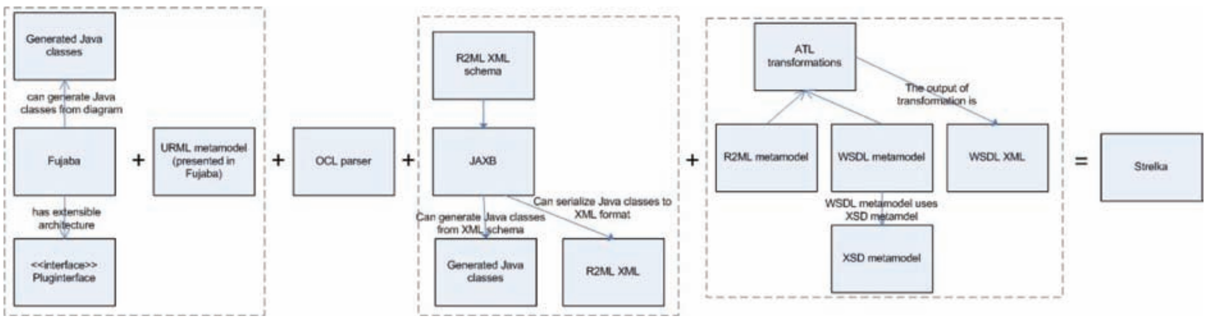
```
ON UserMessageArrival (echoIn-
Param)
IF echoInParam.length < 5 THEN
DO EchoStringResponse (echoIn-
Param)
ON UserMessageArrival (echoIn-
Param)
IF NOT echoInParam.length < 5
THEN
DO EchoStringError (“message too
long”)
```

In order to have these rules represented using our modeling notation, and also to be able to relate the rules with the elements of vocabularies, we model these rules by using URML. It is important to stress that all rules-based systems define business rules on top of business vocabularies. Thus, the URML graphical notation enables us to define business rules regulating the use of Web services by leveraging a human-comprehendible, and yet formally consistent representation, with the underlying business vocabularies (i.e., UML class models). These abovementioned two reaction rules represented in URML are shown in Figure 5.

In this particular case, we have the following situation: triggering event of either rule (i.e., UserMessageArrival) maps to the input message of the Web service operation. The action of the first reaction rule (i.e., EchoStringResponse), which is triggered when a condition is true, maps to the output message of the Web service operation. The action of the second reaction rule (i.e., EchoStringError), triggered on a false condition, maps to the out-fault message of the Web service operation.

As shown, a web service operation may be modeled by using more than one rule, where rules, that define an operation, are grouped in one package by the name of the operation. Otherwise, for modeling the basic in-out MEP, only one (first) reaction rule would be used. Besides the classes that model exchanged messages, the URML diagrams also

Figure 6. Strelka's functionality



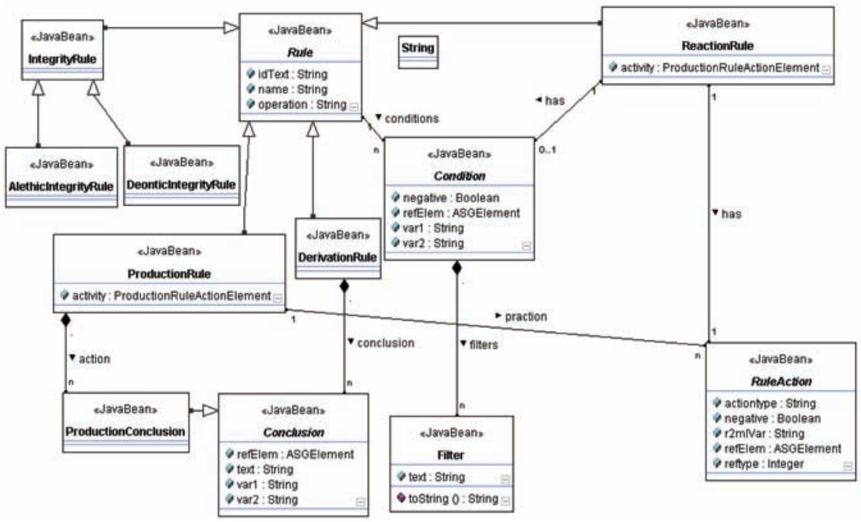
model conditions (e.g., echoInParam.length < 5). Such condition constructs are modeled by using OCL filters. OCL filters are based on a part of OCL that models logical expressions, which can be later translated to R2ML logical formulas, as parts of reaction rules. However, these OCL filters cannot be later translated to Web service descriptions (e.g., WSDL), as those languages cannot support such constructs. But, we can translate our URML models into rule-based languages (e.g., Jess or Drools). This means that for each Web service, we can generate a complementary rule, which fully regulates how the service is used. Another benefit of this approach is that our generated Web services, and their regulating rules, are consistent, as they originate (i.e. they are generated from) the same rule-based Web service model.

4. STRELKA: OUR MODELING ENVIRONMENT

In this section, we describe Strelka – a tool that was implemented as a plugin for the Fujaba tool[1]. Fujaba is a UML modeling tool, developed at the University of Paderborn in Germany. Fujaba allows creation of different types of UML diagrams. It allows generation of Java code based on created model, and it offers a plugin interface that we used for Strelka implementation. Figure 6 shows, symbolically, the architecture of Strelka, as well as the functionality that it provides.

The first part of Figure 6 (leftmost dashed box) shows us two important features Fujaba has: it is capable of generating Java code from diagrams, and it offers easily extensible architecture suitable for plugin implementation. This box also symbolically shows that one of the steps that need to be undertaken when creating a plugin for Fujaba is to create a metamodel for the new elements Fujaba will be extended with. After this first box, in the figure, we have an OCL parser. The OCL parser, which we used in this version of Strelka, is developed at the Technical University of Dresden in Germany[2]. This parser is widely accepted open source tool – besides Fujaba, it can be also integrated into other tools: ArgoUML, Borland Together, Poseidon. The second dashed box on Figure 6 shows that we use JAXB in the process of mapping URML to R2ML. JAXB guaranties that the R2ML XML document we get as the result of serialization of URML diagram created in Strelka, is valid according to the R2ML XML schema. The reason we use JAXB, is that Fujaba does not have explicitly defined metamodel in some of the metamodeling languages (e.g. MOF, or Ecore). So, we could not use model transformations (and ATL), in the mapping process between URML and R2ML like we did in the process of mapping between R2ML and WSDL (this mapping we describe in the next section). And the last dashed box on Figure 6 shows that Strelka has an integrated support for calling model transformations between R2ML and WSDL.

Figure 7. Strelka's rule metamodel



As already mentioned, one of the steps in the process of creating plug-ins for Fujaba is defining a metamodel for all the new elements that we extend Fujaba with. This metamodel is created in Fujaba, and then Java code is generated from it. This Java code we use afterwards, in the process of generating plugin itself. As the process of defining metamodel, and Java code generation is complete, we can create class diagrams that can incorporate those newly created elements, as well as UML elements that were already available (e.g. classes and associations). Figure 7 presents fragment of the Rule metamodel created in Fujaba. Classes that are relevant to the modeling of reaction rules are: ReactionRule, MessageAction, EventCondition, MessageEventType, and FaultMessageEventType. Figure 7 shows relations between the ReactionRule class to the other classes in this metamodel. The ReactionRule class is a subclass of Rule class (just like IntegrityRule, ProductionRule, and DerivationRule classes, which are used for modeling integrity, production, and derivation rules respectively).

5. TRANSFORMING WEB SERVICE MODELS

In this section, we summarize the translation process from URML reaction rules to Web services, and then we describe how this URML model can be transformed to Drools rule language.

The process of getting WSDL documents from URML models is as follows: first, we create an URML diagram in Strelka, after this we serialize it in the R2ML XML concrete syntax (this functionality is built in Strelka, and it relies on the use of JAXB). Once URML reaction rules are serialized to R2ML XML, they are translated into WDSL documents (this translation is done at the level of metamodels as described below). It is important to say that this process allows for two way transformation between URML and WSDL, that is, it is possible to perform reverse engineering of existing services, and extract business rules already integrated to service implementation.

To support this approach, we have implemented a number of transformations between different languages and their representations (these transformations can be found on the ATL ZOO page (WSDL2R2ML, 2006):

- *URML and R2ML XML concrete syntax*. As previously stated, this is the only transformation that is not implemented by using ATL (ATL, ver. 0.7), but instead it is based on JAXB
- *R2ML XML-based concrete syntax and R2ML metamodel*. This transformation is important to bridge concrete (XML) and abstract (MOF) syntax of R2ML. This is done by using ATL and by leveraging ATL's XML injector and extractor.
- *R2ML metamodel and WSDL metamodel*. This transformation is the core of our solution and presents mappings between R2ML and WSDL at the level of their abstract syntax.
- *WSDL XML-based concrete syntax and WSDL metamodel*. This transformation is important to bridge concrete (XML) and abstract (MOF) syntax of WSDL. This is done by using ATL and by leveraging ATL's XML injector and extractor.

Here, we describe the third and the forth transformation: i.e. transformation between R2ML metamodel and WSDL metamodel, and transformation between WSDL XML schema and WSDL metamodel.

Transformation Between WSDL Metamodel and R2ML Metamodel

Figure 8 shows the general principle of mappings between WSDL metamodel and R2ML metamodel (as can be seen on the figure, these mappings can be done in both directions).

Process of transforming WSDL metamodel elements to R2ML metamodel elements is conducted in the MOF technical space (TS) (Kurtev et al., 2002). Table 1 shows a conceptual mapping between these metamodels (this table does not show parts related to the mapping between the XML Schema language, which WSDL uses for defining message types and vocabularies, and the R2ML vocabulary).

Figure 9 shows an example of a matched rule from our (WSDL2R2ML.atl) transformation (for the Input class in the WSDL metamodel), and Figure 10 shows a graphical presentation of this mapping (for this particular class) between WSDL and R2ML metamodel. This matched rule initializes "sender", "objectVariable" and "type" features of the target R2ML MessageEventExpression element (we are putting a string "to_be_defined" in the "sender" feature, as the user will provide this information if required. This feature refers to the requester of the service).

As the result of the execution of WSDL2R2ML.atl transformation, we get an R2ML model. This model (which is an instance of the R2ML metamodel) can later be serialized into the R2ML XML concrete syntax, by our R2ML2XML.atl transformation and the use of the XMLExtractor class.

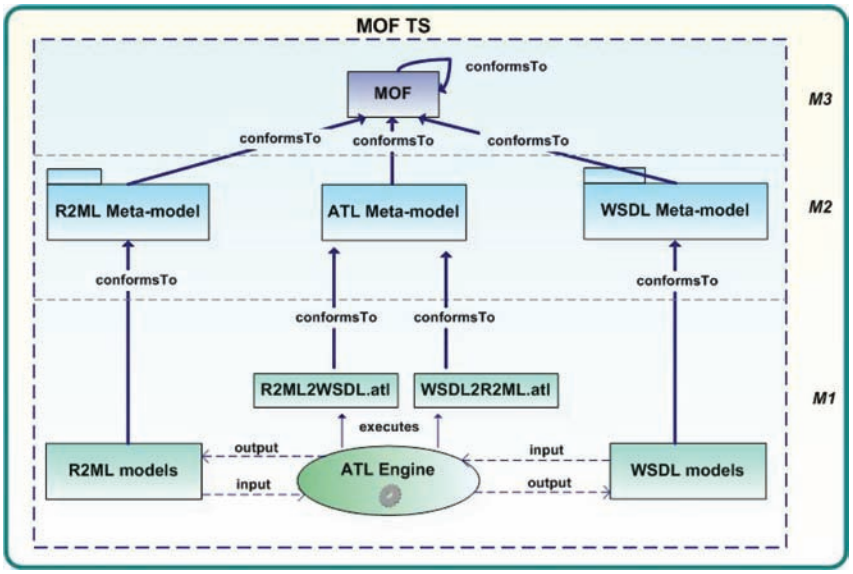
The WSDL2R2ML.atl transformation is defined as a sequence of ATL rules and ATL helpers. It has 11 matched rules, 3 lazy rules and 3 helpers.

Transformation, in the opposite direction, from R2ML metamodel elements to WSDL metamodel elements (R2ML2WSDL.atl) is based on the same conceptual mappings presented in the Table 1. This transformation uses 9 matched rules, 4 lazy rules, 1 unique lazy rule and 5 helpers. Figure 11 shows one matched rule from R2ML2WSDL transformation. This rule is mapping R2ML ReactionRuleSet element into WSDL Service and Interface elements.

Transformation of WSDL XML Schema to WSDL Metamodel

Because the concrete syntax of WSDL is defined in the XML technical space, and its abstract syntax (metamodel) is defined in the MOF TS, it is necessary to make a translation from concrete syntax

Figure 8. Mappings between models that conform to WSDL metamodel to models that conform to R2ML metamodel



elements to the abstract syntax elements. For this translation we have used XML metamodel (XML metamodel, 2006) as a bridge between XML and MOF TS.

To transform an XML document from XML TS to the XML model (instance of a XML metamodel) of the same rule in the MOF TS, we have used the XMLInjector class. By using a SAX parser, this class creates an XML model from an XML file. The next figure (Figure 12) shows general principle of mapping between WSDL XML concrete syntax and WSDL abstract syntax. Figure shows that the transformation can be done in both directions (i.e. we created bidirectional transformations).

Before we describe transformation process between WSDL XML schema and WSDL metamodel, we stress a fact that because WSDL does not define its own language for defining data types, we are referring to the XSD metamodel that can be obtained from the Eclipse Model Development Tool project[3].

Table 1. An excerpt of the mapping between the WSDL metamodel and the R2ML metamodel

WSDL metamodel	R2ML metamodel
Description	RuleBase
ElementType	Vocabulary
Interface	ReactionRuleSet
Operation	ReactionRule
Input	MessageEventExpression
Infault	MessageEventExpression
Output	MessageEventExpression
Outfault	MessageEventExpression

Figure 9. ATL matched rule that transforms WSDL Input element to R2ML MessageExpression element

```
rule TriggeringEvent {
  from i : WSDL!Input (
     i.oclIsTypeOf(WSDL!Input)
  )
  to o : R2ML!MessageEventExpression (
     sender <- 'to_be_defined',
     objectVariable <- thisModule.ObjectVariable(i.element),
     type <- i.element
  )
}
```

Transformation from WSDL XML schema to the WSDL metamodel has two primary steps:

- **Step 1:** This step is about transforming WSDL document from XML TS into MOF TS. In this step WSDL document (EchoService.xml on the Figure 12) is transformed into the WSDL metamodel that represents abstract syntax of the WSDL language in the MOF TS. By using XMLInjector class, WSDL XML document is transformed into model that conforms to the MOF based XML metamodel (step 1 on the Figure 12). As the result of this process we get XML model that can be presented in the XML XMI format (step 2 on the Figure 12).
- **Step 2:** This step is about transforming XML model (EchoService_XML on the Figure 12) into the WSDL based model (EchoService_WSDL on the Figure 12). ATL transformation XML2WSDL.atl (step 3 on the Figure 12) is performing this mapping. The output WSDL model (EchoService_WSDL) conforms to the WSDL metamodel. The XML2WSDL.atl transformation is executed on the model level (M1 level), by using information

Figure 10. Graphical presentation o mapping shown on the Figure 9

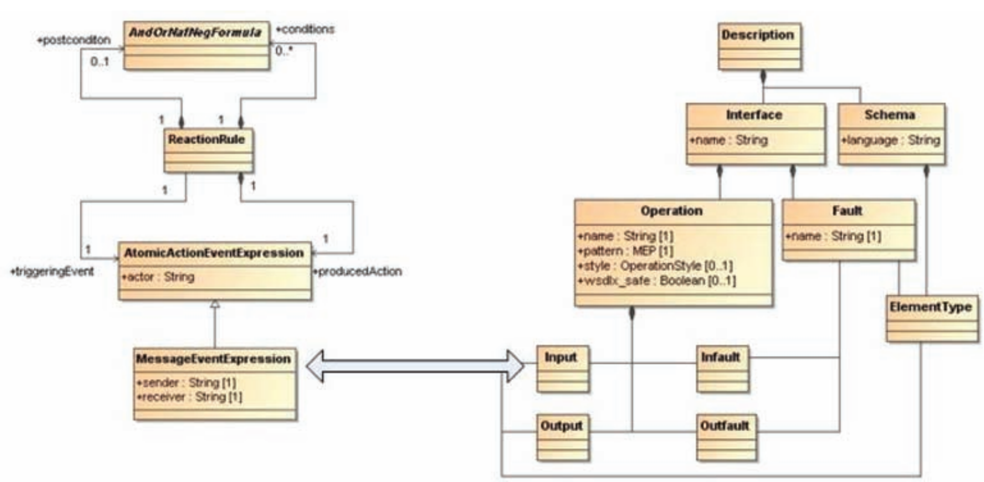


Figure 11. ATL matched rule that transforms R2ML ReactionRuleSet element to WSDL Interface element

```
rule Interface {
  from i : R2ML!ReactionRuleSet (
    i.oclIsTypeOf(R2ML!ReactionRuleSet)
  )

  to serv : WSDL!Service (
    name <- i.ruleSetID,
    interface <- inter,

    endpoint <- en
   ),
   inter : WSDL!Interface (
     name <- thisModule.setElemName(i.ruleSetID, 'Interface'),
     fault <- i.rules,
     operation <- let setOfGroupIDs : Set (String) = i.rules->iterate(p; s :
Set(String) = Set{} | s->including(p.groupID) ) in
     setOfGroupIDs->collect(e| thisModule.Operation(i.rules->asSequence()->
                    select(c| c.groupID = e)->first()) )
    ),
   en : WSDL!Endpoint (
     name <- thisModule.setElemName(i.ruleSetID, 'Endpoint'),
     address <- i.rules->select(c| c.oclIsTypeOf(R2ML!ReactionRule) and not

      c.producedAction.type.oclIsTypeOf(R2ML!FaultMessageType))-> first().pro-
                                ducedAction.sender
   )
}
```

about elements from the (XML and WSDL) metamodel level (M2 level).

The next table (Table 2) shows fraction of mappings between WSDL XML schema and WSDL metamodel.

Figure 13 shows a matched rule from XML2WSDL.atl transformation for the "description" element of the XML metamodel. ATL rule Description transforms Element of the XML model (EchoService_XML) to Description (i.e. WSDL!Description) of the WSDL metamodel.

Figure 12. Transformation scenario: WSDL XML format into WSDL metamodel

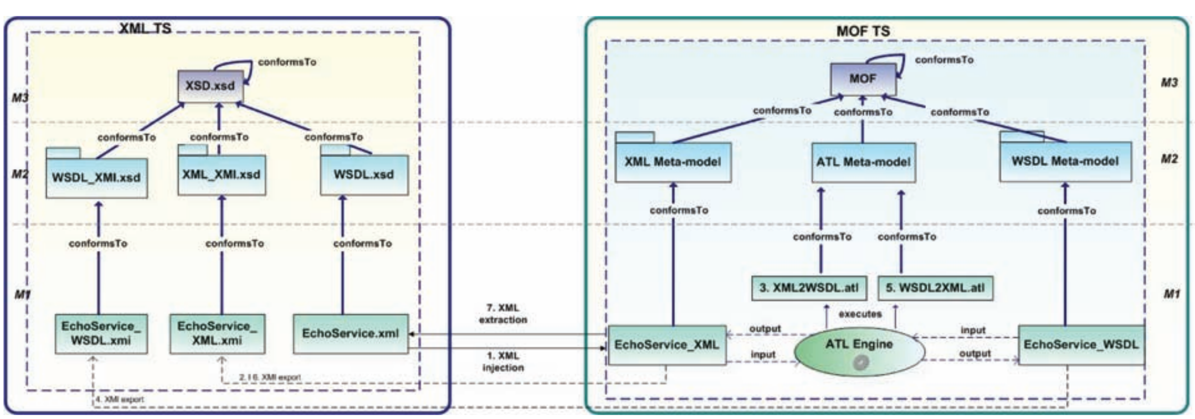


Table 2. Fraction of transformation between WSDL XML schema and WSDL metamodel

WSDL XML schema	XML metamodel	WSDL metamodel	Description
description	Root name='description'	Description	The Root element that contains types, binding, service and interface elements
interface	Element name='interface'	Interface	This element presents WSDL interface, it contains operation and fault elements
operation	Element name='operation'	Operation	This element presents WSDL operation, it contains in/outfault and in/output elements

This rule initializes "types", "interface", "binding" and "service" attributes of Description target model element by calling an ATL helper shown on the Figure 14.

The XML2WSDL transformation has 18 matched rules, one lazy rule, and 17 helpers (we have this number of helpers because we use helpers for mapping enumeration elements to their string representation – e.g. in-out MEP into the 'http://www.w3.org/2006/01/wsdl/in-out' string). The biggest problem in the XML2WSDL transformation was mapping of XML schema complex types. For solving this problem we found the (EclipseDoc, 2002) document very helpful.

Transformation of the WSDL Metamodel into the WSDL XML Schema

Transformation process from WSDL abstract syntax (i.e. WSDL metamodel) into the WSDL XML concrete syntax also contains two primary steps.

- **Step 1:** Transformation of WSDL model into XML model. Here a WSDL model (EchoService_WSDL on Figure 12) is transformed into XML model (EchoService_XML) by using the ATL transformation WSDL2XML.atl (step 5 on the Figure 12). After this transformation is executed on the input WSDL models, XML

Figure 13. Description matched rule

```
rule Description {
from
  i : XML!Element (
    i.name = 'description'
  )
to
  o : WSDL!Description (
    types <- i.getElementsByName('types'),
    interface <- i.getElementsByName('interface'),
    binding <- i.getElementsByName('binding'),
    service <- i.getElementsByName('service')
  )
}
```

Figure 14. Helpers we use in the rule in Figure 13

```
helper context XML!Element def: getAttr(n : String) : Sequence(XML!Attribute) =
  self.getChildren(n)->select(e | e.oclIsKindOf(XML!Attribute))->first();

helper context XML!Element def: getAttrVal(n : String) : String =
  let d : XML!Attribute = self.getAttr(n) in
  if d.oclIsUndefined() then
     OclUndefined
  else
     d.value
  endif;
```

models (EchoService_XML) are placed in the model repository (EchoService_XML.xmi on the Figure 12).

- **Step 2:** Transformation from MOF TS into the XML TS, where an XML model (EchoService_XML on the Figure 12) that conforms to the MOF based XML metamodel (and that is created in the step 1), is transformed into the EchoService.xml document (step 7 on the Figure 12).

In the WSDL2XML transformation, we have 17 matched rules, 3 lazy rules, as well as 5 helpers. The WSDL2XML transformation (i.e. transformation from WSDL metamodel into the WSDL XML schema) is simpler then XML2WSDL transformation (WSDL elements are directly mapped to their corresponding representation from the WSDL XML schema, and the other way around is also true).

Deploying URML Models on a Rule-Based Engine

Here, we describe the transformation process from a URML model to the Drools rule engine. For this purpose we are using an example of the in-out MEP, presented with the following business rule:

When a user sends a request for Social Insurance Number (SIN), if the authentication is successful send the SIN to the user.

This business rule is presented with one reaction rule that has a triggering event SINrequest, condition Person.Address=SINrequest.Address AND Person.Name=SINrequest.Name, and action SINresponse:

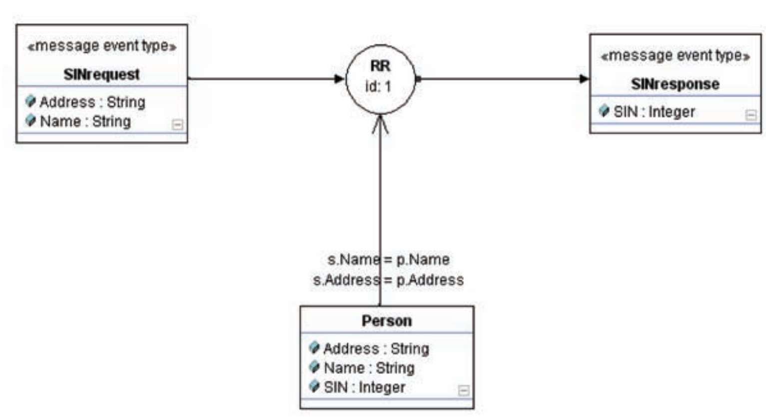
```
ON SINrequest (Address, Name)
IF Person.Address=SINrequest.
Address AND Person.
Name=SINrequest.Name
DO SINresponse(SIN)
```

An URML diagram that models the corresponding Web service operation is presented in Figure 15, and Figure 16 illustrates the R2ML presentation of it that Strelka generated for us.

The first step in getting Drools rules, from the generated R2ML file is transformation of R2ML reaction rules to R2ML production rules. The second step is transformation of R2ML production rules to Drools rules, where we are using the translation provided by (Giurca & Werner, 2006).

In order to successfully realize the first step one must address the major difference between reaction rules and production rules: reaction rules

Figure 15. URML model of the in-out MEP



have triggeringEvent element, while production rules do not have this element. So there is a need for representing triggeringEvent in production rules. If only atomic events are considered, it is legible to map the triggeringEvent element to the condition element in production rule. Figure 17 shows the corresponding R2ML production rule document that we get as the result of our transformation.

As Figure 17 shows, the producedAction part

Figure 16. R2ML serialization of URML model from Figure 15

```
    <r2ml:ReactionRuleSet r2ml:ruleSetID="ClassDiagram1">
        <r2ml:ReactionRule r2ml:ruleID="id469" r2ml:groupID="operation">
            <r2ml:triggeringEvent>
                <r2ml:MessageEventExpression r2ml:sender="not_defined" r2ml:eventType="SINrequest">
                    <r2ml:ObjectVariable r2ml:name="s" r2ml:classID="SINrequest"/>
                </r2ml:MessageEventExpression>
            </r2ml:triggeringEvent>
            <r2ml:conditions>
                <r2ml:ObjectClassificationAtom r2ml:classID="Person" r2ml:isNegated="false">
                    <r2ml:ObjectVariable r2ml:name="p"/>
                </r2ml:ObjectClassificationAtom>
                <r2ml:DatatypePredicateAtom r2ml:datatypePredicateID="swrlb:equal" r2ml:isNegated="false">
                    <r2ml:dataArguments>
                        <r2ml:AttributeFunctionTerm r2ml:attributeID="Name">
                            <r2ml:contextArgument>
<r2ml:ObjectVariable r2ml:name="p" r2ml:classID="Person"/>
                            </r2ml:contextArgument>
                        </r2ml:AttributeFunctionTerm>
                        <r2ml:AttributeFunctionTerm r2ml:attributeID="Name">
                            <r2ml:contextArgument>
<r2ml:ObjectVariable r2ml:name="s" r2ml:classID="SINrequest"/>
                            </r2ml:contextArgument>
                        </r2ml:AttributeFunctionTerm>
                    </r2ml:dataArguments>
                </r2ml:DatatypePredicateAtom>
 <r2ml:DatatypePredicateAtom r2ml:datatypePredicateID="swrlb:equal" r2ml:isNegated="false">
                    <r2ml:dataArguments>
                        <r2ml:AttributeFunctionTerm r2ml:attributeID="Address">
                            <r2ml:contextArgument>
<r2ml:ObjectVariable r2ml:name="p" r2ml:classID="Person"/>
                            </r2ml:contextArgument>
                        </r2ml:AttributeFunctionTerm>
                        <r2ml:AttributeFunctionTerm r2ml:attributeID="Address">
                            <r2ml:contextArgument>
<r2ml:ObjectVariable r2ml:name="s" r2ml:classID="SINrequest"/>
                            </r2ml:contextArgument>
                        </r2ml:AttributeFunctionTerm>
                    </r2ml:dataArguments>
                </r2ml:DatatypePredicateAtom>
            </r2ml:conditions>
            <r2ml:producedAction>
                <r2ml:MessageEventExpression r2ml:eventType="SINresponse"/>
               <r2ml:ObjectVariable r2ml:name="SIN" r2ml:classID="SINresponse"/>
            </r2ml:producedAction>
        </r2ml:ReactionRule>
    </r2ml:ReactionRuleSet>
</r2ml:RuleBase>
```

Figure 17. Generated R2ML production rule document – part of rule containing conditions is shown on left – vocabulary part can be seen on the right

of the generated production rule is the same as in the reaction rule, while the conditions part is constituted from conditions of a reaction rule along with the triggeringEvent of reaction rule. Also, a vocabulary part of production rule is updated with the adequate class generated from triggeringEvent of reaction rule. (i.e. if no such class is already defined. This situation happens with the presentation of rules in older versions of R2ML). Table 3 provides a brief overview of the transformation from Reaction Rules to Production Rules.

Once we have generated a R2ML production rule from the R2ML reaction rule, we use the transformation provided by (Giurca & Werner, 2006) to transform R2ML production rule to a Drools rule. Figure 18 shows the generated Drools rule from the production rule presented in Figure 17. The left part of this figure shows the classes generated for Drools engine, and the right part of the figure shows the rule itself. Each rule fires an action when conditions are met - in the case below, SINrequest trigger this rule, while SINresponse is fired as the response to the information requested.

Table 4 presents the concept of transformation between R2ML production rule elements and the Drools generated code. The table is not presenting all steps of transformation as some of them are not a simple mapping between elements but gives a general idea of what is done.

6. RELATED WORK

In this section, we compare the proposed approach with some relevant solutions to the modeling of Web services that are based on the MDE principles.

(Bezivin et al., 2004) demonstrate how one can take advantage of MDE to develop e-business applications. In their approach, they start from UML and the UML profile for Enterprise Distributed Object Computing (EDOC) to define platform-independent models of e-business applications.

Table 3. Presenting the basic of transformation from reastion rule to production rule

Reaction Rule	Production Rule
ReactionRuleSet	ProductionRuleSet
ReactionRule	ProductionRule
TrigerringEvent & Conditions	Conditions (There will be only one Conditions element generated although both TrigerringEvent and Conditions could be present in the Reaction Rule)
MessageEventExpression	ObjectClassificationAtom

In the next step, they translate such models into models that are based on metamodels of Java, WSDL, and JWSDP. Although this approach uses ATL, similar to what we do, it does not provide two way transformations between models and service implementation. This is unlike our solution, because we support two-way transformations, and thus we enable reverse engineering of the existing Web services. Next, the translation of regular UML models, i.e., UML class related elements into WSDL is limited, as one does not have enough expressivity in UML diagrams to define all details of WSDL (i.e., one can not distinguish between input and output messages). This issue is addressed by using the UML profile for EDOC, but this approach is more focused on modeling distributed components rather on modeling business logics. In our approach, the use of reaction rules in URML models enables us to be closer to business processes, while OCL filter expressions can even further specify conditions under which a message can happen. This is very useful for potential integration of business process based on the use of Web services (Milanovic et al., 2006). Of course, such filter expressions overcome the potentials of the Web service technology, as they presume that there should also be a rule-based engine, which is able to interpret such conditions. However, the research on semantic Web services (Sheth et al., 2006) and Web rules (Nagl et al., 2006) demonstrates that this issue attracts more attention in the Web service research community.

(Vara et al., 2005) define a metamodel for WSDL and its corresponding UML profile that is used for modeling of Web services in the MIDAS-CASE MDA-based tool for Web information system development. They also support automatic generation of the respective WSDL description for

Figure 18. Generated Drools rule (right) and generated class of it (left)

```
public class Person {
    public string Address;
    public string Name;
    public integer SIN;
 }
public class SINrequest {
    public string Address;
    public string Name;
 }
public class SINresponse {
    public integer SIN;
 }
```

```
rule "aRule"

  when
    #conditions
        p : Person()
        s : SINrequest()

        p : Person ( namePerson : Name )
        p : Person ( addressPerson : Address )
        s : SINrequest ( nameSINrequest : Name )
        s : SINrequest ( addressSINrequest : Address )

        eval (nameSINrequest.compareTo( namePerson ) == 0 )
        eval (addressSINrequest.compareTo( addressPerson ) == 0 )

  then
    #actions
        SIN : SINresponse()

end
```

Table 4. Presenting transformation concept of R2ML production rule to Drools

R2ML	Drools
Class	Class
Attribute	Integer, String etc (Based on the value of r2mlv:ID)
Conditions	Conditions (and based on the object within this element the condition element will be generated)
producedAction	Actions (based on the elements residing within this element proper code with regards to the previously generated classes will be generated)

the Web services that is being modeled. Although the MIDAS framework supports platform-independent models of services and service compositions, their definitions are very incomplete and one can hardly generate complete service models automatically. In our approach, we do not strictly base models of services on workflows in which services will be used, as we wanted to focus on how services are used (i.e., MEPs).

(Gronmo et al., 2005) and (Timm & Gannod, 2005) propose an approach to modeling semantic Web services by using UML profiles. Gronmo et al. have the goal to define platform-independent models of services, which can be translated to semantic Web services (OWL-S and WSMO). In their work, they abstract concepts from OWL-S and WSMO, and extend UML activity diagrams (i.e., they define a UML profile) accordingly, while for defining vocabularies they use the Ontology Definition Metamodel, and its corresponding UML profile defined in (Gasevic et al., 2005). Since both these approaches use XSLT, they can hardly translate service pre- and post-conditions from their UML definitions (i.e., OCL) to the languages of the potential target platforms (e.g. Jess). This approach does not consider modeling usage patterns or MEPs or error handling (i.e., business logic) like we do, but instead it focuses on the details specific for service platforms.

(Manolescu, et al., 2005) propose a model-driven approach to designing and deploying service-enabled web applications. This work is different from the ones above in the following aspects: i) it extends WebML, a well-known modeling language for developing Web applications; ii) it uses E-R models for defining data types; and iii) it focuses on the usage patters (MEPs) of Web services in order to define WebML extensions. Their modeling approach is based on the use of MEPs, but they are considering MEPs used in WSDL 1.1, and for each MEP they define corresponding WebML primitive. These new WebML primitives can be used in hypertext models of Web applications. This approach is the most similar to ours; both approaches fully focus on modeling business processes and potential usage patterns. However, the WebML approach does not explicitly define a Web service

Table 5. Summary of solutions for modeling of Web services based on the MDE principles

Approach	Used metamodel	Model description	transformation	output
(Timm & Gannod, 2005)	UML profile	UML XMI	XSLT	OWL-S
(Bezivin et al., 2004)	MOF	MOF XMI	ATL	WSDL, Java, JWSDP
(Manolescu, et al., 2005)	Extended WebML	WebML model	XSLT (through visual editor)	WebRatio output
(Vara et al., 2005)	UML profile	UML XMI	MIDAS-CASE transformation unit	MIDAS output
Our approach	MOF	MOF XMI	ATL	R2ML, WSDL

description, but it infers it from the definitions of hypertext models in which the service is used, which makes this process context dependent. That is, services are defined inside specific WebML workflows unlike our approach where we define workflow-independent services. In addition, they do not consider the use of preconditions of Web services, which is important for the reusability of services in different contexts. Although WebML has support for exception handling (Brambilla et al., 2005), this is also focused on the workflow level rather than on a service level. Finally, the WebML approach does not support reverse engineering of Web services.

The following table (Table 5) summarizes these approaches of modeling of Web services, and shows their relation to our approach we describe in this paper.

7. CONCLUSION

In this paper, we have demonstrated how the use of MDE principles can enable for rule-based modeling of Web services. By using the MDE principles, we have been able to develop a framework for modeling Web services from the perspective of how services are used in terms of message-exchange patterns (MEPs). Our approach enables developers to focus on the definition of business rules, which regulate MEPs, instead of focusing on low level Web service details or on contexts where services are used (i.e., workflows). Business rules are important assets of a business organization that embody valuable domain knowledge. They are stated in a way, which is very close to how users think and talk. Due to the declarative nature of rules, the business logic can easier be updated without the need to change to whole system. Furthermore, integrating business rules in process-oriented Web service composition can improve the quality and ease development. Once business rules are expressed explicitly as first-class entities, they can be reused across several compositions i.e., the same rule can be applied to many web service compositions.

By leveraging MDE, we have defined a rule-based modeling language that can be managed by universal MDE tools. The use of model transformations allows for transforming platform independent models of business logic to specific platforms such as Web services. Moreover, for each Web service, we can also generate a rule that will fully regulate the behavior of the service (i.e., how the service is used), and thus make sure that the business logic is fully followed. In this paper, we have not explained that part of the solution, but we are going to report on that in our future papers.

Our solution has much broader potentials that overcome the pure translation between rule-based business models and Web services. Unlike the WSDL definition of Web services, our models also have an option for defining pre- and post-conditions of services. However, WSDL can not express pre- and post-conditions in the descriptions of Web services. That is, Web service tools can not automatically support conditions under which some services can be used (as defined by OCL filters in Section 4). To address this problem, we expand our approach on W3C's Semantic Annotations for WSDL (SAWSDL) recommendation. In our current activities, we have developed transformations from our R2ML reaction rule-based models to several production rule languages (e.g., Drools, Jena2, and Jess). Our particular focus is on Drools, as the use of Drools allows us to directly enforce business rules to regulate the use of Web services deployed on JBoss' application server.

REFERENCES

Arroyo, S., Toma, I., Roman, D., Drumm, C., Dimitrov, M., Spork, M., et al. (2004). *Report on state of the art and requirements analysis (Deliverable 3.1)*. DIP EU FP6 Integrated Project. ATLAS group, Nantes. (ver. 0.7) (n.d.). *Atlas transformation language - User manual*. Retrieved from http://www.eclipse.org/gmt/atl/doc/ATL_User_Manual[v0.7].pdf

Bézivin, J. (2005). On the unification power of models. *Software and System Modeling, 4*(2), 171–188. doi:10.1007/s10270-005-0079-0

Bezivin, J., Hammoudi, S., Lopes, D., & Jouault, F. (2004). Applying MDA approach for Web Service platform. In *Proc. of the 8th IEEE Int. Conf. on Enterprise Distributed Object Computing Conf.* (pp. 58–70).

Brahe, S., & Osterbye, K. (2006). Business process modeling: Defining domain specific modeling languages by use of UML profiles. In *ECMDA-FA 2006* (LNCS 4066, pp. 241-255). 2006.

Brambilla, M., Ceri, S., Comai, S., & Tziviskou, C. (2005). Exception handling in workflow-driven Web applications. In *Proc. of the 14th International WWW Conference* (pp. 170–179).

Charfi, A., & Mezini, M. (2004). Hybrid Web service composition: Business processes meet business rules. In *Proc. of the 2nd International Conference on Service Oriented Computing* (pp. 30-38).

Chinnici, R., Moreau, J., Ryman, A., & Weerawarana, S. (2007). Web Services Description Language (WSDL) Ver. 2.0 Part 1: Core Language. *W3C Recommendation*. Retrieved from http://www.w3.org/TR/2007/REC-wsdl20-20070626.

Clement, L., Hately, A., Riegen, C., & Rogers, T. (2004). *UDDI Ver. 3.0.2*. Retrieved from http://www.oasis-open.org/committees/uddi-spec/doc/spec/v3/uddi-v3.0.2-20041019.htm.

Coyle, F. (2002). *XML, Web services, and the data revolution*. Addison Wesley.

Czarnecki, K., & Helsen, S. (2006). Feature-based survey of model transformation approaches. *IBM Systems Journal*, *45*(3), 621–645.

EclipseDoc. (2002). *How abstract components are created when reading a schema – an animation in 9 slides (complexType)*. Retrieved from http://help.eclipse.org/help33/index.jsp?topic=/org.eclipse.xsd.doc/references/diagrams/images/xsd-diagram-ct02/img0.html

Erl, T. (2005). *Service-oriented architecture: Concepts, technology, and design*. Prentice Hall.

Falkovych, K., Sabou, M., & Stuckenschmidt, H. (2003). UML for the Semantic Web: Transformation-based approaches. In B. Omelayenko & M. Klein (Eds.), *Knowledge transformation for the Semantic Web* (pp. 92-106). Amsterdam: IOS Press.

Gasevic, D., Djuric, D., & Devedzic, V. (2005). Bridging MDA and OWL ontologies. *Journal of Web Engineering*, *4*(2), 119–134.

Ginsberg, A., Hirtle, D., McCabe, F., & Patranjan, P. (2006). *RIF Use cases and requirements*. W3C Working Draft. Retrieved from http://www.w3.org/TR/rif-ucr.

Giurca, A., Lukichev, S., & Wagner, G. (2006). Modeling Web services with URML. In *Proc. of Workshop Semantics for Business Process Management.*

Giurca, A., & Werner, J. (2006). *Translator: R2ML to JBoss rules*. Retrieved from http://oxygen.informatik.tu-cottbus.de/translator/R2MLtoJBossRules/

Gronmo, R., Jaeger, M. C., & Hoff, H. (2005). Transformations between UML and OWL-S. In *Proc. of the 1st European Conf. on Model Driven Architecture - Foundations and Applications* (pp. 269–283).

Gudgin, M., Hadley, M., Mendelsohn, N., Moreau, J., Nielsen, H. F., Karmarkar, A., & Lafon, Y. (2007). *SOAP Ver. 1.2 Part 1: Messaging framework*. W3C Recommendation. Retrieved from http://www.w3.org/TR/soap12-part1/

Hirtle, D., Boley, H., Grosof, B., Kifer, M., Sintek, M., Tabet, S., & Wagner, G. (2006). *Schema specification of RuleML 0.91*. Retrieved from http://www.ruleml.org/spec/

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004). *SWRL: A Semantic Web rule language combining OWL and RuleML*. W3C Member Submission. Retrieved from http://www.w3.org/Submission/SWRL/

Jovanović, J., & Gašević, D. (2005). XML/XSLT-based knowledge sharing. *Expert Systems with Applications*, *29*(3), 535–553. doi:10.1016/j.eswa.2005.04.024

Kurtev, I., Bezivin, J., & Aksit, M. (2002). Technological spaces: An initial appraisal. In *Proceedings of the Confed. Int'l Conferences CoopIS, DOA, and ODBASE*.

Lewis, A. (2007). Web *Services Description Language (WSDL) Version 2.0: Additional MEPs*. W3C Working Group Note. Retrieved from http://www.w3.org/TR/2007/NOTE-wsdl20-additional-meps-20070626/

Lukichev, S., & Wagner, G. (2006). UML-based rule modeling with Fujaba. In *Proc. of the 4th International Fujaba Days* (pp. 31-35).

Lukichev, S., & Wagner, G. (2006a). Visual rules modeling. In *Proceedings of the 6th Inter-national Conference Perspectives of Systems Informatics* (pp. 467-673).

Manolescu, I., Brambilla, M., Ceri, S., Comai, S., & Fraternali, P. (2005). Model-driven design and deployment of service-enabled Web applications. *ACM Transactions on Internet Technology*, *5*(3), 439–479. doi:10.1145/1084772.1084773

Milanovic, M., Gasevic, D., Giurca, A., Wagner, G., & Devedzic, V. (2006). On interchanging between OWL/SWRL and UML/OCL. In *Proc. of the OCLApps Workshop* (pp. 81-95).

Miller, J., & Mukerji, J. (Eds.). (2003). *MDA Guide Version 1.0.1*. OMG.

MOF. (2005). Object Management Group, Meta Object Facility (MOF) Core, v2.0. *OMG Document formal/06-01-01*. Retrieved from http://www.omg.org/cgi-bin/doc?formal/2006-01-01.

Nagl, C., Rosenberg, F., & Dustdar, S. (2006). VIDRE– A distributed service oriented business rule engine based on RuleML. In *Proc. of the 10th IEEE International Enterprise Distributed Object Computing Conference* (pp. 35-44).

Object Management Group. (2008). *Business Process Model and Notation (BPMN) specification 2.0*. Retrieved from http://www.omg.org/cgi-bin/doc?bmi/08-02-06

OCL (2006). Object constraint language. *OMG Available Specification, Version 2.0, formal/06-05-01*.

Proctor, M., Neale, M., Frandsen, M., Griffith, S., Jr., Tirelli, E., Meyer, F., & Verlaenen, K. (n.d.). *Drools documentation*. Retrieved from http://downloads.jboss.com/drools/docs/4.0.7.19894.GA/html/index.html

Ribaric, M., Gasevic, D., & Milanovic, M. (n.d.). *R2ML: A rule-based approach to modeling of semantically-enriched Web services.*

Rosenberg, F., & Dustdar, S. (2005). Business Rules Integration in BPEL – A service-oriented approach. In *Proc. of the 7th International IEEE Conference on E-Commerce Technology* (pp. 476-479).

Rupp, N. A. (n.d.). *The login of the bottom line: An introduction to The Drools Project*. Retrieved from http://www.theserverside.com/tt/articles/article.tss?l=Drools

SBVR. (2006). Object Management Group. *Semantics of business vocabulary and business rules (SBVR) – Interim specification.* Retrieved from http://www.omg.org/docs/bei/05-08-01.pdf

Schneier,B.(1992). The Rete matching algorithm: A description of the pattern-matching algorithm underlying Drools.

Seidewitz, E. (2003). What models mean. *IEEE Software*, *20*(5), 26–32. doi:10.1109/MS.2003.1231147

Sheth, A., Verma, K., & Gomadam, K. (2006). Semantics to energize the full services spectrum. *Communications of the ACM*, *49*, 55–61. doi:10.1145/1139922.1139949

Tabet, S., Wagner, G., Spreeuwenberg, S., Vincent, P., Jacques, G., de Sainte Marie, C., et al. (2007). *Object Management Group: Production Rule Representation (PRR)*. Retrieved from http://www.w3.org/2004/12/rules-ws/paper/53/

Timm, J., & Gannod, G. (2005). A model-driven approach for specifying Semantic Web services. In *Proc. of IEEE International Conference on Web Services* (pp. 313-320).

UML (2005). Object Management Group, Unified Modeling Language 2.0. Docs. *formal/05-07-04 & formal/05-07-05*.

Vara, J., Castro, V., & Marcos, E. (2005). WSDL Automatic Generation from UML Models in a MDA Framework. *International Journal of Web Services Practices*, *1*(1-2), 1–12.

Wagner, G., Giurca, A., & Lukichev, S. (2006). R2ML: A general approach for marking up rules. In *Dagstuhl Seminar Proc. 05371*.

WSDL2R2ML (2006). *The WSDL to R2ML transformation scenario*. Retrieved from http://www.eclipse.org/m2m/atl/atlTransformations/#WSDL2R2ML

XML metamodel (2006). Retrieved from http://mdr.netbeans.org/mdrxml.html

KEY TERMS AND DEFINITIONS

Business Rules: Business rules or business rule sets describe the operations, definitions and constraints that apply to an organization in achieving its goals.

Drools: A business rule management system (BRMS) with a forward chaining inference based rules engine, more correctly known as a production rule system, using an enhanced implementation of the Rete algorithm.

Reaction Rules: Also known as Event-Condition-Action Rules, are three components rules i.e. the *event* part specifies the signal that triggers the invocation of the rule, the *condition* part is a logical test that, if satisfied or evaluates to true, causes the action to be carried out and the *action* part consists of updates or invocations on the local data.

R2ML: REWERSE I1 Rule Markup Language, a rule markup language allowing rules interoperability.

URML: UML-based Modeling Language, a visual language for rules modeling.

Web Service: A software system designed to support interoperable machine-to-machine interaction over a network.

WSDL: Web Service Description Language is a language for describing both the abstract functionality of a service and the concrete details of a Web Service.

ENDNOTES

1 http://www.Fujaba.de

2 http://dresden-ocl.sourceforge.net/

3 http://www.eclipse.org/xsd/s

This work was previously published in Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches, edited by A. Giurca, D. Gasevic, and K. Taveter, pp. 422-446, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 2.14

A Semantic Web-Based Approach for Building Personalized News Services

Flavius Frasincar
Erasmus University Rotterdam, The Netherlands

Jethro Borsje
Erasmus University Rotterdam, The Netherlands

Leonard Levering
Erasmus University Rotterdam, The Netherlands

ABSTRACT

This article proposes Hermes, a Semantic Web-based framework for building personalized news services. It makes use of ontologies for knowledge representation, natural language processing techniques for semantic text analysis, and semantic query languages for specifying wanted information. Hermes is supported by an implementation of the framework, the Hermes News Portal, a tool which allows users to have a personalized online access to news items. The Hermes framework and its associated implementation aim at advancing the state-of-the-art of semantic approaches for personalized news services by employing Semantic Web standards, exploiting domain information, using a word sense disambiguation procedure, and being able to express temporal constraints for the desired news items.

INTRODUCTION

The simplicity, availability, reachability, and reduced exploitation costs have made the Web one of the most common platforms for information publishing and dissemination. This is particularly true for news agencies that use Web technologies to present emerging news regarding different types of events as for example business, cultural, sport, and weather events. Most of this information is published as unstructured text that is made available to a general audience by means of Web pages.

The heterogeneity of the Web audience and the diversity of the published information asks for more refined ways of delivering information that would enable users to access news items that interest them. For this purpose the Really Simple Syndication (RSS) (Winer, 2003) standard was developed that publishes information in a semi-structured format that supports machine processing. This format is based on metadata that (1) associates news items with channels (feeds) that have properties like categories (e.g., business, sport, politics, etc.), title, publication date, etc., and (2) describes news items by means of their properties as categories (e.g., online business, business system, Internet marketing, etc.), release time, title, abstract, link to the original published information, etc.

Most of the annotations supported by the RSS feeds are coarse-grained in the sense that they fail to identify the different topics addressed in a certain news item. Also, the current annotations are only partially processable by machines as the tags do not have unique semantic meaning associated to them and thus have different interpretations. Being able to understand the semantic content of a news item would enable a fine-grained categorization of this information, thus better supporting the users (casual users, media analysts, stock brokers, etc.) information needs.

In order to make the Web data not only machine readable but also machine understandable the World Wide Web Consortium proposes the Semantic Web (Berners-Lee, Hendler, & Lassila, 2001), a sequence of technologies that allow for self-describing content. On the Semantic Web metadata is defined using semantic information usually captured in ontologies. Some of the most popular formats to describe ontologies on the Semantic Web are RDF(S) (Klyne & Carroll, 2004) (Brickley & Guha, 2004) and OWL (Bechhofer et al., 2004).

A special class of users who make daily use of (emerging) news is that of stock brokers. Because news messages may have a strong impact on stock prices, stock brokers need to monitor these messages carefully. Due to the large amounts of news information published on a daily basis, the manual task of retrieving the most interesting news items with respect to a given portfolio is a challenging one. Existing approaches such as Google Finance or Yahoo! Finance are developed to meet these personalization needs by supporting automatic news filtering on the Web.

Current approaches to news filtering are able to retrieve only the news that explicitly mention the companies involved, failing to deliver indirect information which is also deemed relevant for the considered portfolio. For example, for a portfolio based on Google shares, such systems fail to deliver news items related to competitors of Google, such as Yahoo! or Microsoft, which might have an indirect influence on the share price of Google. Exploiting the semantic contextual information related to companies such as its competitors, CEO's, alliances, products, etc., enables a more comprehensive overview of relevant news with respect to a certain portfolio.

Another limitation of existing news filtering systems is their inability to cope with delivering news items satisfying temporal constraints. The time aspect is of utmost importance when, for example, one considers the fact that news items usually have an immediate impact on stock prices, or when one desires to do a historical analysis of past news and stock price evolutions. Being able to exploit the timestamps associated to news items enables retrieving only news that obey user-determined time-related constraints.

In this article we propose the Hermes framework, a semantic-based approach for retrieving news items related, directly or indirectly, to the concepts of interests from a domain ontology. In addition these news items might need to satisfy temporal constraints. For illustration purposes we focus here on the NASDAQ stock market domain (Kandel & Marx, 1997), but the genericity of our approach makes it applicable also to other domains, as, e.g., tourism or scientific domains. The Hermes News Portal (HNP) is an implementation of the Hermes framework, which allows the user to specify queries for the concepts of interest and temporal constraints, and retrieve the corresponding news items.

For HNP we make use of Semantic Web technologies like OWL (Bechhofer et al., 2004) for formally defining the semantics of the concepts of interest in the ontology. We employ natural language processing (NLP) technologies as, e.g., lexical analysis, gazetteering, word sense disambiguation, etc., for indexing news items based on ontology terms. The most popular Semantic Web query language SPARQL (Prud'hommeaux & Seaborne, 2008) is used for expressing queries using the previously identified concepts. In order to simplify the representation of temporal constraints we propose time-related extensions to SPARQL. HNP is a generic platform that could easily be applied to other domains than the financial one.

The structure of the article is defined as follows. The first section discusses related approaches for personalized news services. The second section presents the Hermes framework identifying the proposed methodological steps. The third section describes HNP, an implementation of the proposed framework. The last section concludes the article and discusses future work.

RELATED WORK

Among the methods that aim at personalizing news information we distinguish two types: non-semantic approaches and semantic approaches. In the followings we will present short descriptions of two non-semantic methods: Server for Adaptive News and YourNews, and two semantic methods: MyPlanet and SemNews. For each presented method we give the differences compared to our approach and at the end of this section we highlight the main contributions of the Hermes framework.

Server for Adaptive News (SeAN) (Ardissono, Console, & Torre, 2001) enables a personalized access to news servers on the Web. The generated views are composed of sections, as in newspapers, on which customized news items are embedded. The news items are viewed as complex entities in which attributes define different components, e.g., title, abstract, text, photos, videos, commentaries, etc. The system employs a user model initialized using orthogonal stereotypes (interests, domain expertise, cognitive characteristics, and life styles) for which the user is asked to provide input and is further updated using rules that exploit the user behavior with the application. Taking into account the user model, the system builds a presentation based on relevant news items, each news item being shown at an appropriate level-of-detail (based on the user model). Differently than SeAN, our framework uses standard Semantic Web technologies for representing knowledge and employs NLP techniques for automatic annotation of news items, instead of using a manual approach.

YourNews (Ahn, Brusilovsky, Grady, He, & Syn, 2007) proposes an open and editable user model for personalizing news items. The user model is open in the sense that users can view the list of keywords stored in the individual profiles. The user model is also editable as it allows users to add/delete keywords from their associated profiles. As an additional feature which also contributes to the transparency and control over adaptation, YourNews shows the key terms present in news items. The representation of news items is given by weighted vectors of terms (Salton, 1971), where the weights are computed using TF-IDF (Salton & McGill, 1983). The visited news items are used for building a weighted term vector which is the user model. The similarity between a news item and the user model is defined by the cosine metric between their associated vectors. This measure allows the system to recommend news items that are considered relevant for the user. Despite the users' interest to view and edit their profiles, there is a decrease in performance (e.g., precision, recall, etc.) for recommended items compared to the same system using a closed user model (where the user is not able to view/edit the user model). While YourNews uses a keyword-based approach for modeling news items and user interests, Hermes employs a semantic approach based on ontology concepts for modeling similar aspects.

MyPlanet (Kalfoglou, Domingue, Motta, Vargas-Vera, & Shum, 2001) aims at providing users with news items relevant for their topics of interest. MyPanet is an extension of PlanetOnto, an integrated suite of tools used to create, deliver, and query internal newsletters of the Knowledge Media Institute (KMi). Similar to our approach an ontology is used for classifying news items and allowing the user select his topics of interest. Nevertheless, the classification process is based on the heuristics of cue phrases attached to ontology concepts, while we have a more systematic approach to classification by employing NLP techniques (e.g., exploiting the WordNet term synonyms, performing word sense disambiguation, etc.) that improve classification results. In addition, our implementation is based on the standard ontology language OWL instead of the specific ontology language, OCML (Motta, 1999), used in myPlanet. We also did choose to present the ontology as a graph instead of a tree as it allows the user to have a more comprehensive overview of the ontology structure.

SemNews (Java, Finin, & Nirenburg, 2006) proposes a framework for understanding and querying news items. As in Hermes, the monitored news are made available by RSS feeds. The news items are analyzed by OntoSem (Nirenburg & Raskin, 2001), SemNews' natural language processing engine. OntoSem converts the textual representation of news into Text Meaning Representation (TMR), a specific format for knowledge representation. The TMR descriptions are subsequently converted to OWL and published on the Web. The OWL news representation can be used for querying using RDQL (Seaborne, 2004), one of the precursors of the SPARQL query language. Differently than SemNews, Hermes uses a semantic lexicon (e.g., WordNet) for performing word sense disambiguation, and allows for a more intuitive way of building queries by letting the user make his selections in a graphical way.

The contributions that Hermes brings to building personalized news services compared to existing approaches are fivefold. First, Hermes makes a strict distinction between the framework (Hermes framework) and its implementation (HNP), allowing for possible different technologies (as these evolve) to be used with the same framework. Second, Hermes uses an advanced NLP methodology (e.g., tokenization, part-of-speech tagging, word sense disambiguation, etc.) for news understanding employing a semantic lexicon (e.g., WordNet). Third, the implementation is based on the most up-to-date Semantic Web standards (OWL and SPARQL). Fourth, we allow news querying using temporal constraints by providing temporal extensions to SPARQL. Fifth, and the last contribution, the user is provided with a graphical query interface to specify the concepts of interest in a direct (using concept selections) or indirect manner (using relationship selections).

HERMES FRAMEWORK

The Hermes framework proposes a sequence of steps to be followed in order to build a personalized news service. The input for the constructed system comprises RSS news feeds and the output consists of news items fulfilling user needs. The Hermes framework is centered around a domain ontology which is used for indexing news items and helping the user formulate queries based on his concepts of interest. In addition, the user can specify temporal constraints that news items need to satisfy. The resulting news items are sorted based on their relevance for the user queries.

For illustrative purposes we chose a financial domain example, i.e., a personalized news service which can help the stock brokers in their daily decisions. More precisely we opted for portfolios based on stocks of NASDAQ companies. For this purpose we developed a domain ontology, which captures companies, products, competitors, CEO's, etc. In addition we have developed a news ontology able to store news items and their metadata such as title, abstract, time stamp, etc.

The domain ontologies are developed by domain experts. The process of developing the ontology is an incremental middle-out approach. First the most salient concepts are defined and then these are refined using generalization/specialization towards the top/bottom of the ontology. As the news information can contain additional concepts not a priori known, the ontology needs to be regularly maintained by the domain experts. We validated our domain ontology using the OntoClean methodology (Guarino & Welty, 2002).

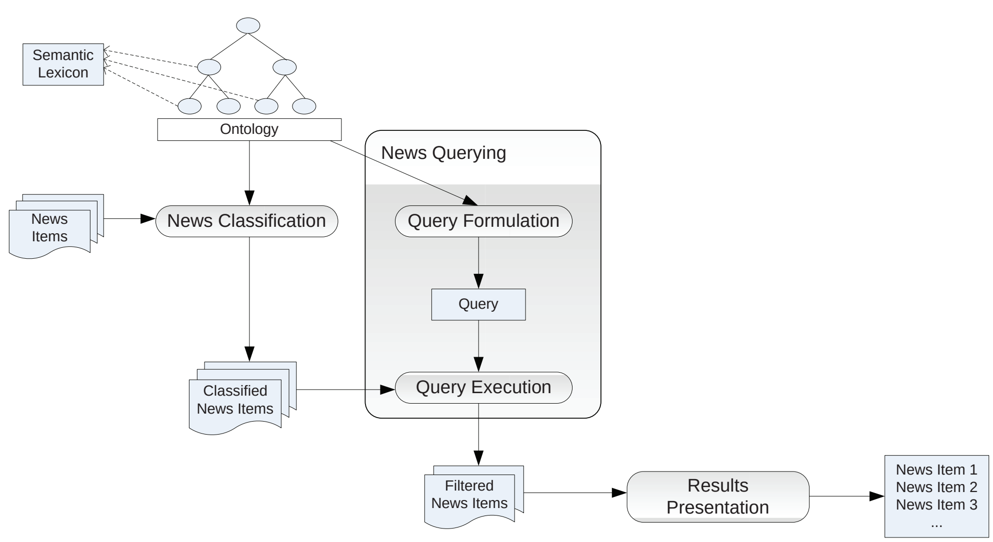
As news items might come from different RSS feeds, it is possible that the same news item has been published through different channels. After aggregating the news items one needs to remove the duplicate ones. In order to speed-up this process we employed the heuristics to use only the title for identifying news items, even so we acknowledge that in few cases news items might have the same title and still have different content, i.e., they represent different news items.

The architecture of the Hermes framework is described in Figure 1 and is composed of three main steps: *news classification*, *news querying*, and *results presentation*. *News classification* is responsible for indexing the news items based on ontology concepts. *News querying* consists of two substeps: *query formulation*, i.e., helping the user build the query that expresses the items of interest, and *query execution*, i.e., computing the results of query evaluation. In the last step, *results presentation* the computed news items are presented based on their relevance to the user interests.

NEWS CLASSIFICATION

The ontology concepts used for news items classification are classes and individuals from the domain ontology. Concepts are linked to synsets, i.e., sets of synonyms, from a semantic lexicon, which identify their unique meaning. The synonyms stored in a synset in the semantic lexicon are used as lexical representations of the associated

Figure 1. Hermes architecture



concept. In this case a lexical representation has a sense associated with it, i.e., the one given by the corresponding synset. As the ontology is specific to a domain, while the semantic lexicon is domain independent, we associate additional domain specific lexical representations to the concepts in our ontology. The lexical representations are composed only of word lemmas (the canonical word form appearing in dictionaries).

In addition, for classes without subclasses and individuals we decided to consider the hyponyms associated to their corresponding semantic lexicon synset. For these classes and individuals, the domain expert, who devised the ontology, is probably not interested in more refined definitions of these. However, the lexical representations of these concepts can be enlarged by considering also the corresponding hyponyms synsets from the semantic lexicon.

The classification approach is *ontology-centric*, in the sense that the ontology concepts are loaded one at-a-time and their lexical representations are matched against the news items. This approach is opposed to a *news items-centric* one where the words in the news items are matched against the lexical representations of the concepts from the ontology. We opted for an ontology-centric approach in order to speed-up the classification process as in this case we need to traverse the ontology only once. The number of concepts in the ontology is considerably larger than the number of words in the news items.

First the tokenization, sentence splitting, part-of-speech tagging, and morphological analysis are performed. The tokenization precedes sentence splitting as sentence splitting needs the identification of the punctuation signs from tokenization (as ".", ",", etc.). Morphological analysis follows part-of-speech tagging because the lemma of a word depends on its part-of-speech tag. For example "reading" as a verb has lemma "read" but as noun it has the lemma "reading". In this way, all words in a news item are reduced to their canonical form, a form shared also by the lexical representations of concepts stored in the domain ontology. For lexical representation identification we use the maximal group of words (compound words) found in sequence in a news item that stand for a concept's lexical representation. For example "European Central Bank" would be identified as a compound word representing the ECB concept, e.g., a longer match supersedes a shorter match.

Each time a lexical representation of a concept is matched a *word sense disambiguation* procedure takes place. As the same lexical representation can belong to different concepts (present or not in the ontology) this procedure checks if the match indeed corresponds to the meaning of the found ontology concept. If the check is positive a *hit* is stored in the ontology, i.e., a link between the news item and the corresponding concept is defined. The hit also stores the found lexical representation, as classification evidence.

For word sense disambiguation we use a variant of the SSI algorithm (Navigli & Velardi, 2005). In this process we also consider lexical representations for concepts that are not stored in the ontology but are present in the semantic lexicon as these are relevant when computing the sense of a found lexical representation from the ontology. These lexical representations help in better determining the context of a sentence and thus computing the sense of an ontology-based lexical representation.

The algorithm determines, per news sentence, the sense of a lexical representation (*lex*) by computing the sum of the distances between one of the senses of the considered lexical representation (s_j) and the senses of the previously disambiguated lexical representations from the context sentence (sc_i). The sense corresponding to the smallest sum is the chosen one (*selectedSense*). The algorithm starts with monosemous lexical representations (i.e., lexical representations which correspond to only one concept) and in case that such representations do not exist a guess is made by picking the most common sense for one of the found lexical

Box 1.

$$selectedSense(lex) = \arg \min_{s_j \in senses(lex)} \sum_{sc_i \in I} d(s_j, sc_i) \qquad (1)$$

representations. These senses are added to the context (I) of the sentence. For each remaining polysemous lexical representations two steps take place: a disambiguation step to find the correct sense and an insertion step which adds the newly computed sense to the context. In this way the context gets enlarged with new senses that model the meaning of the sentence. Formula (1) specifies what sense is selected for each lexical representation (See Box 1).

The distance between senses is defined as being inverse proportional with the length of the shortest path between the corresponding synsets ($sense_i$ and $sense_j$) in the semantic lexicon graph. The graph is based on the hypo/hypernyms relationships stored in the semantic lexicon. Path lengths larger than a predefined constant (e.g., 4 or 5) are not used (for these cases the distance is considered infinite). In this way we employ only semantically close concepts that truly help in the disambiguation procedure and also improve the speed of the process by not using arbitrarily large paths. Formula (2) shows how to compute the distance between two senses (See Box 2).

As the distances between synsets are not changing, one can pre-compute these, thus further reducing the time needed for the disambiguation step. The Hermes framework can be used with other methods for computing the similarity between concepts as for example string metrics (e.g., Levenshtein, Editex, etc.) or lexical co-occurrences in a corpus (e.g., pairwise mutual information, Google distance, etc.). Nevertheless, many of these methods are less precise than the graph-based method used here, as they do not take in account senses, comparing only the lexical representations of concepts.

NEWS QUERYING

The user expresses the topics of interests by posing queries using concepts from the domain ontology. In addition the user can express constraints that the timestamps associated to news items need to satisfy. The news querying step consists of two substeps: query formulation, i.e., supporting the user to build queries, and query execution, i.e., computing the results of query evaluation. Below we present in detail each of these substeps.

Query Formulation

In order to assist the query construction process we present to the user the ontology graph. We decided to show a graph-based representation of the ontology instead of a tree-based representation, as it gives more insight in the overall structure of our domain. For example a graph representation captures more relationship types instead of a singular relationship type, often the subsumption relationship, from a tree-based representation. The

Box 2.

$$d(sense_i, sense_j) = \frac{1}{length(shortestPath(synset(sense_i), synset(sense_j)))} \qquad (2)$$

user needs to understand the different relationship types in order to be able to build his query.

By using the ontology graph the user can select the *direct concepts of interest*. In addition he is able to specify concepts of interest using a keyword search facility through the graph. For this purpose the input keyword is checked for possible inclusion in the lexical representations of concepts. If such inclusions are found the corresponding concepts are being returned as possible direct concepts of interest. It is the task of the user to accept these as direct concepts of interest or to reject them.

One of the important functionalities of the Hermes approach is that it allows for the selection of concepts indirectly linked to the selected ones, concepts which are not a priori known to the user. We call these concepts *indirect concepts of interest*. For the selection of the indirect concepts of interest the user can state the type of the relationships that links the direct concepts of interest to the indirect concepts of interests or leave this undefined in which case all relationship types are being considered.

Suppose that the user has selected the direct concept of interest Google from the NASDAQ domain ontology. The user also specifies that he is interested in news related to the competitors of `Google` by selecting the `hasCompetitor` relationship. That means that `Yahoo!`, `Microsoft`, and `EBay` will be selected as indirect concepts of interests, without the user having to know the exact names of Google's competitors. All this background information is being extracted from the ontology.

The direct, indirect, keyword-based search concepts of interest, and the other concepts from the ontology need to be emphasized in a graph by using for example different colors. In this way the user is able to know, by analyzing the graph, why a certain concept is being highlighted. As the size of the graphs can be very large the user is provided with zooming/panning facilities for visualizing the ontology.

The original graph of the domain ontology is also called the *conceptual graph*. Based on the user selection, a new graph is generated, the so-called *search graph* that contains only the concepts and concept relationships relevant for the query. The user can go back and forth between the two graphs performing new selections and thus updating the search graph with new concepts.

The search graph is given by the subgraph of the conceptual graph that models the user interests which is equivalent to the answer of a conjunctive query based on selected binary predicates and concepts (that form graph patterns) while keeping the selected relationships between the concepts in the result set. The search graph has disjunctive semantics with respect to the included concepts which means that the user is interested in *any* of the search graph concepts.

Another crucial functionality of the Hermes approach is that it allows the specification of temporal constraints for news items. As news items appear at a certain moment in time and have certain time validity, it is important to be able to restrict the timestamps associated with the news. For this purpose the user can employ time comparison/arithmetic operators and retrieve the current time in order to build complex time expressions. In addition the system provides predefined temporal constraints such as: last day, last week, last two weeks, last three months, last quarter, last half year, and last year. The temporal conditions that model these constraints have conjunctive semantics as they need to be fulfilled in the same time.

Query Execution

Based on the previously selected concepts and specified temporal constraints the system can support the generation of the corresponding query in a semantic query language. This translation process involves mapping concepts and temporal constraints to query restrictions. After that, the user can trigger the query evaluation and the relevant news items are retrieved. The order of the retrieved news items is not relevant, at this stage.

RESULTS PRESENTATION

The results returned from query evaluation are presented in the order of their relevance for the user query. For this purpose, for each returned news item a *relevance degree* is computed based on all the hits between the news item and the query concepts. News items with high relevance degree are placed at the top of the retrieved news items list.

Based on previous work (Micu, Mast, Milea, Frasincar, & Kaymak, 2008) the relevance degree is defined as a weighted sum of the number of hits ($n(c_i)$), where the weights (w) depend on hits location (*title* or *body* of a news item). From our experimental results we have determined as acceptable values for w_{title} to be *2* and for w_{body} to be *1.* Formula (3) presents how to compute the relevance degree (See Box 3).

News items that have the same relevance degree are sorted in descending order based on the associated timestamps (the most recent news items are presented first).

In addition to presenting the relevant news items, the system shows the query concepts in order to provide cues of the current query for which the results are computed. Also, for each returned news item, the found lexical representations stored in the hits are emphasized in the news item text, thereby offering to the user an explanation of why a certain news item is considered to be relevant.

HERMES NEWS PORTAL

The Hermes News Portal (HNP) is an implementation of the Hermes framework, which allows the user to specify queries on the considered domain using temporal constraints and subsequently retrieve the relevant news items. The presentation of HNP follows closely the steps proposed by the Hermes framework. Note that HNP is one of the possible implementations of the Hermes framework, with specific design choices, query/programming languages, and libraries used.

Operating on the Semantic Web we chose as ontology language OWL due to its expressivity and standard status. We did not opt for RDFS because OWL specific features, as for example relationships inverses (`hasCompetitor` has as inverse `isCompetitorOf`), are exploited in the conceptual graph. Lacking a true OWL query language we used SPARQL as the query language, an RDF query language that we extended with time-related functionality. This functionality is provided by implementing comparison/arithmetic time operators and functions for retrieving current time information.

The chosen implementation language is Java due to the availability of powerful libraries for manipulating, reasoning with, querying, and visualizing OWL ontologies. For manipulating

Box 3.

$$relevanceDegree(news\ item) = \sum_{\substack{c_i\ found\ in\ title \\ c_i \in O \cap news\ item}} w_{title}\, n(c_i) + \sum_{\substack{c_i\ found\ in\ body \\ c_i \in O \cap news\ item}} w_{body}\, n(c_i) \qquad (3)$$

Figure 2. News item example

Google to broker print ads in newspapers
6 November 2006 17:41 CET

SAN FRANCISCO (Reuters) - Google Inc. is set to begin helping customers buy advertisements in 50 U.S. newspapers in a test of how the Web search leader can extend its business into offline media, the company said on Sunday.

and reasoning with OWL ontologies we used Jena (Jena Development Team, 2008a). For querying we employed ARQ (Jena Development Team, 2008b), the SPARQL implementation available in Jena. For visualizing ontologies we adapted the generic graph visualization library Prefuse (The Berkeley Institute of Design, 2008) for visualizing OWL graphs (Borsje & Giles, 2008). For part-of-speech tagging we used the Stanford parser (The Stanford Natural Language Processing Group, 2008). As a semantic lexicon we employed WordNet (Princeton Cognitive Science Laboratory, 2008), the largest database available online for the English language. JWI (Finlayson, 2008) was used for the morphological analysis (finding lemmas of words) and the communication with WordNet.

We illustrate the HNP by means of the following user query: *retrieve all news items related to Google or one of its competitors that appeared in the last three months.* For this query we will go through all the different phases of Hermes: news classification, news querying, and results presentation. In the current HNP the news items duplicates removal has not been yet implemented.

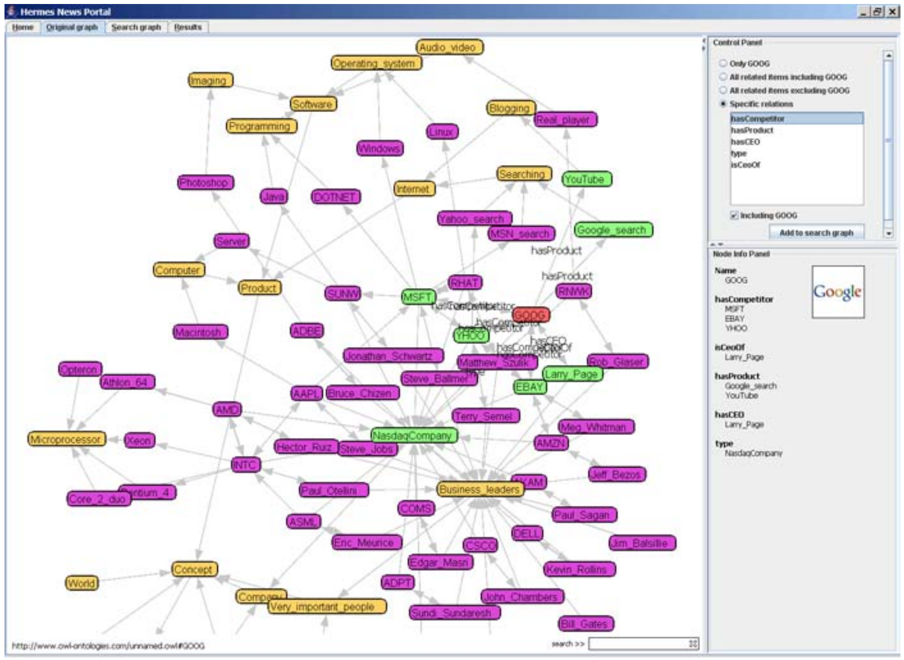
NEWS CLASSIFICATION

The news classification step is responsible for indexing news items based on the domain ontology. We present this process by means of the news item example depicted in Figure 2. The first line describes the title, the second line shows the timestamp, and the remaining text represents the content of the news item.

The news classification step starts by identifying lexical representations of the ontology concepts in the news item. First the tokenization, sentence splitting, part-of-speech tagging, and morphological analysis take place. Then, the concepts from the ontology are traversed once and their lexical representations are matched against the content of the news item. The following lexical representations "Google", "extend", and "company" are found. Next, per sentence, for each of the lexical representations with multiple senses, the word sense disambiguation procedure takes place in order to identify the used senses. The noun "Google", having only one sense, doesn't undergo the word sense disambiguation procedure and is mapped to the Google concept.

For "extend" and "company" a word sense disambiguation procedure is needed. In this process we do consider also lexical representations of concepts outside the ontology, as for example the nouns "customer" and "business", or the verb "buy" that do appear in the news item. We select the sense that yields the smallest sum of similarities to previously disambiguated lexical representations. For "extend" it is determined as representing the `extend-verb-#1` concept with lemma extend, part-of-speech tag `verb`, and the first sense from WordNet. For "company" the corresponding concept is `company-noun-#1`. "customer", a lexical representation outside the ontology, is determined as having the sense `customer-noun-#1`.

Figure 3. Conceptual graph example



After identifying an ontology concept in a news item, a hit is stored. This hit is defined as a link between the news item and the concept together with the found lexical representation. For this purpose we decided to model a hit as an instance of the `Relation` class, which uses different properties for storing the involved news item, concept, and found lexical representation. This modeling choice is based on a best practice for modeling N-ary relations on the Semantic Web (Noy & Rector, 2006).

NEWS QUERYING

Query Formulation

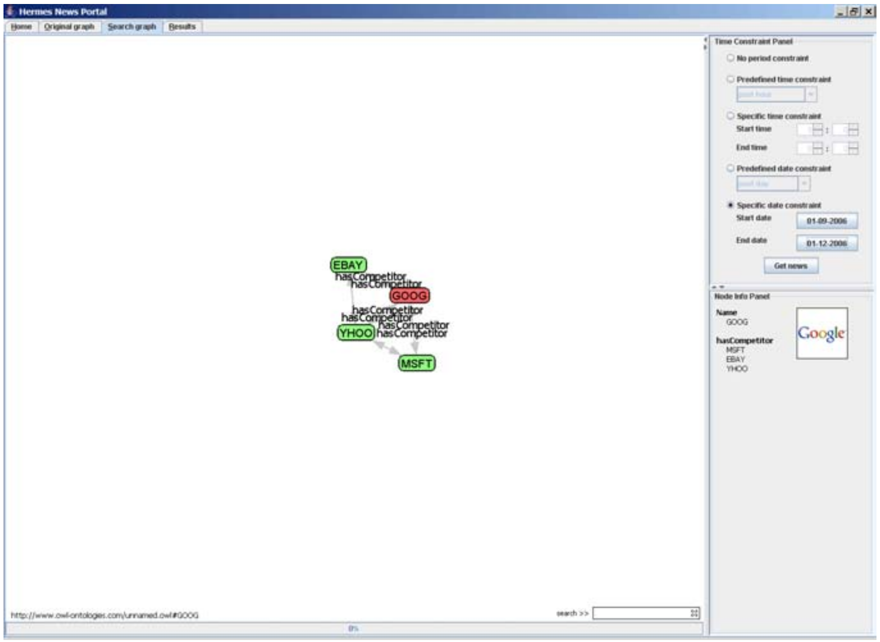
Figure 3 shows the conceptual graph from which the user can select concepts of interest. Once a user selects a concept, the control panel gets activated by means of which the user can add to the search graph his concepts of interest.

The concepts of interest can only be the current concept, all related concepts including the current one, all related concepts excluding the current one, or all related concepts by means of specified relationships including/excluding the current concept. The node info panel shows information regarding the selected concept.

The local name of the different concepts is displayed using ovals. In order to emphasize the different types of concepts we use the following coloring scheme for ovals. The selected concepts are displayed in red, the concepts related to the selected one are shown in green, and the ones returned by the keyword search are presented in pink. The other concepts are displayed in yellow for classes and magenta for individuals.

In the example from Figure 3 the user has directly selected the `Google` concept. In the node info panel all the information related to the `Google` concept is displayed: name, competitors, CEO, etc. Then, the user can select the indirect concepts by specifying the `hasCompetitor`

Figure 4. Search graph example



relationship between the direct concept and the indirect ones. The user also specifies that the `Google` concept should be kept in the search graph.

After selecting the concepts of interest the user can visualize the search graph. The user can refine its query by deleting some of the concepts, resetting the search, or adding new concepts from the conceptual graph to the search graph. After a number of iterations the user has added all the concepts of interests to the search graph.

Figure 4 shows the search graph, which presents the selected concepts, in this case `Google` and its competitors: `Microsoft`, `Yahoo!`, and `EBay`. In the time constraint panel the user can specify the desired time restriction. The user can choose between predefined temporal constraints as the past hour, past day, past week, past two weeks, past three months, past quarter, past half year, and past year, and specific time/date constraints.

As before the node info panel displays information about the currently selected node. However, differently than in the previous situation, only the information given by the specified relationships, in this case the `hasCompetitor` relationship, is displayed.

In this example the user has specified that the news items have to be between 1st of September 2006 and 1st of December 2006, where the current day is 1st of December 2006 (the *last three months* in the user query). Alternatively the user could have selected the past three months option from the predefined temporal constraints.

Query Execution

For each search graph a SPARQL query is generated. This query is a SELECT query as it retrieves the news items in which any of the search graph concepts are present. The disjunctive semantics of the search graph with respect to its embedded concepts is naturally specified as an ‘or’ filter in the SPARQL query. Also, we have decided to use filters to specify the time restrictions that news timestamps need to satisfy. Due to the conjunctive

Figure 5. SPARQL query example

```
PREFIX hermes: <http://hermes-news.org/news.owl#>
SELECT ?title
WHERE {
   ?news hermes:title ?title .
   ?news hermes:time ?date .
   ?news hermes:relation ?relation .
   ?relation hermes:relatedTo ?concept .
   FILTER (
   ?concept = hermes:Google ||
   ?concept = hermes:Microsoft ||
   ?concept = hermes:Ebay ||
   ?concept = hermes:Yahoo
   ) .
   FILTER (
   ?date > “2006-09-01T00:00:00.000+01:00” &&
   ?date < “2006-12-01T00:00:00.000+01:00”
   )
}
```

Figure 6. Custom time functions

```
    xsd:date currentDate()
    xsd:time currentTime()
xsd:dateTime dateTime-add(xsd:dateTime A, xsd:duration B)
xsd:dateTime dateTime-substract(xsd:dateTime A, xsd:duration B)
xsd:duration dateTime-substract(xsd:dateTime A, xsd:dateTime B)
```

Figure 7. tSPARQL query example

```
PREFIX hermes: <http://hermes-news.org/news.owl#>
SELECT ?title
WHERE {
   ?news hermes:title ?title .
   ?news hermes:time ?date .
   ?news hermes:relation ?relation .
   ?relation hermes:relatedTo ?concept .
   FILTER (
   ?concept = hermes:Google ||
   ?concept = hermes:Microsoft ||
   ?concept = hermes:Ebay ||
   ?concept = hermes:Yahoo
   ) .
   FILTER (
      ?date > hermes:dateTime-substract(hermes:now(), P0Y3M) &&
      ?date < hermes:now()
   )
}
```

semantics of the time restrictions we modeled them as an ‘and’ filter in the SPARQL query.

Figure 5 shows the SPARQL query corresponding to the search graph given in Figure 4. This query is hard-coded with XML Schema `dateTime`s specifying the desired temporal boundaries of the interval in which the timestamps of the desired news items need to be contained.

The first part of the SPARQL query defines that the returned news items should be related to the concepts of interest. The second part of the query is composed of two filters. The first SPARQL filter defines the concepts of interest to be `Google`, `Microsoft`, `Ebay`, and `Yahoo!`. The second SPARQL filter specifies that the timestamp of news items should be between 1st of September 2006 and 1st of December 2006, where 1st of December 2006 is the current day. Both dates are specified using XML Schema `dateTime` format.

In order to ease the specification of temporal constraints in queries we have extended SPARQL with custom functions. We call the SPARQL language extended with time functions tSPARQL. Please note that SPARQL does naturally support such extensions, tSPARQL being backwards compatible with SPARQL. Figure 6 shows the signature of the time functions that we have added. These functions relate to retrieving the current `date` and `time`, the current `dateTime` instance, adding/substrating to a `dateTime` instance a `duration`, and subtracting two `dateTime` instances (`date`, `time`, `dateTime`, and `duration` are defined by XML Schema).

Figure 7 depicts the same query as in Figure 5 but now written in tSPARQL. Different than in the previous case the tSPARQL query is not hard-coded with times and dates, but makes use of custom functions and durations. The semantics of the query is closer to its representation, in our current example that is retrieving the news items *that appeared in the last three months.*

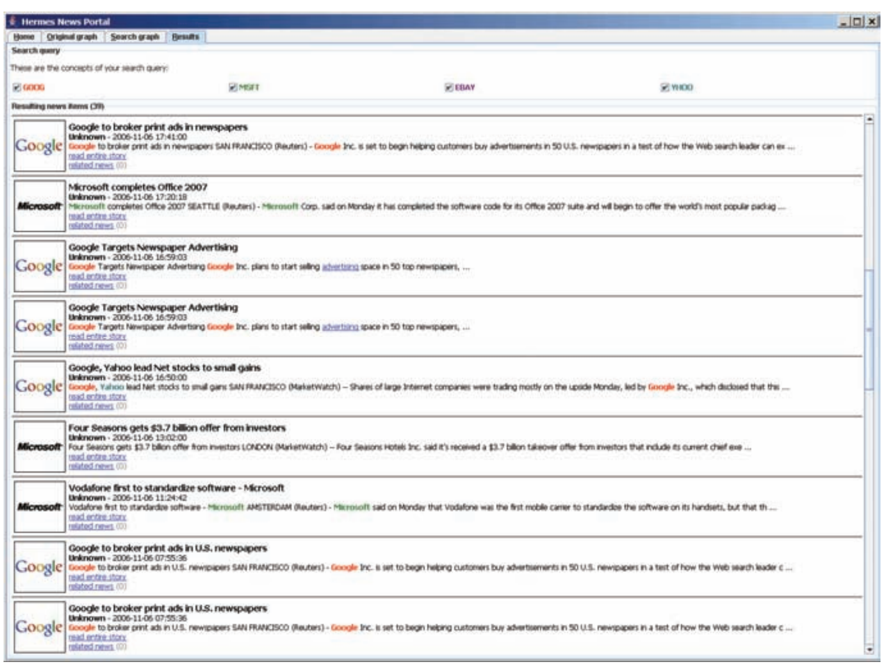
The tSPARQL query uses the `dateTime-substract()` to determine the `dateTime` of three months ago, `now()` is used to obtain the current `dateTime`, and it specifies that the news timestamps should be between these two `dateTimes`. `P0Y3M` is an XML Schema `duration` constant that specifies a period with 0 number of years and 3 months.

After its creation, the tSPARQL query is executed. As a result, the news items that match the query constraints (concepts of interest and temporal constraints) are being returned. The order of the results is not relevant here.

RESULTS PRESENTATION

Figure 8 shows the results after the query execution. It lists the concepts of interest from the search

Figure 8. Results presentation example



graph (top) and the retrieved news items (bottom). In addition, the system shows the found lexical representations of the concepts of interests in the returned news items using different colors. The user is able to deselect some of the concepts of interest in order to refine his query and thus limit the result set. At the current moment the relevance degree and timestamps-based sorting of relevant news items as proposed in the Hermes framework have not been yet implemented in the HNP.

EVALUATION

In order to evaluate the performance of the implementation we measured the concept identification precision. Precision was defined as the number of concepts correctly identified in the news items divided by the number of concepts identified in news items. If we define recall as the number of concepts correctly identified in the news items divided by the number of concepts that should have been identified in the news item, one can notice that for our application precision and recall are the same. The reason is that we are only looking for given concept lexical representation in news items which means that we identify all the concepts that should be recognized possibly with some errors. For our current implementation the precision (recall) is 85% for a given repository of around 200 news items.

Precision (recall) are based on cumulative errors through our application pipeline based on the HNP's part-of-speech recognition, morphological analysis, and word sense disambiguation algorithm. Despite using only WordNet as a semantic lexicon (compared to other approaches which combine several semantic lexicons, some domain specific (Navigli & Velardi, 2005)) we obtained high values for precision as many of our concepts' lexical representations are named entities (names of companies, CEO's, locations, etc.) that usually have only one meaning. The high value of recall can be explained by the fact that we do not aim at providing meaning for each (compound) word in news messages but only for the ones that correspond to lexical representations of ontology concepts. The meanings of the news' (compound) words that are not present in the ontology are used only to help the disambiguation process of found concept lexical representations.

A different metric for the performance evaluation is the latency of a news item in the concept identification phase. The obtained average latency time is around 30s which represents the time needed to process a news item from tokenization to concept recognition. The bottleneck lies in the disambiguation step for which distances between synsets need to be computed. As identified in the Hermes framework these distances can be precomputed (given a certain limit for the shortest path length between synsets) which would further reduce the disambiguation time.

Regarding usability we have asked 9 users (students at Erasmus University Rotterdam following a course on Semantic Web technologies including RDF(S), OWL, and SPARQL) to find news items for 3 given natural language queries in two ways: (1) using the Hermes implementation and (2) a SPARQL engine. Most students were able to correctly build the search graph and specify the temporal constraints, as well as the corresponding SPARQL query. All queries were faster specified using the Hermes framework than using SPARQL. Note that we do not claim that it is easier to use Hermes instead of SPARQL for querying RDF graphs, but for expressing a certain set of RDF queries (the ones supported by the search graph with temporal restrictions, which we consider typical for news querying) Hermes seems to be easier to use than SPARQL. Among the features mostly appreciated by students in Hermes were the graphical representations of the conceptual graph, the predefined time functions, and the visual cues employed for emphasizing concepts in returned news items.

Compared with traditional keyword-based search engines for news items (e.g., YourNews, SeAN, Google News, Yahoo! News, etc.) our semantic approach benefits from better precision, as it is able to disambiguate (compound) words, more query expressive power, because it allows the selection of indirect concepts (i.e., concepts not directly related to the items of interest), and the support for temporal constraints. A quantitative comparison with non-semantic based approaches is difficult to achieve due to the query limitations that these systems have and the impossibility of using the same news items as inputs in the compared systems.

In the tradeoff between expressivity and usability we decided to keep our queries simple with intuitive semantics so that a broad range of users (casual users, media analysts, stock brokers, etc.) should be able to use Hermes. Nevertheless, we acknowledge that an expert user might need more query expressivity (e.g., optional graph patterns, disjunctive semantics for temporal constraints, etc.) which contributes to the increase in complexity (and thus to the decrease in usability) of the framework. For this purpose we plan to extend the Hermes framework in the future with additional powerful functionality that would enable the generation of a news personalization service family (services targeting novice, average, or expert users).

CONCLUSION

The Hermes framework proposes a sequence of steps to be followed for building personalized news services. The input for these systems comprises RSS news feeds and the output are news items fulfilling user needs. The Hermes approach is based on a domain ontology used for classifying news items and to support the user define his concepts of interest. In addition the user can specify temporal constraints that the news item needs to obey. The Hermes News Portal (HNP) is an implementation of the Hermes framework. The domain ontology is specified in OWL and as a query language we used SPARQL. As a semantic lexicon we employed WordNet, one of the most popular English dictionaries available online. For representing temporal constraints we have extended the SPARQL language with temporal functions.

Differently than Google News and Yahoo! News, Hermes is able to exploit the background information stored in ontologies for retrieving user's items of interest. In this way the user doesn't need to explicitly define all the instances involved in the query by making use of the concept relationships for specifying his concepts of interest. In addition to the concepts of interest, the user is able to specify temporal constraints in his query. Another key feature of Hermes is the word sense disambiguation procedure, which is not used in related approaches as SeAN, YourNews, MyPlanet, or SemNews. The word sense disambiguation step increases the accuracy of news classifications, by making sure that the found lexical representations indeed correspond to the meaning of the domain ontology concepts.

As future work, we would like to extend the Hermes framework by employing multiple semantic lexicons and adding specific concepts to the ontology that are not captured in existing semantic lexicons. Some domain specific concepts (e.g., domain neologisms) are used in news items while current semantic lexicons, which are not up-to-date, do not include them. We also plan to exploit the structure of the domain ontology in order to compute the similarity between concepts. In this way we enrich our knowledge base and thus are better equipped in determining concept (synset) similarity.

In addition, we would like to introduce a learning step in which new ontology instances and relations are extracted from news items. For this purpose we envisage the use of lexico-semantic rules that would extract the relevant information from news items. For example a proper noun followed by "Inc." in a news item as in "ClearForest Inc." would indicate that the proper noun is a company which needs to be inserted in the ontology if it is not already present there. In addition we would like to explore the possibilities to redefine the domain ontology as a time-based representation, where instances have a certain time validity associated with them (Milea, Frasincar, & Kaymak, 2008). These temporal extensions to the ontology would enable us to better reason with the temporal contextual information available for our domain.

Another direction that we would like to pursue is that of semantic adaptation of news items based on a user model. The user preferences now represented in the (temporary) search graph would be represented in a (stored) user model which is continuously adapted based on user behavior. In order not to bother the user with already seen content, we would like to be able to filter news items that provide new information by using a novelty control mechanism (Gabrilovich, Dumais, & Horvitz, 2004). We believe that our semantic approach can be successfully applied for modeling dissimilarities between news items and thus be able to recommend only news carrying novel content. In a different scenario, by measuring the similarities between news items, we would be able to recommend news items related to the same story but issued at different moments in time and/or by different institutions.

Regarding HNP we would like to implement news items duplicates removal, and the relevance degree and timestamps-based sorting of relevant news items, as proposed in the Hermes framework. Additionally, we also wish to implement the previously proposed extensions to the Hermes framework: enriching our knowledge sources with multiple lexicons and domain-specific concepts, updating the domain ontology based on news information, adding a user model and employing it to adapt system functionality, and filtering news that provide novel content. Also, we would like to test the usage of data structures for fast data access (e.g., hash maps) for ontology access in a news-centric approach where the concept lexical representations are identified during a single news item traversal. Having a constant access time to concept lexical representations and taking in consideration that the number of lexical representations in a news item is smaller than in an ontology might reduce the time needed for the news classification step.

Additionally, we would like to conduct a more extensive evaluation procedure of the Hermes implementation. Based on detailed questionnaires and measuring the time spent on building queries given in natural language, we can obtain more empirical evidence on the system usability. The accuracy of the proposed relevance degree (based on concept identification) could be determined by measuring the access order and reading time of news item in the result list (accessing the first items first and spending substantial time for reading them are good indications that the returned items are relevant). In addition, we want to experiment with other domains (e.g., politics, sports, etc.) and analyze the precision and latency of our implementation for these new fields. The genericity of our approach only asks for the definition of a new domain ontology and domain-specific news feeds that need to be plugged into our implementation.

ACKNOWLEDGMENT

The authors are supported by the EU funded IST-STREP Project FP6-26896: Time-Determined Ontology-Based Information System for Real Time Stock Market Analysis (TOWL). More information is available on the official website of the TOWL project (TOWL Consortium, 2008). Also, we would like to thank Wouter Rijvordt, Maarten Mulders, and Hanno Embregts for their contribution to the Hermes framework.

REFERENCES

Ahn, J.-w., Brusilovsky, P., Grady, J., He, D., & Syn, S. Y. (2007). Open User Profiles for Adaptive News Systems: Help or Harm? In *16th International Conference on World Wide Web (WWW 2007)* (pp. 11-20). New York, NY: ACM.

Ardissono, L., Console, L., & Torre, I. (2001). An Adaptive System for the Personalized Access to News. *AI Communications, 14*(3).

Bechhofer, S., Harmelen, F. v., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., et al. (2004). *OWL Web Ontology Language Reference* W3C Recommendation 10 February 2004.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 284*(5), 34-43.

Borsje, J., & Giles, J. (2008). OWL2Prefuse. from http://owl2prefuse.sourceforge.net/index.php

Brickley, D., & Guha, R. V. (2004). *RDF Vocabulary Description Language 1.0: RDF Schema*: W3C Recommendation 10 February 2004.

Finlayson, M. (2008). The MIT Java Wordnet Interface (JWI). from http://www.mit.edu/~markaf/projects/wordnet/

Gabrilovich, E., Dumais, S., & Horvitz, E. (2004). Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty. In *13th International Conference on World Wide Web (WWW2004)* (pp. 482-490). New York, NY: ACM.

Guarino, N., & Welty, C. A. (2002). Evaluating Ontological Decisions with OntoClean. *Communications of the ACM 45*(1), 61-65.

Java, A., Finin, T., & Nirenburg, S. (2006). Text Understanding Agents and the Semantic Web. In *39th Hawaii International Conference on Systems Science (HICSS 2006)* (Vol. 3, pp. 62.62). Washington, DC: IEEE Computer Society.

Jena Development Team. (2008a). A Semantic Web Framework for Java (Jena). from http://jena.sourceforge.net/

Jena Development Team. (2008b). A SPARQL Processor for Jena (ARQ). from http://jena.sourceforge.net/ARQ/

Kalfoglou, Y., Domingue, J., Motta, E., Vargas-Vera, M., & Shum, S. B. (2001). *myPlanet: An Ontology-Driven Web-Based Personalized News Service*. Paper presented at the Workshop on Ontologies and Information Sharing (IJCAI 2001).

Kandel, E., & Marx, L. M. (1997). NASDAQ Market Structure and Spread Patterns. *Journal of Financial Economics*, *45*(1), 61-89.

Klyne, G., & Carroll, J. J. (2004). *Resource Description Framework (RDF): Concepts and Abstract Syntax*: W3C Recommendation 10 February 2004.

Micu, A., Mast, L., Milea, V., Frasincar, F., & Kaymak, U. (2008). Financial News Analysis Using a Semantic Web Approach. In A. Zilli, E. Damiani, P. Ceravolo, A. Corallo & G. Elia (Eds.), *Semantic Knowledge Management: An Ontology-Based Framework* (pp. 311-328). Hershey, Pennsylvania: IGI Global.

Milea, V., Frasincar, F., & Kaymak, U. (2008). Knowledge Engineering in a Temporal Semantic Web Context. In *The Eighth International Conference on Web Engineering (ICWE 2008)* (pp. 65-74). Washington, DC: IEEE Computer Society Press.

Motta, E. (1999). *Reusable Components for Knowledge Modelling: Case Studies in Parametric Design Problem Solving* (Vol. 53). Amsterdam, the Netherlands: IOS Press.

Navigli, R., & Velardi, P. (2005). Structural Semantic Interconnections: a Knowledge-Based

Approach to Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(7), 1063-1074.

Nirenburg, S., & Raskin, V. (2001). Ontological Semantics, Formal Ontology, and Ambiguity. In *Formal Ontology in Information Systems (FOIS 2001)* (pp. 151-161). New York, NY: ACM.

Noy, N., & Rector, A. (2006). *Defining N-ary Relations on the Semantic Web*: W3C Working Group Note 12 April 2006.

Princeton Cognitive Science Laboratory. (2008). A Lexical Database for the English Language (WordNet). from http://wordnet.princeton.edu/

Prud'hommeaux, E., & Seaborne, A. (2008). *SPARQL Query Language for RDF*: W3C Recommendation 15 January 2008.

Salton, G. (1971). *The SMART Retrieval System—Experiments in Automatic Document Processing*. Upper Saddle River, NJ: Prentice-Hall.

Salton, G., & McGill, M. J. (1983). *Introduction to Modern Retrieval*. New York, NY: McGraw-Hill.

Seaborne, A. (2004). *RDQL - A Query Language for RDF*: W3C Member Submission 9 January 2004.

The Berkeley Institute of Design. (2008). The Prefuse Visualization Toolkit. from http://prefuse.org/

The Stanford Natural Language Processing Group. (2008). The Stanford Parser: A Statistical Parser. from http://nlp.stanford.edu/software/lex-parser.shtml

TOWL Consortium. (2008). Time-Determined Ontology-Based Information System for Real Time Stock Market Analysis (TOWL). from http://www.towl.org/

Winer, D. (2003). *RSS 2.0 Specification*: Harvard Law School.

This work was previously published in IGI Publishing, edited by I. Lee, pp. 35-53, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 2.15
A Service Oriented Ontological Framework for the Semantic Validation of Web Accessibility

Rui Lopes
LaSIGE, University of Lisbon, Portugal

Konstantinos Votis
Pattern Recognition Laboratory, University of Patras, Greece Informatics and Telematics Institute, CERTH, Greece

Luís Carriço
LaSIGE, University of Lisbon, Portugal

Spiridon Likothanassis
Pattern Recognition Laboratory, University of Patras, Greece

Dimitrios Tzovaras
Informatics and Telematics Institute, CERTH, Greece

ABSTRACT

The Web serves as the principal mediator for information sharing and communication on a worldwide scale. Its highly decentralized nature affords a scale free growth, where each endpoint (i.e., Web site) is created and maintained independently. Web designers and developers have the onus of making sure that users can interact without accessibility problems. However, coping with users with disabilities poses challenges on how to ensure that a Web site is accessible for any kind of user. When (and if) this is done, designers and developers do it in a post-hoc way, (i.e., verify and tweak Web sites according to guidelines such as WCAG). In this Chapter the authors present SWAF, the Semantic Web Accessibility Framework, a base framework for supporting the integration of accessibility services into Web design and development processes. SWAF affords both tailoring accessibility to user needs and specifying the semantic validation of accessibility guidelines in different application situations.

DOI: 10.4018/978-1-60566-650-1.ch003

INTRODUCTION

The increasing adoption of technologies from users puts the Internet in a central spotlight. The Web, as its major application, is accessed and interacted by users at constantly increasing pace, allowing them to quickly seek information, interact with their peers through social networks, or perform transactions from the comfort of their homes. For this reason, the way information is structured and presented is critical for the success of accessing it.

However, users have their own specific characteristics (e.g., abilities, impairments, preferences, knowledge). Consequently, the accessibility of each piece of information – such as a Web page – can differ significantly from user to user. While often dismissed in the Internet realm and, more specifically, on the Web, people with disabilities are not just a small population minority. If one takes into account people with mild disabilities, the slice of the population that requires some sort of software-based accessibility ramp is of the uttermost importance.

The most important way to mitigate this problem is making sure that information providers (ranging from the individual to large corporations) do not overlook such accessibility issues. Internationally recognized organizations such as the World Wide Web Consortium (W3C, n.d.) play a critical role on helping information providers to cope with accessibility. Traditionally, this goes in the form of specifying accessibility-centric best practices, guidelines, and vocabularies to augment already existing Web languages.

Evangelization of accessibility practices, coped with the progressive intertwining of accessibility features in Web languages has brought Web accessibility more close to information providers. Consequently, each day, Web accessibility is gaining awareness. Guidelines such as WCAG, the Web Content Accessibility Guidelines (Chisholm, Vanderheiden, & Jacobs, 1999), are being followed more often, leading more users with disabilities to have access to information without barriers. In order to be so, these guidelines are presented as straightforward as possible, geared towards the largest set of Web designers and developers. However, mostly due to financial, human resources, and technological expertise problems, several companies (and individuals) totally dismiss the adequacy of Web sites to the different requirements of accessibility-dependent audiences, despite the fact that legislation is being pushed in several countries, in order to promote the rights of people with disabilities.

The dismissal of accessibility from information providers leverages the fact that such guidelines and standards for accessibility have inherent problems. Since they are specified in such a way that they require manual inspection of their conformance, developers have an increased effort on coping with accessibility issues. Furthermore, by being informally described (i.e., in natural language), they tend to lead to different interpretations from developers and accessibility experts and, for this reason, different and incoherent ways to ensure that a given Webpage is accessible.

When Webpages and Websites require an additional effort of supporting more specific and fine-grained audiences, the development of accessibility-centric solutions becomes cumbersome. Guidelines have an implicit assumption of which audiences they target to (and more often than not, in a very informal and loose way). Typically, these are often geared towards people with visually impairments (e.g., different kinds of blindness). The way users from this audience interact with a webpage is also left implicit, but often assumed to include specific devices (e.g., screen readers). This leads to a lack of understanding what is the role of guideline checkpoints for each user and device characteristic they are tackling, thus posing more difficulties on developers on how to tackle fine-grained accessibility analysis and consequent development of accessible Websites and Webpages. Consequently, software developers need to have a conceptual framework in which to situate disabled-related guidelines, which they often do not have due to lack of experience with disabled people and their technologies.

This Chapter proposes SWAF (Semantic Web Accessibility Framework) as an ontological framework targeted to accessibility-aware Web design and development processes. This will enable large organizations, small enterprises, or even individuals (developers, designers, etc.) to produce Web sites of superior accessibility and usability, accompanied with appropriate measures, technologies and tools that improve their overall quality. This ontological framework can be used to answer questions about common accessibility standards, user abilities and disabilities, as well as about the technical capabilities and constraints of appropriate assistive devices, thus forming the context for semantic validation of Web accessibility. In This Chapter we present the general overview of the framework, and detail it in the context of existing Web accessibility standards, in order to facilitate accessibility assessment of Web sites across different audiences.

BACKGROUND

Designing for people with disabilities is becoming an increasingly important topic for a variety of reasons, but especially due to recent legislation in many countries that aims at promoting and enforcing the rights of people with disabilities. A number of philosophies and methodologies have been developed to support this process.

Firstly there has been the development of the *universal design*, *design for all*, and *universal usability* philosophies, as detailed by Shneiderman (2000). Many developers worry that they will be expected to produce a system that will be usable by every user, regardless of their abilities, and that they might have to seriously compromise their overall design to achieve this aim. Clearly this would be in no-one's interest. With the increasing ability to personalize interfaces to meet the requirements of different users, this is not necessary.

Secondly, there have been numerous sets of guidelines to help developers produce systems that are accessible and usable by people with disabilities. These range from very general guidelines to the very specific guidelines for Web user agents (e.g., Web browser), authoring tools, and content creators. However, it is not clear whether providing guidelines is an effective method for ensuring usable designs, since these might be differently interpreted by developers and designers. Developers need to have a conceptual framework in which to situate disabled-related guidelines, which they often do not have due to lack of experience with people with disabilities and assistive technologies.

Consequently, even if people with disabilities want to be independent and do things for themselves by themselves, unfortunately, most Web sites and Web applications are not fully accessible today. There are a number of reasons for this as explained below:

- The Web has evolved over the latest years, and the importance of accessibility has only begun to be appreciated and encouraged in recent years. Older solutions are unlikely to be fully accessible (or accessible *at all*). Making an existing Web site accessible is often very difficult and expensive, in much the same way as making an existing building wheelchair-friendly can be very difficult, as well as un-aesthetic. Although efforts should be made to improve accessibility, it will be typically easier to do so during a major refurbishment.
- Many developers and, more surprisingly, designers, are not aware of the importance or need for accessibility. Consequently, new developments are being built in blissful ignorance, as many of them do not have the necessary knowledge or skills for building accessible Web sites.
- Some market stakeholders believe that creating accessible solutions will have prohibited costs and, at the same time, make them boring and less attractive to the majority of users (read: the non-impaired).
- Existing design and development tools give little out-of-the-box assistance in

most cases or, at worst, make it impossible to develop accessible solutions.

Accessibility Standardization

Up to now, there are several initiatives concerning guidelines, tools and technologies for Web accessibility. The major steering body for accessibility is the World Wide Web Consortium and its Web Accessibility Initiative (WAI, n.d.). WAI has three main tracks: the Web Content Accessibility Guidelines (Chisholm et al., 1999), the Authoring Tool Accessibility Guidelines (Treviranus, Richards, Jacobs, & McCarthieNevile, 1999), and the User Agent Accessibility Guidelines (Jacobs, Gunderson, & Hansen, 2002). The activities of W3C and WAI are the result of collaboration of groups and organizations from different countries, like the TRACE Research and Development Centre (TRACE, n.d.), which is responsible for compiling and publishing the original set of Web accessibility guidelines that provided the backbone of WAI guidelines.

Apart from the guidelines, there are also legislative and standards initiatives for accessibility. Strong governmental support in the Unites States has led to initiatives such as the Americans with Disabilities Act (ADA, 1990). The UK equivalent is the Disability Discrimination Act (DDA, 1995) amended in 1999, and now extended to the Special Educational Needs and Disability Act (SENDA, 2001). The rulings of ADA are also extended to Section 508 of the US Rehabilitation Act (Section 508, 1998). This legislation defines processes and the monitoring role of the US federal government in the procurement of electronic and information technology. Regarding accessibility, it states that regardless of medium, government must ensure that disabled federal employees and members of the public have the same accessibility as non-disabled members. Where accessibility is not present, government is directed to provide alternative means. Although, Section 508 is intended for the US federal government, many organizations and software houses worldwide are making efforts to address its mandate, which puts it as a central destination for Web accessibility verification practices.

Major standards bodies such as the US Human Factors and Ergonomics Society (HFES, n.d.) are engaged in furthering the accessibility drive. Their efforts extend to features and functions of the operating systems, drivers, application services, other software layers upon which the application depends and applications that increase accessibility with a general aim of reducing the need for add-on assistive technologies. The International Standards Organisation ISO/TS 16071:2003 (ISO 16071, 2003), Ergonomics of Human–system Interaction also provides guidance on accessibility for HCI interfaces. The guidelines were designed to complement general design for usability covered by related standards on usability.

As thoroughly discussed in this Section, there exist several initiatives and standardization bodies concerning guidelines, standards and methodologies for accessibility assessment that can be effectively applied in the context of Web technologies. It is also a fact that the existing standards and best practices concerning accessibility are in most cases confusing and incomplete (Lopes & Carriço, 2008). Therefore, developers need to have a conceptual framework in which to situate Web accessibility-related guidelines, which they often do not have, due to lack of experience on technologies for the disabled. The fundamental aspect of pushing forward accessibility on Web site design and development practices is to provide concrete and objective rules and standardized guidelines that homogenize accessibility assessment and quality control procedures. Consequently, existing software that aims at assessing accessibility based on such guidelines (a thorough list of such software packages can be found at http://www.w3.org/WAI/ER/tools/complete) will provide incomplete and overly generalized answers to whether a given Web page or Web site is accessible.

While it is clear that determining what truly represents accessibility in the customer's view can be elusive, it is equally clear that the number and frequency of problems and defects associated with a Web site are inversely proportional to its accessibility. Software problems and defects regarding accessibility are among the few direct measurements of software processes and products. Such measurements allow us to quantitatively describe trends in defect or problem discovery, repairs, process and product imperfections, and responsiveness to customers. Problem and defect measurements also are the basis for quantifying several significant software accessibility attributes, factors, user characteristics and criteria.

Although the advantages of measurement in the Web site design and development process are indisputable, the popularity of measurement methods, within accessibility terms, in practice is rather limited (McGarry, 2002; Varkoi, 1999; SEI, 2006). Very often difficulties arise when trying to focus the measurement. In many cases it is unclear what should be measured and also how the measurement data obtained should be interpreted (Habra, Abran, & Lopez, 2004; Kulik, 2000). Choosing the correct measurement entities and ranking the importance of measurement accessibility indicators is still a challenging task (Neely, 1998). Despite the difficulties, metrics such as those defined by Vigo et al. (2007) provide insightful cues on how to approach the problem of measuring the accessibility of Web sites, based on WCAG standards. This will help designers and developers to have a better (and measurable) understanding of accessibility on Web technologies.

Ontologies for Disability and Accessibility

There are several efforts towards the direction of the definition of ontological concepts and taxonomies for people with disabilities. These efforts try to cover adequately the personal requirements of the end users, including the person's disabilities and individual preferences.

A central reference for classifying disabilities concerns the World Health Organization's International Classification of Functioning, Disability and Health (ICF, n.d.), particularly tailored to impairment qualification on medical diagnosis tasks. Consequently, it stresses just on profound disabilities, leaving out several impairments such as color blindness. Obrenovic, Abascal, & Starcevic (2007) have leveraged ICF concepts into an accessibility description framework to help designers and developers discuss and describe multimodal interaction issues.

Gruber (1993) proposes an ontology architecture that tries to cover comprehensively the situation of persons with special needs for the purpose to utilize this information for customization of their home environment's services is proposed. This approach tries to combine contextual information like personal aspects (e.g., disabilities, preferences), technical aspects (e.g., equipment, services, network) and natural aspects (e.g., location, time) in a way that the smart home environment's services can adapt to the end user more or less automatically while keeping the user in control.

Several pre-existed ontologies for supporting context-aware smart environments, like CoOL (Strang, Linnhoff-Popien, & Fank, 2003), COBRA-ONT (Chen, Finin, & Joshi, 2003), CONON (Wang, Zhang, Gu, & Pung, 2004), SOUPA (Chen, Perich, Finin, & Joshi, 2004), and UbisWorld (Heckmann, 2005). All these ontologies share common concepts and structures. From these, SOUPA incorporates most concepts of previously defined ontologies and seems to be the most elaborated one of the listed ontologies. However, all of them still lack a specific support for persons with special needs towards a comprehensive specification.

A major contribution to the field of ontologies for disabilities was made from EU's FP6 ASK-IT project (ASK-IT, n.d.). Within ASK-IT, ontology modeling and mapping produced a collection of shared sub-ontologies, which reflect mobility impaired people user needs, and relationally map available services to them. These needs were initially specified and afterwards, the ontology authoring procedure was based on content models derived from these specifications. It also defines the interrelationships that may rationally hold between user groups of people with disabilities and various user information needs of different content types, including multi-modal content.

The potential for applying ontologies in end user diverse environments and their potential for promoting a unified methodology is exemplified by the ontology devised by Uschold, King, Moralee, & Zorgios (1998). This ontology includes lexical and relational terms based on the idea of the activity (anything that involves doing) linked to the doer or operative unit which may be a person, organizational-unit or machine said to have capability and on occasion possessing roles in respect of an activity such as activity-owner.

Wooldridge, Jennings, & Kinny (2000) also adopted a role-oriented analysis as a natural step in the Gaia methodology. Another example, the Framework for Distributed organizational Memories (Abecker et al., 2001), describes the various actors in domain ontologies according to their goals, knowledge and competencies. Van Heijst, Schreiber, & Wielinga (2000) also capture the role of an ontology in the accessibility requirements specification process, where they illustrate how a methodology can extract semantics from an ontology at different levels of depth to produce conceptual models.

As stated by Masuwa-Morgana & Burrell (2004), an ontology for accessibility requirements could be centered in a similar fashion, on an activity such as a use case in which there are doers (people and access technologies). The only difference is that in that ontology for accessibility requirements there would be a need to reliably furnish (with clear identities and essence) descriptions of doers-people and patterns of doer-access technology and subsequent competencies and demands on interface design and interaction styles. Abecker et al. (2001) propose "AccessOnto" as an accessibility requirements theoretic ontological framework consisting of four components: a requirements elicitation subsystem, an inference engine, a requirements explanation subsystem and an accessibility knowledge base. It has the intention of extrapolating a requirements specification based on rules extracted from the accessibility knowledge database based on end user traits data elicited by the end user.

The aforementioned ontological frameworks emphasize the fact that there is little coupling between ontologies regarding accessibility and disabilities, and Web accessibility assessment practices (as they tend just to frame different accessibility scenarios). Our proposed ontological framework will be based on existing ontological models, as well as in best practices for ontology engineering, affording the design of a multi-layer knowledge base for accessibility and disability requirements mapping into Web accessibility verification procedures that can provide support for the requirements and needs of different accessibility-centric user groups.

SEMANTIC WEB ACCESSIBILITY FRAMEWORK

For many people, in particular for groups at risk of exclusion, the complexity and lack of accessibility and usability of Web sites is a major barrier to information access. We respond to this challenge by proposing a tailored accessibility assessment ontological framework, the Semantic Web Accessibility Framework (*SWAF*), which affords the specification of user characteristics and their requirements, and associate them to specific accessibility assessment procedures.

The main goal of SWAF is to provide support for the formal and unambiguous definitions of accessibility domains, as well as the possible semantic interactions between them. We have specified SWAF to be integrated into accessibility verification environments (e.g., authoring tools, Web accessibility evaluators, integrated development environment - IDE). This will establish a common vocabulary for exchanging and describing the complex information that is related to accessibility assessment of Web sites. The framework aims to formalize conceptual information about:

- The *characteristics* of users with disabilities, devices, applications, and other aspects that should be taken into account when describing an audience with disabilities and developing tailored Web sites.
- Web accessibility *standards* and associated checkpoints.
- Semantic *verification rules* to help describing requirements and constraints of audiences, and associating them to accessibility checkpoints.

In order to cope with these goals, the framework must comply with the following requirements:

- To be as *formal* as possible, thus providing all the necessary definitions in a concise, unambiguous, and unified form;
- Provide information that can be *easily processed* by software applications and integrated into accessibility validation processes;
- *Easily implemented* by software developers and other users involved in the software development process of Web accessibility tools.

One of the main issues in designing and developing the proposed framework was to make it maintainable and extensible, while assuring model consistency within the framework. Therefore, we have separated SWAF into two distinct dimensions: *Web Accessibility Descriptions*, and *Web Accessibility Mapping*, as depicted in Figure 1. Each dimension is further explained in the following Sections.

Web Accessibility Descriptions

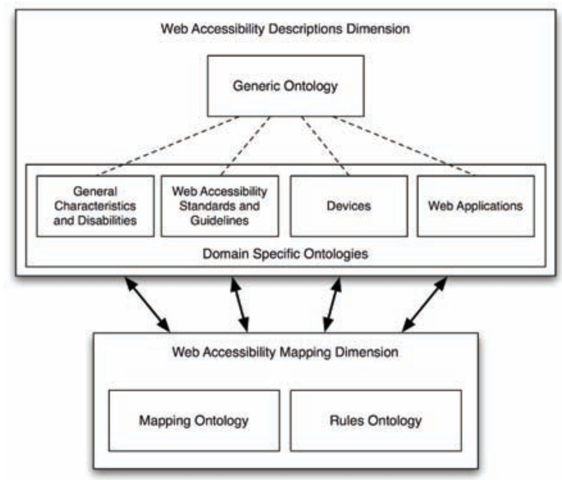
The first dimension provides constructs to describe different Web accessibility concepts (WAD). To explore the differences and synergies between Web accessibility fields, and to support the inclusion of external concepts from other domains, the WAD dimension cuts the concept space into a Generic ontology and a set of Domain Specific ontologies, as detailed next.

Generic Ontology

The Generic ontology forms the core ontology and describes top-level entities and concepts that are critical for the semantic validation of Web accessibility. Thus this ontology provides more abstract and generic knowledge such as general characteristics and disabilities of users, devices, Web accessibility standards, and other main aspects that constitute the basis for applying accessibility-based approaches into the accessibility validation field.

Domains are specified in classes and subclasses providing a hierarchical model representing all the knowledge fields that are necessary for the accessibility validation. There are also a number of properties denoting the relationship between classes. A part of the Generic Ontology is depicted in Figure 2. This partial snapshot of the ontology consists of the main classes *Disabilities*, *WAI_WCAG* and *Devices*. The *Disabilities* class contains three subclasses: *HearingImpairment*, *SpeechImpairment*, and *VisualImpairment*. There is also a property of the type *hasIncludedDomains* denoting that the classes *WAI_WCAG* and *Devices* include disabilities.

Figure 1. Semantic Web accessibility framework

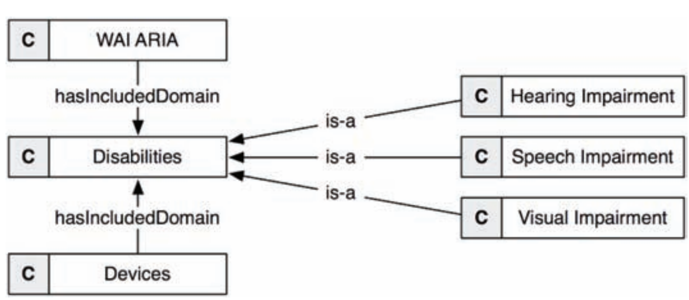


Domain Specific Ontologies

To better illustrate how the Generic Ontology can cope with real scenarios within the Web accessibility domain, we defined several domain specific ontologies and integrated them into SWAF. These domain specific ontologies (DSO) cope with the key aspects that are part of the integration of Web accessibility into Web design and development processes. This way, ontologies are able to represent a more detailed description of their corresponding domain, fruitful for extensibility scenarios (e.g., using Web accessibility validation ontologies in Mobile Web tailoring scenarios). The purpose of distinguishing a generic ontology from the domain specific ones is to facilitate the extension of SWAF to different application domains (e.g., outside the scope of Web related accessibility guidelines and applications).

Each DSO uses the basic entities of the Generic Ontology to describe the specific concepts and structures that are needed for the semantic validation of Web accessibility. This ensures that all terms and their relationships utilized by each accessibility approaches separately are included in the generic ontology scheme. To cover every spectrum of applicability of accessibility assessment procedures, there should exist a corresponding domain specific ontology. Next, some of these ontologies are described.

Figure 2. Excerpt of the generic ontology



General Characteristics and Disabilities Ontology

As discussed previously, validating accessibility is a process that must cope with user's disabilities, as well as with each individual's preferences. Thus it is of great importance to consider the users' personal capabilities determined by her/his impairments. Consequently, different categories of disabilities (based on the ICF categorization) are incorporated within this ontology, such as:

- *Visual impairments*. Disorders in the functions of the eye ranging form reduced capability of sight, color-blindness to total disability to see (e.g., cataracts or retinal detachment).
- *Hearing impairments*. Disorders in perceiving audio, ranging from problems in understanding normal conversations to complete deafness (e.g., high or low tone hearing loss).
- *Specific learning impairments*. Disorders manifested by significant difficulties in the acquisition and use of listening, speaking, writing, reading, reasoning, or mathematical abilities.

The degree of the users' disabilities determines the extent of Web accessibility concepts and guidelines that must be followed by Web sites (e.g., enlarging font sizes) to suit to users' computing environment and usage context.

To afford the specification of such concepts, the General Characteristics and Disabilities Ontology provides a set of supportive constructs at a meta-level (e.g., generalizations). The main concepts are *User* and *Characteristic*. A *hasCharacteristic* property maps characteristics to users, thus affording the description of users. We have further detailed several meta-concepts under the *Characteristic* umbrella. Since this ontology is tailored to accessibility scenarios (in the broad sense of *ability to access*), we introduced a small taxonomy to afford the classification of *Characteristic* instances. This taxonomy distinguishes *Ability*, *Disability*, and *Preferences*, as well as more specific concepts (e.g., *SensorialAbility* or *LearningDisability*). Accordingly, the *hasCharacteristic* property has been refined to cope with the three main domains.

Lastly, to afford a semantic extensibility and proper categorization of user characterization concepts, we defined an *extendsCharacteristic* property that maps between *Characteristic* instances in a taxonomical way.

All of these concepts provided by the ontology strive for a strong and expressive tool for Web designers and developers to describe and characterize their target users in a clean, thorough way, along the line of the descriptive ontologies devised by Obrenovic et al. (2007). Furthermore, by affording an extensible way of organizing user characterization concepts independently from users/audiences, Web designers and developers can build their own taxonomies with respect to their particular needs without being tied to a particular way of thinking and organizing information typical of stricter solutions.

Web Accessibility Standards and Guidelines Ontology

This domain ontology covers the main evaluation guidelines for Web accessibility assessment devised in the Web Accessibility Initiative, such as WCAG. These guidelines are divided into checkpoints and arranged based on their impact and priority. The combination of these factors is given in levels (none, A, AA, or AAA), depending on their evaluation outcomes. For instance, to claim conformance on level A, all the priority one checkpoints must be satisfied.

The table presented in the Appendix of this chapter reproduces the most fundamental Priority 1 checkpoints, which have been incorporated into this ontology. It is important to notice that some of these checkpoints may be irrelevant in different situations. Certain Web site instances might not have markup that can trigger accessibility problems. Furthermore, they might also be irrelevant based in the particular constraints and preferences of a user (from an accessibility point-of-view).

In particular, the checkpoints 1.2 and 9.1 apply (only) if image maps are used, the checkpoints 5.1 and 5.2 apply if tables are used, the checkpoint 12.1 applies if frames are used, the checkpoint 6.3 applies if applets or scripts are used, and finally the checkpoints 1.3 and 1.4 apply if multimedia is used. Most of these checkpoints are just relevant for those with visual impairments.

On the meta-level, this ontology introduces the *Guideline* and *Checkpoint* concepts, which can be mapped through an *includesCheckpoint* property. While Web designers and developers can leverage verification processes with the already supplied instances for WCAG, the extensibility provide by this meta-level affords the addition of new guidelines and checkpoints to their development processes in an effective way. Furthermore, this will allow them to leverage out-of-the-box all domain independent verification rules provided in SWAF's Web Accessibility Mapping Dimension.

Devices Ontology

Owing to the rapid development of electronic technologies, it tends to be common to access Web sites outside the traditional field of a desktop PC and a computer screen (e.g., PDAs, mobile phones, assisting devices, etc.). This has brought more specific assistive technologies to improve interactivity for users with disabilities, as well as broad personal preferences. This includes the ability of coping with diverse input/output modalities combination within interactive scenarios. Since the diversity of these technologies varies along different axes (e.g., display resolution, images coloring, multimedia process, etc.), the way accessibility is assessed for Web pages must also cope with these differences.

This ontology provides a simple set of meta-level concepts to describe devices ecology: a *Device hasFeature DeviceFeature*. Like in the description of users and accessibility scenarios, this ontology affords out-of-the-box instances for common cases of devices and device characteristics without closing the door to extensibility and odd-case scenarios.

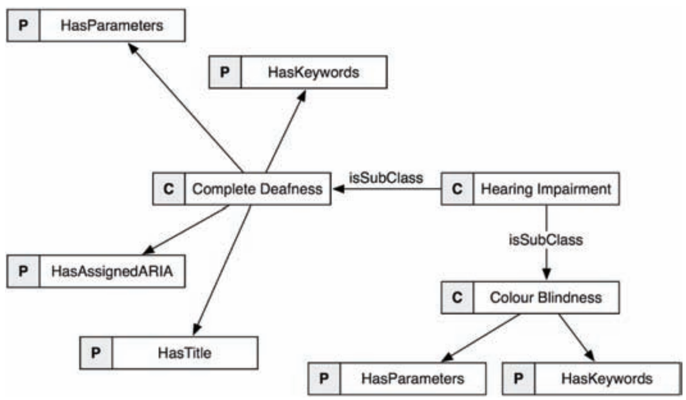
The description of devices can be used at two different fronts within development processes: it can help triggering semantic validation rules on user/device mismatches (e.g., user with total blindness and computer screen), and tying Web accessibility guidelines and checkpoints to particular devices and/or device features.

Web Applications Ontology

Web pages are not the only end for Web technologies. Nowadays, an increasingly number of applications is being ported from "traditional" desktop environments into Rich Internet Applications (RIA), by taking advantage of the richness of Web browsers, thus becoming easily available to any kind of users. However, since these technologies cannot cope with specific semantics of desktop applications widgets, accessibility issues may arise. The proposed domain ontology supports the development of accessible Web applications by affording the inspection of ARIA keywords (Accessible RIA) within Web pages, according to application requirements.

Figure 3 represents a partial snapshot of the ontology describing the knowledge domain of *HearingImpairment*. One of the main classes is the *CompleteDeafness* that consists of four subclasses *Title*, *Parameters*, *Keywords* and *AssignedARIA*. Each of these subclasses is characterized by a set of properties that can be either a simple data property or an object property that denotes the relationship between two classes.

Figure 3. Example of domain description ontology



This ontology supports both HTML-specific concepts (such as key HTML terms that have influence on accessibility issues), besides Web application domain concepts. This way, Web pages are perceived as a subset of Web applications from the point-of-view of key concepts, thus affording reuse scenarios of the ontology.

For example, the Web Applications Ontology provides the *GUICharacteristic* and *GUITechnology* abstract concepts (coupled with more fine-grained concepts such as *HTML*), to support the description of the Web Applications domain.

Web Accessibility Mapping

Finally, SWAF is completed with the Web Accessibility Mapping (WAM) dimension. This dimension aims to cover the establishment of mapping relationships between the ontologies of the first dimension, and to validate the semantics of these relationships. These relationships can be used, e.g., for efficient navigation and searching inside the ontologies, as well as to afford the creation of semantic rules-based accessibility verification. Two ontology layers are provided in this dimension (as detailed next): *Mapping Ontology*, and *Rules Ontology*.

Mapping Ontology

The WAM dimension provides a mapping ontology comprising of a set of lexical and notational synonyms to express the semantics of the relationships between concepts within the *General Ontology* and *Domain Specific Ontologies*. This mapping is necessary, since each ontology domain represents the semantics of different knowledge domains. It is important to notice that these mapping concepts can be use to tie terms from the General Ontology to any Domain Specific Ontology, as well as between different Domain Specific Ontologies. This way, the Semantic Web Accessibility Framework can support different interdependent relationships between DSOs, thus affording richer Web accessibility validation scenarios.

Since the proposed integration needs require that information be passed seamlessly among the different layers, generic and domain ontology mapping is absolutely necessary. This semantic information stems from the semantic metadata description of the content and has to be mapped to the corresponding classes and properties of the relevant domain description ontology. Therefore, each domain specific ontology will provide a set of concepts to the mapping ontology to support this type of mappings.

For instance, the Web applications ontology provides properties to map *Checkpoint* instances (inherent from the Web Accessibility Standards and Guidelines Ontology) to *Application* instances (described with concepts from the Web Applications Ontology). This mapping property allows the specification of which checkpoints are valid to the particular set of technologies available for the design and development of Web applications. Another set of mapping concepts provide support for bridging *Checkpoint* instances with *Device* instances and *User* instances, thus closing the loop between characterization of devices and audiences and tailored Web accessibility assessment processes.

Rules Ontology

The last piece in the Web Accessibility Mapping dimension of the Semantic Web Accessibility Framework concerns the specification of semantic validation rules for Web accessibility. This ontology will provide the required set of rules that go beyond the syntactic analysis of Web accessibility processes, such as the description of checkpoints, users, etc. The role of this ontology is, therefore, to bridge the semantic verification gap between the Web Accessibility Description domain and the Mapping Ontologies.

We have devised this ontology as a set of rules based on SWRL (Horrocks, Patel-Schneider, Boley, Tabet, Grosof, & Dean, 2004), a rules language that affords the specification Horn-like rules with OWL predicates. The Rules Ontology can be used to reason which concepts from other ontologies (both at the instance and meta levels) and which combinations of them are satisfied by accessibility validation procedures. By setting up these rules within an inference engine, relevant accessibility rules will be reasoned out according to the information residing within the ontologies of the Semantic Web Accessibility Framework's Web Accessibility Domain dimension.

While some rules can be specified with General Ontology concepts, its use is fairly limited, as they are not tied to particular application domains. By using terms originated from Domain Specific Ontologies, and by combining them according to the semantics of existing validation processes, Web design and development processes can be augmented with more interesting verification rules that are triggered according to specific application/audience requirements. It is worth mentioning the fact that this ontology serves as an entry point for semantic validation processes. We devised it as a placeholder upon which the SWAF ontology can (and, in fact, *should*) be enriched with application-specific and technology-specific semantic Web accessibility validation rules.

As a simple example, we present the description of a set of rules for users that have been characterized as having some sort of visual disability, and how to cope with content presentation. This is one of the critical rule types that are to be supported within Web accessibility validation scenarios. User rules are defined as the set *UR* = {*U1*, *U2*, *U3*}, where each one of the rules represents a single semantic validation according to a specific user audience:

U1: if user is color blind then content of black and white images and black text are preferred.
U2: if user is partially sighted then content of audio and appropriate image is preferred.
U3: if user is totally blind then pure audio content is preferred.

The same approach can be used in other domains, e.g., for devices. When verifying if Web sites can cope with device capabilities, one can check if content can be appropriately fit into the constraints imposed by devices. Such rules domain, e.g., DR = {*D1*, *D2*, *D3*} can be defined as:

D1: if device is a mobile phone then image depth must be black and white.
D2: if device is a PDA image color contrast must be high.
D3: if device is a PC then image depth can be of any size.

When defining such semantic rules, the constructs available on the different ontologies can be used to express the concrete situations that help Web designers and developers in the tailored Web accessibility verification processes in an effective, unambiguous way. As an example, we detail how to express the Checkpoint 2.1 from the Web Accessibility Content Guidelines 1.0 (as shown in the Appendix), which state: "Ensure that all information conveyed with color is also available without color, for example from context or mark-up". When targeting to audiences composed by individuals with color blindness, this checkpoint can be expressed in SWRL as (compact syntax):

```
wao:hasGUITechnology(?APP, ?x1)
&
wao:HTML(?x1) &
wao:hasGUICharacteristics(?x1,
?x2) &
wao:characteristicName(?x2, "al-
ternativeNonColoredInformation")
&
wao:characteristicValue(?x2,
"true") &
gdco:hasDisability(?USER,
Individual("colorBlindness"))
=>
wasgo:isDefiningValidApplica
tion(Individual("WAI_check-
point2.1"), ?APP)
```

Here we are ensuring that conforming to Checkpoint 2.1, only for users (?USER) characterized by color blindness (colorBlindness), applications (?APP) encompassing HTML technologies (wao:HTML) must provide alternative non colored information. It is worth mentioning that the constructs provided by the Web Accessibility Ontology (wao:hasGUITechnology, wao:HTML, wao:hasGUICharacteristics, wao:characteristicName, and wao:characteristicValue) form the core validation rule, since they bind more general concepts (color blindness, WCAG checkpoint) to concrete concepts inherent of the application domain. All of these concepts would have to be substituted, if targeting the specific WCAG rule to other application technology (e.g., outside the scope of Web accessibility). Likewise, this type of rule can be easily adapted to device constraints such as the device domain rules described above, as follows:

```
wao:hasGUITechnology(?APP, ?x1)
&
wao:HTML(?x1) &
wao:hasGUICharacteristics(?x1,
?x2) &
wao:characteristicName(?x2, "al-
ternativeNonColoredInformation")
&
wao:characteristicValue(?x2,
"true") &
do:hasFeature(?DEV,
Individual("colorDepth1bit"))
=>
wasgo:isDefiningValidApplica
tion(Individual("WAI_check-
point2.1"), ?APP)
```

Here we attach Checkpoint 2.1 just to those devices (?DEV) that are severely constrained by color depth, e.g., just black and white displays ("colorDepth1bit"). All other atomic rules can be retained, thus showing that Web accessibility verification semantics are similar between different characterization and verification domains (e.g., color blind vs. color depth). In both rules, the test for alternativeNonColoredInformation provides the bridge towards an actual accessibility check. In this case, a software component attached to a SRWL rules processor is triggered and analyses an HTML document accordingly.

More complex rules can be built on simpler rules, thus affording sharing semantics between verification scenarios. This will afford a richer and more complete approach to the implementation and integration of SWAF into Web accessibility aware design and development tools. In the next example we rework the two rules presented above by refactoring the common set of rules in a modular way:

```
wao:hasGUITechnology(?APP, ?x1)
&
wao:HTML(?x1) &
wao:hasGUICharacteristics(?x1,
?x2) &
wao:characteristicName(?x2, “al-
ternativeNonColoredInformation”)
&
wao:characteristicValue(?x2,
“true”)
=>
ex:verifyHTMLColor(?APP)
ex:verifyHTMLColor(?APP) &
gdco:hasDisability(?USER,
Individual(“colorBlindness”))
=>
wasgo:isDefiningValidApplica
tion(Individual(“WAI_check-
point2.1”), ?APP)
ex:verifyHTMLColor(?APP) &
do:hasFeature(?DEV,
Individual(“colorDepth1bit”))
=>
wasgo:isDefiningValidApplica
tion(Individual(“WAI_check-
point2.1”), ?APP)
```

FUTURE TRENDS

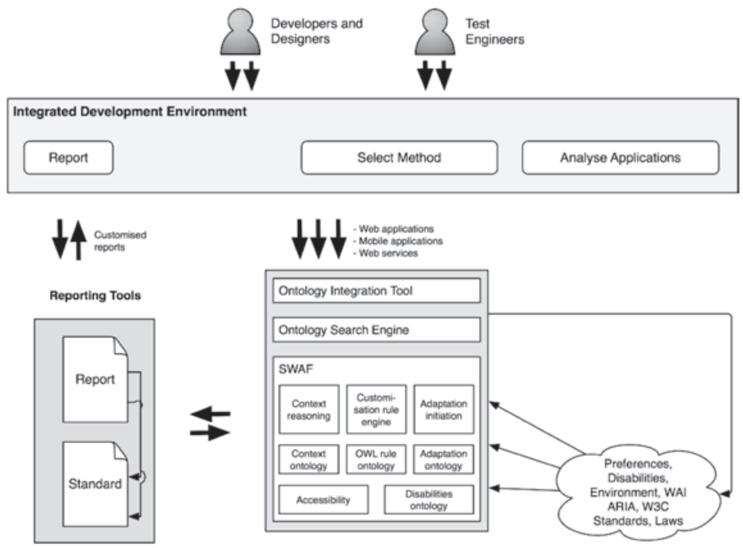
As described earlier, Web accessibility is gaining traction as time goes by, and as Web technologies mature. Web designers and developers are becoming more receptive to accessibility and inclusive design practices, to broaden the spectrum of users that can be targeted by a Web site or Web application. Still, they tend to follow blindly guidelines such as WCAG, thus lacking the perception that these do not cope with a high range of accessibility situations that are not taking into account.

Furthermore, the lack of integration of accessibility tools within the design and development process of Web sites and Web applications tend to leave accessibility assessment procedures to quality assurance tasks or, at best, usability testing tasks. This fact sets Web accessibility assessment as a patch to existing development processes, which has consequences on the adequacy of Web sites and applications.

We believe that the entry point for disrupting how Web accessibility is perceived nowadays must come from proper tool support, e.g., by means of open and free Web accessibility tools that can be plugged into existing Integrated Development Environments (e.g., NetBeans, Visual Studio, etc.) and design tools (e.g., Dreamweaver). This way, Web designers and developers will have an acute sensibility for Web accessibility issues during the design and development processes they are working in.

As an extra point, the Semantic Web Accessibility Framework detailed in this Chapter provides a fine-grained control of audiences and their characteristics, and how these can cope with existing Web accessibility guidelines. We expect that by providing this feature out-of-the-box, Web design and development teams will bring audience-modeling procedures to their development processes. This will give them more control on implementing Web sites and Web applications that are accessible and verifiable during development stages.

Figure 4. Architecture for accessibility validation services



Lastly, while Web accessibility is an important issue to take into account, it is just the starting point for providing digital services to end-users that are totally accessible and universally usable. Other domains, such as mobile phones, desktop applications, or even embedded services must also be targeted by accessibility assessment procedures during early design and development stages. The SWAF framework described in this Chapter will be extended in the future to cope with these scenarios in a very effective way, since extensibility is one of the core concerns inherent to it. On Figure 4 we present our vision of the application of SWAF in the context of application development (not just to Web sites and Web applications), and how it can be tied to Integrated Development Environments.

This architecture builds on the core technologies and concepts defined in SWAF in different fronts. First, different application domains are supported through a plug-in fashion (e.g., Web accessibility would be one of the plug-ins), based on semantic technologies such as the Mapping and Rules ontologies, or the General Characteristics and Disabilities ontology. An inference engine would provide context-reasoning features tied to particular application and technology-dependent ontologies (e.g., Web sites and HTML). We envision three extra components in this architecture that complement SWAF: (1) an ontology-oriented search engine, where developers can search for information residing in SWAF-based knowledge bases, (2) an ontology integration tool, to afford the specification of new domain-specific ontologies, and (3) a set of reporting tools centered on providing concise information about accessibility assessment procedures.

A supportive Integrated Development Environment will tie these technologies to already existing features as a complementary facet of development (e.g., similar to a debugging/helper feature). We believe that enriching IDEs with such features, as well as supportive accessibility simulation and reporting facilities will bring Web accessibility and general accessibility assessment procedures to a wide range of designers and developers, thus lowering the burden of providing accessible applications to all users without any kind of barriers.

CONCLUSION

This chapter presented SWAF, the Semantic Web Accessibility Framework, as the foundation for the semantic description of Web accessibility audiences, concepts and verification rules. This framework provides the basic constructs for the creation of Web accessibility verification engines that are capable of performing assessments tailored to specific user audiences and interaction devices. We have divided SWAF into two dimensions, Web Accessibility Descriptions (which includes general and domain specific concepts) and Web Accessibility Mapping (which affords semantic mapping and rules between concepts from the first dimension), in order to afford the extension of SWAF into different domains in the scope of Web accessibility.

Ongoing work is currently being done in several fronts in the SWAF realm, including: (1) building a comprehensive set of concept instances for user and device feature characterization, (2) providing support for guidelines and standards other than WCAG, (3) improving the Mapping Ontology to cover more situations, (4) implement a robust inference engine supporting SWAF concepts, (5) integrate this inference engine into existing Integrated Development Environments and other Web site development tools, and (6) extend the SWAF ontology to cover other application domains outside the scope of the Web.

REFERENCES

Abecker, A., Bernardi, A., Van Elst, L., Lauer, A., Maus, H., Schwarz, S., & Sintek, M. (2001). *FRODO: A framework for distributed organizational memories* (DFKI Document D-01-01, GFKI GmbH). Milestone M1: Requirements Analysis and System Architecture. Retrieved May 7, 2008, from www.dfki.uni-kl.de/frodo/ontology

ADA. (1990). *Americans with disabilities act*. Retrieved May 7, 2008, from http://www.usdoj.gov/crt/ada

ASK-IT. (n.d.). *ASK-IT ontological framework* (Public Deliverable D1.7.1). ASK-IT FP6 Integrated Project.

Chen, H., Finin, T., & Joshi, A. (2003). An ontology for context-aware pervasive computing environments. In *Proceedings of the IJCAI-03 Workshop on Ontologies and Distributed Systems (ODS 2003)*, Acapulco, Mexico.

Chen, H., Perich, F., Finin, T., & Joshi, A. (2004). SOUPA: Standard ontology for ubiquitous and pervasive applications. In *Proceedings of the First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (Mobiquitous 2004)*, Boston, MA, USA.

Chisholm, W., Vanderheiden, G., & Jacobs, I. (1999). *Web content accessibility guidelines 1.0*. Retrieved May 6, 2008, from http://www.w3.org/TR/WAI-WEBCONTENT

DDA. (1995). *Disability discrimination act*. Retrieved May 7, 2008, from http://www.disability.gov.uk/dda/index.html

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Source, Knowledge, Acquisition, 5*(2).

Habra, N., Abran, A., & Lopez, M. (2004). *Towards a framework for measurement lifecycle* (Tech Rep TR37/04). University of Namur.

Heckmann, D. (2005). *Ubiquitous user modeling*. Unpublished doctoral dissertation, Department of Computer Science, Saarland University.

HFES. (n.d.). *US human factors and ergonomics society*. Retrieved May 7, 2008, from http://www.hfes.org

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004). *SWRL: A Semantic Web rule language combining OWL and RuleML*. Retrieved June 16, 2008, from http://www.w3.org/Submission/SWRL/

ICF. (n.d.) *International classification of functioning, disability and health. World Health Organization*. Retrieved May 7, 2008, from http://www.who.int/classifications/icf/site/index.cfm

ISO 16071. (2003). *ISO/TS 16071: 2003 ergonomics of human–system interaction*. International Standards Organisation.

Jacobs, I., Gunderson, J., & Hansen, E. (2002). *User agent accessibility guidelines 1.0*. Retrieved May 6, 2008, from http://www.w3.org/TR/WAI-USERAGENT/

Kasunic, M. (2006). *The state of software measurement practice: Results of 2006 survey* (Tech. Rep. CMU/SEI-2006-TR-009, ESC-TR-2006-009). Software Engineering Institute.

Kulik, P. (2000). *Software metrics: State of the art*. Retrieved May 7, 2008, from http://www.klci.com

Lopes, R., & Carriço, L. (2008). The impact of accessibility assessment in macro scale universal usability studies of the Web. In *Proceedings of the 5th International Cross-Disciplinary Conference on Web Accessibility*, Beijing, China.

Masuwa-Morgana, K. R., & Burrell, P. (2004). Justification of the need for an ontology for accessibility requirements (theoretic framework). *Interacting with Computers*, *16*(3), 523–555. doi:10.1016/j.intcom.2004.04.001

McGarry, J., Card, D., Jones, C., Layman, B., Clark, E., Dean, J., & Hall, F. (2002). *Practical software measurement. Objective information for decision makers*. New York: Addison-Wesley.

Neely, A. (1998). *Measuring business performance. Why, what and how?* London: Profile Books Ltd.

Obrenovic, Z., Abascal, J., & Starcevic, D. (2007). Universal accessibility as a multimodal design issue. *Communications of the ACM*, *50*(5), 83–88. doi:10.1145/1230819.1241668

Section 508. (1998). *Civil rights division home page*. Retrieved May 7, 2008, from http://www.usdoj.gov/crt/508/508home.html

SENDA. (2001). *Special educational needs and disability act*. Retrieved May 7, 2008, from http://www.dcs-gb.net/part4dda.html

Shneiderman, B. (2000). Universal usability. *Communications of the ACM*, *43*(5), 84–91. doi:10.1145/332833.332843

Strang, T., Linnhoff-Popien, C., & Fank, K. (2003). CoOL: A context ontology language to enable contextual interoperability. In *Proceedings of the 4th IFIP International Conference on Distributed Applications and Interoperabile Systems (DAIS2003)*, Paris, France. *TRACE*. (n.d.). *Designing more usable websites*. Retrieved May 6, 2008, from http://trace.wisc.edu/world/web

Treviranus, J., Richards, J., Jacobs, I., & McCathieNevile, C. (1999). *Authoring tool accessibility guidelines 1.0*. Retrieved May 6, 2008, from http://www.w3.org/wai/au/wai-autools-19990621/wai-autools.html

Uschold, M., King, M., Moralee, S., & Zorgios, Y. (1998). The enterprise ontology. *The Knowledge Engineering Review, Special Issue on Putting Ontologies to Use, 13*.

Van Heijst, G., Schreiber, A. T., & Wielinga, B. J. (2000). *Understanding, building, and using ontologies: A commentary to using explicit ontologies in KBS development*. LADSEB-CNR, National Research Council.

Varkoi, T. (1999). Development of measurement programs to support process improvement in small software companies. In *Proceedings of the FESMA'99 Conference*, Amsterdam, The Netherlands (pp. 141-149).

Vigo, M., Kobsa, A., Arrue, M., & Abascal, J. (2007). Quantitative metrics for measuring Web accessibility. In *Proceedings of the 4th International Cross-Disciplinary Conference on Web Accessibility*, Banff, Canada. *W3C*. (n.d.). *World Wide Web Consortium – Web standards*. Retrieved May 6, 2008, from http://www.w3.org

WAI. (n.d.). *Web accessibility initiative*. Retrieved May 6, 2008, from http://www.w3.org/WAI

Wang, H. X., Zhang, Q. D., Gu, T., & Pung, H. K. (2004). Ontology based context modeling and reasoning using OWL. In *Proceedings of 2nd IEEE Conference on Pervasive Computing and Communications Workshops (PerCom 2004 Workshops)*, Orlando, FL, USA.

Wooldridge, M., Jennings, N., & Kinny, D. (2000). The Gaia methodology for agent-oriented analysis and design. *Journal of Autonomous Agents and Multi-Agent Systems*, *3*(3), 285–312. doi:10.1023/A:1010071910869

KEY TERMS AND DEFINITIONS

Accessibility: The *ability to access*. Often tied to people with disabilities (e.g., total blindness), accessibility thrives to break the barriers to information access. We follow the strict sense of accessibility by embracing any situation where the ability to access information can be disrupted by device or even surrounding environment constraints.

Accessibility Guidelines: A set of best practices that must be followed by designers and developers when implementing software solutions (e.g., Web site) that will help on providing accessible information. By being guidelines, it should not be assumed that content is accessible just by following them.

Checkpoint: A concrete verification task that materializes a (part of a) guideline. Checkpoints can be fully automated if application technology provides corresponding support (e.g., verifying if all images have associated textual captions).

Integrated Development Environment: A computer application used by developers that provides several features to ease the task of developing applications, such as text editor, compiler, automation features, etc.

Universal Usability: a research field that studies the adequacy of user interfaces and information to all users, regardless of their characteristics, knowledge, or mean of interaction (Shneiderman, 2000).

Usability: A research field that studies how adequate user interfaces are to users, how easily can they learn to perform tasks, and what are their levels of satisfaction when interacting with user interfaces.

User Interface: The "visible" side of an application, where users can acquire and interact with information.

Web Accessibility: The subfield of accessibility that is targeted to the specific technologies and architecture that compose the World Wide Web. This includes technologies such as HTML, CSS and JavaScript, as well as the HTTP protocol.

APPENDIX

Table 1. WAI Web Content Accessibility Priority 1 Checkpoints

Cp	Description
Cp1.1	Provide a text equivalent for every non-text element (e.g., via "alt", "longdesc", or in element content).
Cp2.1	Ensure that all information conveyed with colour is also available without colour, for example from context or mark-up.
Cp4.1	Clearly identify changes in the natural language of a document's text and any text equivalents (e.g., captions).
Cp6.1	Organize documents so they may be read without style sheets.
Cp6.2	Ensure that equivalents for dynamic content are updated when the dynamic con-tent changes.
Cp7.1	Until user agents allow users to control flickering, avoid causing the screen to flicker.
Cp14.1	Use the clearest and simplest language appropriate for a site's content.
Cp1.2	Provide redundant text links for each active region of a server-side image map.
Cp9.1	Provide client-side image maps instead of server-side image maps except where the regions cannot be defined with an available geometric shape.
Cp5.1	For data tables, identify row and column headers.
Cp5.2	For data tables that have two or more logical levels of row or column headers, use mark-up to associate data cells and header cells.
Cp12.1	Title each frame to facilitate frame identification and navigation.
Cp6.3	Ensure that pages are usable when scripts, applets, or other programmatic objects are turned off or not supported. If this is not possible, provide equivalent information on an alternative accessible page.
Cp1.3	Until user agents can automatically read aloud the text equivalent of a visual track, provide an auditory description of the important information of the visual track of a multimedia presentation.
Cp1.4	For any time-based multimedia presentation (e.g., a movie or animation), synchronize equivalent alternatives (e.g., captions or auditory descriptions of the visual track) with the presentation.
Cp11.4	If, after best efforts, you cannot create an accessible page, provide a link to an alternative page that uses W3C technologies, is accessible, has equivalent in-formation (or functionality), and is updated as often as the inaccessible (original) page.

This work was previously published in Handbook of Research on Social Dimensions of Semantic Technologies and Web Services, edited by M.M. Cruz-Cunha, E.F. Oliveira, A.J. Tavares, and L.G. Ferreira, pp. 49-67, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 2.16
Building Semantic Web Portals with a Model-Driven Design Approach

Marco Brambilla
Politecnico di Milano, Italy

Federico M. Facca
Leopold-Franzens-Universität Innsbruck, Austria

ABSTRACT

This chapter presents an extension to Web application conceptual models toward Semantic Web. Conceptual models and model-driven methodologies are widely applied to the development of Web applications because of the advantages they grant in terms of productivity and quality of the outcome. Although some of these approaches are meant to address Semantic Web applications too, they do not fully exploit the whole potential deriving from interaction with ontological data sources and from semantic annotations. The authors claim that Semantic Web applications represent an emerging category of software artifacts, with peculiar characteristics and software structures, and hence need some specific methods and primitives for achieving good design results. In particular the contribution presented in this chapter is an extension of the WebML modeling framework that fulfils most of the design requirements emerging in the new area of Semantic Web. The authors generalize the development process to cover Semantic Web needs and devise a set of new primitives for ontology importing and querying. The chapter also presents a comparison of the proposed approach with the most relevant existing proposals and positioned with respect to the background and adopted technologies.

INTRODUCTION AND MOTIVATION

Modern Web applications comprise distributed data integration, remote service interaction, and management of workflow activities, possibly spawned on different peers. In this scenario, a wider attention to the semantics of data and applications is mandatory to allow effective design and evolution of complex systems. Indeed, if semantics of data and applications is known, their integration becomes more feasible. Moreover, explicit semantic annotation of Web applications can facilitate content search and access, and foster a future generation of Web clients that exploit the semantic information to provide better browsing capabilities to customers.

The Semantic Web aims at bringing formal “semantics” to the human-readable information so as to make it machine-readable and allow better and easier automatic integration between different Web applications. To address this challenge, many semantic description languages arose, like RDF, OWL and WSML; some of them are currently W3C Recommendations. All these languages allow to formally model knowledge by means of ontologies: the resulting formal models are the starting point to enable easy information exchange and integration between machines.

These languages are suitable for reasoning and inference, i.e., to deduct more information from the model by applying logical expressions. This makes the modeling task easier since not all the knowledge has to be modeled. These languages are supported by a wide range of tools and APIs, that cover design of knowledge (e.g., Protégé (Noy et al., 2001) and OntoEdit (Sure et al., 2002)), provide storing facilities (e.g., Sesame (Aduna, 2007) and Jena (HP, 2007)), and offer reasoning on the data (e.g., Racer (Racer Systems, 2007) and Pellet (Sirin et al.,2007)). Based on these modeling languages, a set of querying languages have been devised too; among them, we can mention TRIPLE (Sintek & Decker, 2002) and SPARQL (W3C, 2007), a W3C candidate recommendation.

Unfortunately, although the theoretical bases and some technological solutions are already in place for Semantic Web support, the techniques and methodologies for Semantic Web application design are still rather rough. This leads to high costs of implementation for Semantic Web features, even if embedded within traditional Web applications. These extra costs are related not only to the design of the architecture and deployment of the Semantic platforms, but also to the repetitive and continuous task of semantic annotation of contents and application pages.

We claim that conceptual modeling and model-driven development can increase dramatically the efficiency and efficacy of the design and implementation of such applications, by offering tools and methodologies to the designer for specifying semantically-rich Web applications.

The model-driven approach to software development has been proven valid in several application fields and is currently one of the best practices of the software engineering discipline. Developing a Semantic Web application, as with any other kind of software system, is a complex achievement that requires the ability to master a broad spectrum of tasks, jointly performed by a number of persons with different skills for a long timeline. Software engineering and Web engineering (Ceri et al., 2002) demonstrated that following a well organized development process, centered on the appropriate modeling concepts, is essential to overcome the complexity inherent to such kind of developments.

This chapter aims at demonstrating how model-driven design can impact on specification, design, and implementation of Semantic Web portals as well. In the proposed approach, we leverage a conceptual modeling approach for visually designing the Web application domain model and hypertext model. Conceptual modeling works at higher abstraction levels with respect to direct implementation design, thus allowing to specify the application design with a top-down philosophy. The first design steps aim at describing the platform-independent domain and hypertext models, disregarding the low-level details. Further refinements can take more and more into account such details, finally leading to the expected outcome of the work. These approaches take great advantage from the adoption of C.A.S.E. (Computer Aided Software Engineering) tools, which provide facilities for designing and implementing applications according to specific methods and notations. C.A.S.E. tools sometimes include partially or completely automatic code generation too.

As a concrete example of modeling language, we pick WebML (Web Modeling Language) (Ceri et al., 2002), a well known methodology and set of metamodels in the Web engineering environment, and its companion C.A.S.E. tool WebRatio (WebModels s.r.l., 2007).

To discuss our approach and related background technologies, a running example will be used throughout the chapter. We discuss a realistic scenario based on the reuse of existing ontologies available on the Internet that can be easily exploited to create new Semantic Web Portals. In particular, we will consider two ontologies for the musical domain to build a Web application offering access to music contents, considering also users profile information. We combine the MusicBrainz ontology (http://www.musicbrainz.org) for the music domain information; the MusicMoz (http://www.musicmoz.org/) hierarchy to classify music genres; the RDF Site Summary (RSS-DEV Working Group, 2000) for music news; and the Friend Of A Friend (Foaf) ontology (Miller & Brickley, 2000) to describe for user's profiles and relationships among them. The case study application is similar to other existing Semantic Web applications (e.g., http://Foafing-the-music.iua.upf.edu/, http://squiggle.cefriel.it/music/), that provide personalized access to the contents exploiting distributed semantic information. The presented application, although rather simple because of space reasons, can be considered a full-fledged Semantic Web Portal since it aggregates different sources of information spanned across the Internet and presents them in a structured and user-friendly manner.

Structure of the Chapter

This chapter is organized as follows. The Section *Background* contains a description of the background technologies, metamodels and methodologies for Web engineering and Semantic Web applications (i.e., WebML, OWL, RDF, and so on). The Section *Requirements for Semantic Web engineering* presents a set of requirements that Web Engineering approaches for modeling Semantic Web applications must comply with, according to our analysis; The Section *Modeling Semantic Web application with WebML* proposes our extension of a metamodel for the design of Semantic Web portals (i.e., new primitives and metamodels within the WebML framework); in Section *Case Study: A Music Semantic Portal* we present a realistic Semantic Web portal scenario designed and implemented with the proposed methodology. The Section *Implementation Experience* discusses the architectural and implementation aspects of the proposed approach. Finally Section *Conclusion* and *Future Research Directions* propose a summary of the lesson learned from the research presented and possible research direction starting from the work reported in the chapter.

BACKGROUND

This section presents an overview of the background technologies and relevant work in the field of engineering Semantic Web applications. In particular we first introduce an overview of the most relevant languages adopted for defining and querying ontologies; then we present a review of relevant Web engineering methodologies that cover Semantic Web technologies; a subsequent paragraph is dedicated to introduce WebML, the Web application modelling language that we extend for supporting Semantic Web features (extensions are discussed in Section *Modeling Semantic Web application with WebML*).

Languages for the Semantic Web

Several ontology languages have been developed during the last few years, and they will surely be exploited in the context of the Semantic Web. Some of them are based on XML syntax, such as Ontology Exchange Language (XOL), SHOE (which was previously based on HTML), and Ontology Markup Language (OML), whereas Resource Description Framework (RDF) (W3C, 2004b) and RDF Schema (W3C, 2004c) are languages created by World Wide Web Consortium (W3C) working groups. Later, two additional languages have been built on top of RDF(S) - the union of RDF and RDF Schema - to improve its features: Ontology Inference Layer (OIL) and DAML+OIL. DAML+OIL was submitted to W3C as starting proposal for a new ontology language: OWL (W3C, 2004a). Recently, a new emerging initiative, the Web Service Modeling Ontology (WSMO), introduced a new ontology language called WSML (Fensel et al., 2006). A more detailed discussion can be found in Chapter *Ontology Languages for the Semantic Web: An Overview*. The ontologies adopted in our case study are a mix of RDF(S) and OWL, that are shortly presented in the next paragraphs.

Resource Description Framework and RDF Schema

RDF (W3C, 2004b), developed by the W3C for describing Web resources, allows the specification of the semantics of data based on XML in a standardized, interoperable manner. It also provides mechanisms to explicitly represent services, processes, and business models, while allowing recognition of non explicit information. The RDF data-model is based on *subject, predicate, object* triples, so called RDF statements, to formalize meta-data. RDF is domain independent in that no assumptions about a particular domain of discourse are made. It is up to the users to define their own ontologies for the user's domain in an ontology definition language such as RDF Schema (RDFS) (W3C, 2004c).

RDF statements are based on the concept of *resources*, which can be used in the different roles of the statement. A resource can be everything; a book, a person, a Web page, a CD, a track on a CD, and so on. Every resource is identified by a Uniform Resource Identifier (URI). In the case of a Web page, the URI can be the Unified Resource Locator (URL) of the page. The URI does not necessarily enable the access via the Web to the resource; it simply unambiguously identifies the resource.

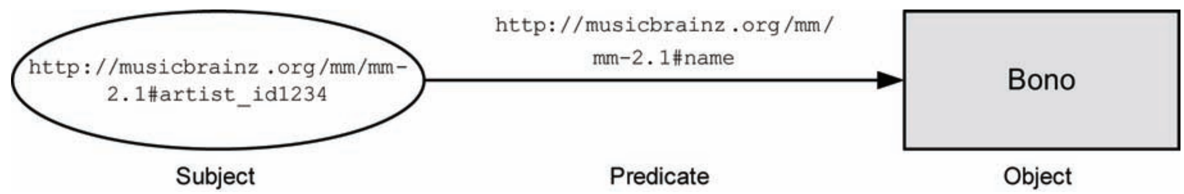
The *subject* of a statement is the resource we want to make a statement about. For example if we want to make a statement about a music artist, we need to identify it with a URI (e.g. http://musicbrainz.org/mm/mm-2.1#artist_id1234). The predicate defines the kind of information we want to express about the subject. For example if we want to make a statement about the name of the Artist we can use the URI http://musicbrainz.org/mm/mm-2.1#name that references the property defined in the ontology. The *object* defines the value of the predicate, for example in our case we want to state that the name of artist is "Bono". The object can be a literal, like in this example, or another resource represented by the object's URI.The statement presented can be written in triple notation as[a]:<http://musicbrainz.org/mm/mm-2.1#artist_id1234> <http://musicbrainz.org/mm/mm-2.1#name> "Bono".RDF supports also an XML syntax for writing and exchanging RDF graphs, called RDF/XML. The RDF/XML representation of the previous statement is:<rdf:RDF xmlns:mm="http://musicbrainz.org/mm/mm-2.1#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"><rdf:Description rdf:about="http://musicbrainz.org/mm/mm-2.1#artist_id1234"><mm:name>Bono</mm:name></rdf:Description></rdf:RDF>

A statement can be represented as a graph too, as depicted in Figure 1.

The RDF data model does not provide mechanisms for defining the relationships between properties (attributes) and resources - this is the role of RDFS (W3C, 2004c). RDFS offers primitives for defining knowledge models that are closer to frame-based approaches[b].

In particular RDFS introduces "meta"-classes that are used to define resource types: rdfs:Class

Figure 1. The graphical representation of a RDF statement.



to define RDF classes (concepts); rdfs:Literal to define literal values such as strings and integers; rdfs:Property to define properties. The full RDFS vocabulary and its description can be found in (W3C, 2004c). The following fragment defines the concept artist as previously introduced:<rdfs:Class rdf:about="http://musicbrainz.org/mm/mm-2.1#Artist"rdfs:label="Artist"><rdfs:subClassOfrdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/></rdfs:Class><rdf:Propertyrdf:about="http://musicbrainz.org/mm/mm-2.1#name" rdfs:label="name"><rdfs:domain rdf:resource="http://musicbrainz.org/mm/mm-2.1#Artist"/><rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/></rdf:Property>

RDF and RDFS are not rich enough to completely describe an ontology as they still miss some important concepts within their Description Logic.

Ontology Web Language

OWL (W3C, 2004a) is designed for applications that need to process the content of information instead of just presenting information to humans. Furthermore OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDFS) by providing additional vocabulary along with a formal semantics.

OWL provides a number of additional modelling primitives that increase the expressiveness compared to RDFS and solve some shortcomings of RDFS:

- *Cardinality restrictions*. RDFS does not provide any means to restrict the number of distinct values a property may or must take.
- *Disjoint classes*. RDFS does not provide primitives to declare two classes to be disjoint.
- *Set combination of classes*. RDFS does not allow to define new classes by building the union, intersection, or complement of other classes.
- *Special characteristics of properties*. In RDFS is not possible to define that a property is transitive (e.g. greater than), unique (e.g. is mother of), or the inverse of another property.

The following OWL fragment represents a OWL properties and its inverse. The inverseOf element specifies that playedBy is the inverse of plays.<!-- Artist plays a Track--><owl:ObjectProperty rdf:ID="http://musicbrainz.org/mm/mm-2.1#plays"><rdfs:domain rdf:resource="http://musicbrainz.org/mm/mm-2.1#Artist"/><rdfs:range rdf:resource="http://musicbrainz.org/mm/mm-2.1#Track"/></owl:ObjectProperty><!-- Track played by an Artist --><owl:ObjectProperty rdf:ID="http://musicbrainz.org/mm/mm-2.1#playedBy"><rdfs:domain rdf:resource="http://musicbrainz.org/mm/mm-2.1#Track"/><rdfs:range rdf:resource="http://musicbrainz.org/mm/mm-2.1#Artist"/><owl:**inverseOf** rdf:resource="http://musicbrainz.org/mm/mm-

2.1#plays" /></owl:ObjectProperty>OWL has three increasingly-expressive sublanguages: OWL Lite, OWL DL, and OWL Full.

- *OWL Lite* supports those users primarily needing a classification hierarchy and simple constraint features.
- *OWL DL* includes all OWL language constructs with restrictions such as type separation (a class can not also be an individual or property; a property can not also be an individual or class). OWL DL is so named due to its correspondence with *description logics*, a field of research that has studied a particular decidable fragment of first order logic.
- *OWL Full* gives the maximum expressiveness and the syntactic freedom of RDF with no computational guarantees. For example, in OWL Full a class can be treated simultaneously as a collection of individuals and as an individual in its own right. It is unlikely that any reasoning software will be able to support every feature of OWL Full. For instance, general inference in OWL Full is clearly undecidable as OWL Full does not include restrictions on the use of transitive properties which are required in order to maintain decidability.

Query Languages for the Semantic Web

With the diffusion of RDF data sources, the problem of querying them in a easy way caused the flourishing of many different query languages for RDF based on different principles: some are inspired by SQL, like SPARQL (W3C, 2007) and RQL (Karvounarakis, 2004), others are based on different principles like graph patterns - e.g. TRIPLE (Sintek & Decker, 2002) - and reactive rules - e.g. Algae (Prud'hommeaux, 2004). In depth comparison of the most relevant RDF query languages can be found in (Haase et al., 2004; Furche et al., 2006); while a broader discussion of Web and Semantic Web query languages is presented in (Bailey et al., 2005). In the next paragraph we give some more details on SPARQL since it is probably the most adopted query language for RDF and it is the language adopted by our solution.

SPARQL

SPARQL (W3C, 2007) is a query language that has already reached candidate recommendation status at the W3C, and is on a good way to become the W3C recommendation for RDF querying. Querying RDF data with languages in the SPARQL family amounts to matching graph patterns that are given as sets of triples of subjects, predicates and objects. The triple syntax adopted by SPARQL is based on Turtle (Beckett, 2004). Solutions to SPARQL queries are given in the form of result sets: each result set contains a set of mappings from the variables occurring within the query to nodes of the queried data. For instance, the query that extracts the name and the nationality of all the Artists from the ontology previously presented can be written as:BASE <http://musicbrainz.org/mm/mm-2.1#>PREFIX mm: <http://musicbrainz.org/mm/mm-2.1#>PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>SELECT ?name ?nationalityWHERE{ ?id rdf:type mm:Artist .?id mm:name ?name .OPTIONAL{ ?id mm:nationality ?nationality .}}

The WHERE clause specifies the graph pattern to selected data to be mapped to the variables; variables are identified by either ? or $ prefix. The OPTIONAL clause specifies that a certain triple is optional: it is not compulsory that all the returned results contains the triples stated has optional. If the OPTIONAL clause is removed, the results will include only the triples that contain the whole graphical pattern requested. The FROM clause can be used to specifies the URL (or some other identifier) of the data to be queried.

SPARQL includes also other three types of query: the CONSTRUCT query clauses, that create new RDF graphs with data from the RDF graph queried; the DESCRIBE query that return RDF "descriptions" of the resources matching the query part (e.g., the query DESCRIBE mm:u2 returns a RDF graph representing the mm:u2 resource); and the ASK query that return true or false according to the fact that the specified pattern has a solution or not (e.g., the query ASK mm:u2 rdf:type mm:Artist returns true if the resource mm:u2 is an instance of the class mm:Artist).

Methodologies to Design Semantic Web Applications

While design methodologies for traditional Web applications offer rather mature and established solutions methodologies for developing Semantic Web applications are still in an early development phase. Realizing the benefits of the Semantic Web platform (e.g., interoperability, inference capabilities, increased reuse of the design artifacts, etc.) traditional design methodologies are now focusing on designing Semantic Web applications: e.g., OOHDM (Schwabe & Rossi 1998) evolved in SHDM (Lima & Schwabe, 2003). New methodologies like XWMF (Klapsing et al., 2001), OntoWebber (Jin et al., 2001) and Hera (Vdovjak et al., 2003) were specifically designed by considering the Semantic Web peculiarities. Among them, the most complete are Hera and SHDM that are shortly described in the next paragraph.

The Semantic Hypermedia Design Method

SHDM (Lima & Schwabe, 2003) is an ontology-based design methodology. It extends the expressive power of OOHDM (Schwabe & Rossi, 1998) by defining ontologies for each of the OOHDM models. These ontologies are specified in OWL, a more expressive language than RDFS. In the same way as OOHDM, SHDM identifies four different phases: conceptual design, navigation design, abstract interface design, and implementation.

The conceptual design builds the conceptual class schema for the application domain. This schema is described in UML extended with a few new characteristics like the ability to specialize relations. The UML diagram is mapped to an OWL model according to some heuristics rules.

The navigation design defines the navigational class schema and the navigational context schema. The main navigational primitives are navigational classes (nodes), navigational contexts, and access structures. In the same way as for the conceptual class schema, one can specialize navigational relations. The mappings between the conceptual schema and navigational class schema are defined using RQL. The navigation context allows the description of sets of navigational objects.

The abstract interface design defines the visual aspect of the deployed application by means of the abstract widget ontology and concrete widget ontology. The implementation phase produces a Semantic Web application based on the previous SHDM specifications.

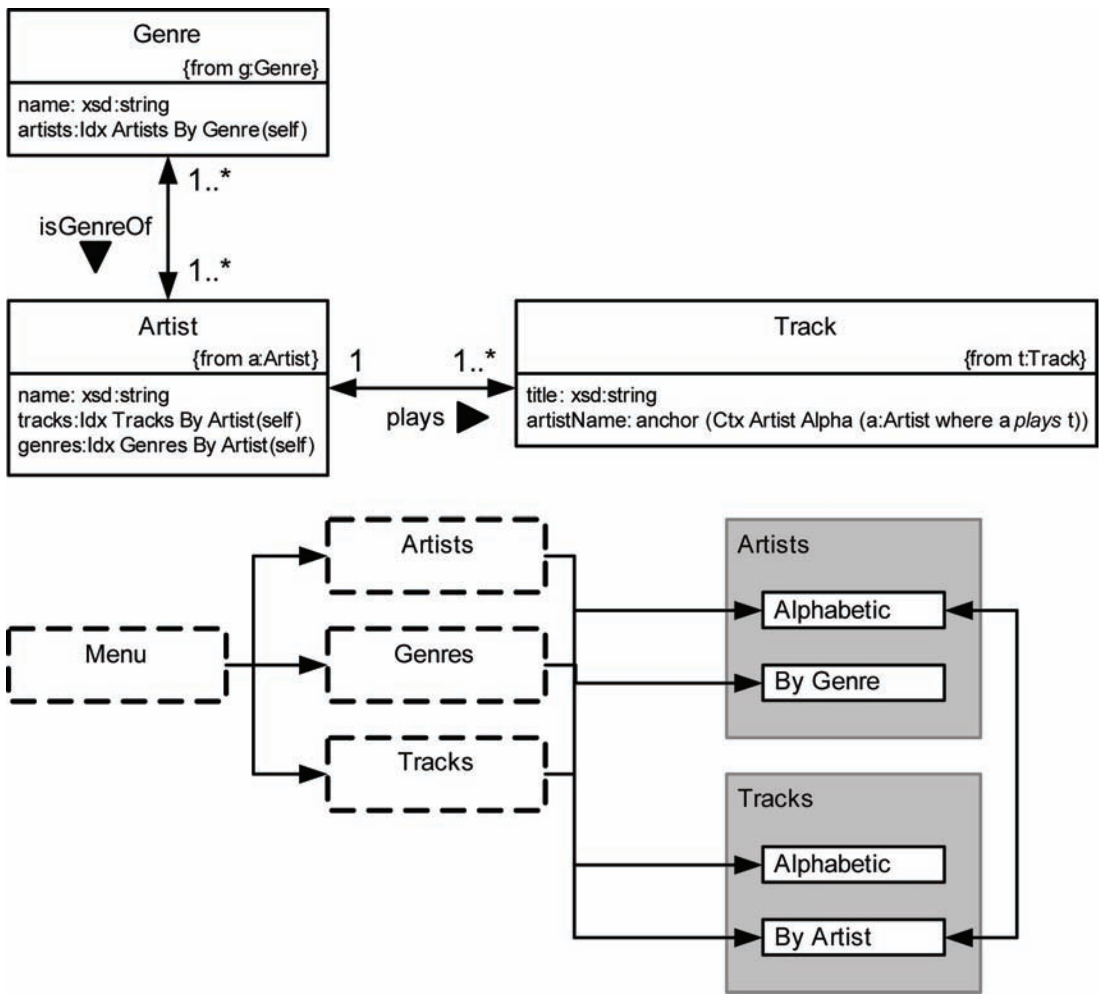
Figure 2 presents the Navigational Class Schema (top part) and the Navigational Context Schema (bottom part) of a simple Web application based on the musicbrainz and musicmoz ontology.

HERA

The Hera Methodology (Vdovjak et al., 2003) is a model driven methodology for designing and developing Web applications using Semantic Web technologies. It is developed from a database perspective; the conceptual modelling (similar to the ER modelling) and the querying of the different Hera models are important issues in the proposed methodology.

The Hera methodology has three main layers: *the conceptual model layer (CM), the application model layer (AM)*, and the *presentation model*

Figure 2. SHDM navigational design for a Web application based on the case study ontologies: Navigational Class Schema (top part) and Navigational Context Schema (bottom part).

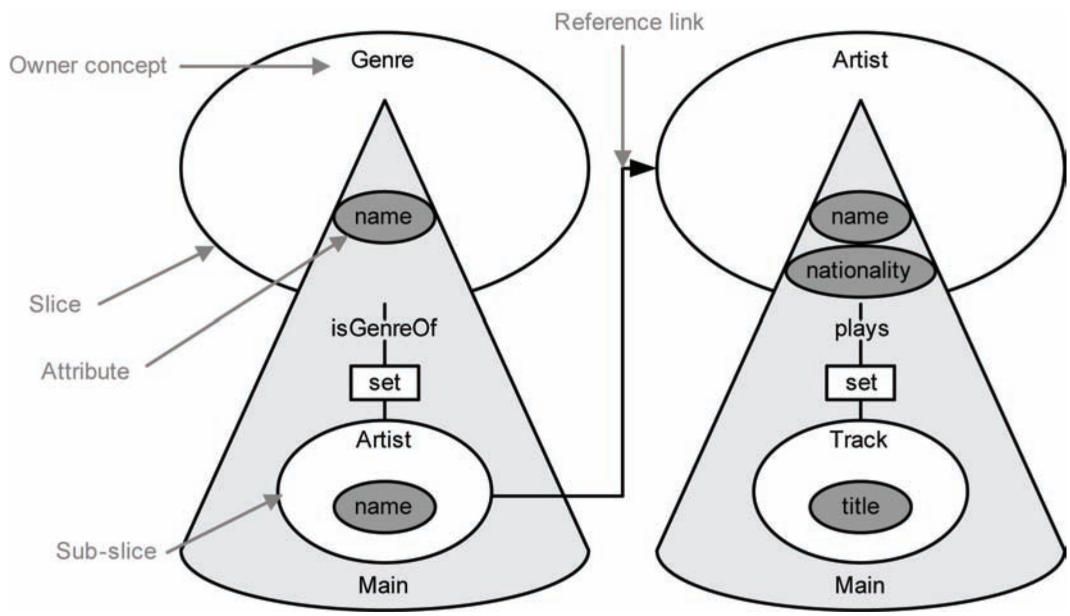


layer (PM). The first layer describes data content used for generation of hypermedia presentations and makes available data coming from different, possibly heterogeneous data sources; the integration of different data sources is possible thanks to a data integration model (a sub-layer that allows decoupling between the conceptual model and the data sources). The second layer describes the navigation structure and functionality; finally the PM layer describes spatial layout and rendering of hypermedia presentations. An orthogonal layer, called adaptation layer, captures adaptation issues in all the above layers.

By distinguishing these different layers, Hera differentiates at design level between the semantic aspects, the navigational aspects, and the interface aspects of a Semantic Web application.

The core components of the application model layer are called slices. They are associated with a concept from the CM model and may contain properties of the concept or other slices. Different not nested slices may be interconnected by slice relationships, that can be classified in aggregation relationships (e.g., index, tour, indexed guided tour, etc.) or reference relationships (i.e., links with an anchor specified). Slices represent views

Figure 3. An Hera application model based on the case study ontologies



over the conceptual model and are translated to RQL queries.

All the HERA models are, as matter of fact, ontologies, and hence a Web application model is a collection of instances of the different model ontologies. Therefore, it is possible to define a fragment of an application model as subclass of another fragment (this applies also the other two models).

Figure 3 presents a fragment of an application model based on the musicbrainz and musicmoz ontology: the first slice on the left represents a page that contains a genre and the set of artists linked to that genre; each artist is linked with a reference link to the artist slice that collects information about the artist an the set of tracks played by the artist.

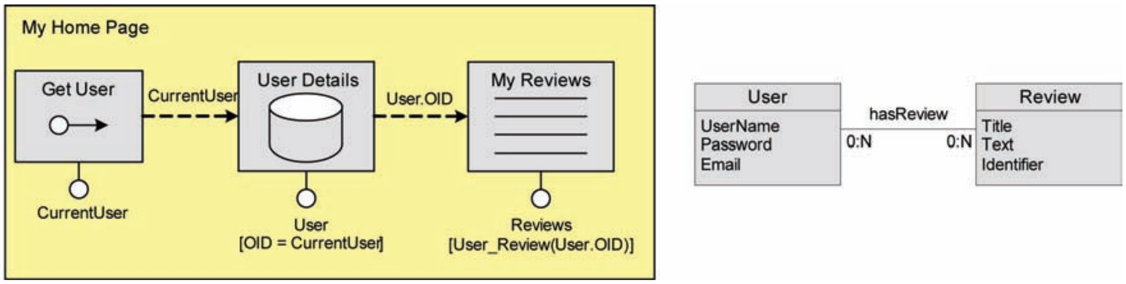
WEBML

Our approach to Semantic Portals specification is based on WebML (Ceri et al., 2002). In this paragraph we introduce the basic features of WebML; the extension we introduced to model Semantic Web applications is presented later in this chapter. WebML uses conceptual modeling techniques for describing web applications. The WebML design methodology comprises three main phases: data design, hypertext design, and implementation. It provides a visual notation and a XML serialization for the proposed models.

For specifying the data underlying the application, WebML exploits an extended version of the *Entity-Relationship* model, which consists of entities (classes of data elements), and relationships (semantic connections between entities).

The hypertext design defines the navigational structure of the application. WebML also allows designers to describe hypertexts, called *site views*, for publishing and managing content. A site view is a piece of hypertext, which can be browsed by a particular class of users. Multiple site views can be defined for the same application. Site views are then composed of *areas* and *pages*. Areas and pages can be nested in areas. Finally pages are the containers of elementary pieces of content, called *content units*, typically publishing data

Figure 4. An example Web page modeled in WebML (left) and its underlying data model (right)



retrieved from the database, whose schema is expressed through the data model. In particular, WebML primitives for content publishing denote alternative ways for displaying one or more entity instances: e.g., the *data unit* publishes a single instance of data, the *index unit* a list of data. Unit specification requires the definition of a source and a selector: the source is the name of the entity from which the unit's content is extracted; the selector is a condition, used for retrieving the actual objects of the source entity that contribute to the unit's content.

Between units/pages one can define hypertextual links as oriented connections. WebML distinguishes several types of links: *navigational*, *automatic*, and *transport* links. These links can carry information from the source to the destination. The information is stored in the link parameters. The navigational links require user intervention, while both automatic and transport links are traversed without user intervention: for automatic links once the source is presented also the associated destination is shown; the transport links do not define navigation and are solely used to transport information.

WebML also supports the specification of content management operations. They allow creating, deleting or modifying an instance of an entity (respectively through the *create*, *delete*, and *modify* units), or adding or dropping a relationship between two entity instances (respectively through the *connect* and *disconnect* units).

Figure 4 presents an example of WebML hypertext model and its underlying E-R model. It shows a single page, which allows the user to see his reviews about music albums. In the *My Home Page* a *Get* unit retrieves the identifier of the current user from the session parameter *CurrentUser* and provides it in input to the following *Data* unit, which publishes the user's personal profile data. The data to be published are restricted by means of a selector condition, specified below the unit. The user identifier is further propagated to the *Index* unit *My Reviews* by means of a transport link. The index unit shows the list of reviews related to the received user identifier.

WebML-based development is supported by the WebRatio CASE tool (WebModels s.r.l., 2007), which offers a visual environment for designing the WebML conceptual schemas, storing them in XML format, and automatically generates the running code (through XSLT model transformations), which is deployed as pure J2EE code.

REQUIREMENTS FOR SEMANTIC WEB ENGINEERING

While Web Engineering managed to bring software engineering practices to the Web development area, no discipline has addressed the peculiar needs of Semantic Web application design. A good methodology for the design of Semantic Web applications must provide additional facilities to the developer, specifically addressing the new needs. To understand them, we define the features of these applications, and then infer the characteristics of the methodologies and models that must be provided to the designer. To collect the requirements that a Semantic Web application should comply with, we analyzed some existing Semantic Web Portals (e.g., http://www.mindswap.org, http://Foafing-the-music.iua.upf.edu, and http://ontoworld.org) and we extracted the following set of needs:

- **Support of semantic languages:** Semantic Web applications should be aware of and support (i.e., be able to query and manage) different Semantic Languages and metamodels (RDFS, OWL, WSML, …).
- **Semantic application models:** Semantic Web applications should be designed and specified by means conceptual models that include and support semantic descriptions.
- **Flexible integration:** Semantic Web applications should embrace the philosophy of flexibility and heterogeneity integration of Semantic Web.
- **Classes and instances access and queries:** Both domain ontology classes and instances should be easily and seamlessly accessible by Semantic Web applications, through appropriate querying primitives, including data instance and structure queries.
- **Inference and verification:** Ontology-based web applications should exploit available inferring systems on ontological data, both for semantic queries and verification of data.
- **Semantic data sources:** A Semantic Web application relies on semantic data (e.g., ontologies) that offer a machine understandable data description that may be used to populate and generate Web pages and also to provide semantic annotations.
- **Importing and reuse of ontologies:** Semantic Web applications shall allow to: (1) import new (possibly distributed) data conforming to the Web application ontology; (2) to seamlessly integrate new ontologies, not fitting the default ontology; and (3) to reuse existing and shared ontologies.

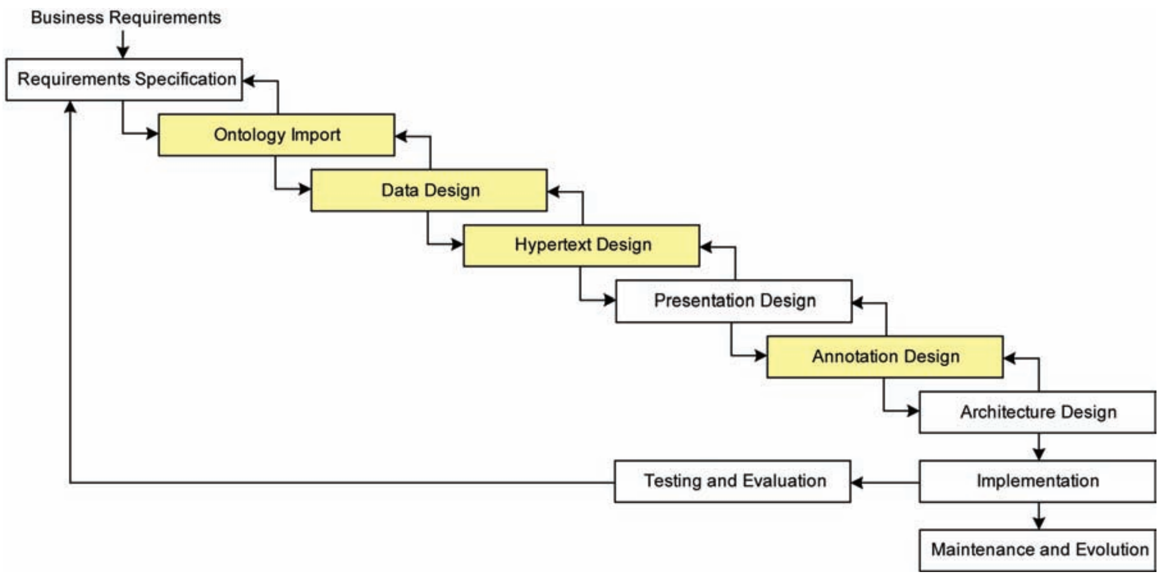
From the previous set of characteristics, we derived the following requirements for the conceptual models pursuing the design of Semantic Web applications:

- Semantic Web application models should be aware of and support semantic languages.
- The models themselves should be "semantic", i.e., grant self-annotation and explicit semantic extraction.
- The models should allow flexible integration of heterogeneous sources and applications.
- The models should allow transformations towards a query language able to capture all the aspects of ontologies, including inference, verification, query on instances, and query on classes.
- The models should easily allow: to specify semantic data sources as underlying level of the application; to exploit these sources for populating Web pages, and for (automatically) annotating such Web pages.
- The models shall be able to import and reference distributed data and ontologies, aiming at the reuse and sharing of the knowledge.

MODELING SEMANTIC WEB APPLICATIONS WITH WEBML

In this section we give an overview of extensions to the WebML methodology and models that we developed for complying with the requirements of Semantic Web applications (see previous section). WebML showed its flexibility and ease of extension in many other contexts (e.g., Web Services, Processes, Adaptive Web applications, Rich

Figure 5. The extended development process for Semantic Web applications, new and modified steps have yellow background



Internet applications): this allowed to consolidate and standardize the extension process. Basically, extending WebML for a new design domain (like Semantic Web portals) requires to introduce changes to the overall development process and to the related metamodels (and if required to introduce new metamodels in the development process). In particular, to deal with Semantic Web applications, we analyzed the current aspects of the WebML methodology and provided the proper extensions for each of them: the development process is enriched with steps that allow to describe the tasks related to the design of ontologies and semantics of the web applications/services; the data model is extended to support semantic data sources (i.e., ontologies); the hypertext model is enriched with new primitives that support ontology querying, with particular attention to advanced and inferring queries. Finally the presentation model is extended to allow for semantic annotations of the applications.

Extending the WebML Development Process

The injection of semantics within Web applications requires the extension of the methodology adopted in the development of "traditional" Web applications with additional tasks that formalize the new design steps. Figure 5 depicts the extended version of the development process for Web applications (yellow blocks represent the new tasks we introduced to fulfill Semantic Web application requirements). The original version was proposed in (Ceri et al., 2002) and is adopted with slight variations by most of the existing Web Engineering approaches. We want to stress that this development process is of general purpose and can be used for developing Semantic Web applications regardless of the use of any model-driven technique (actually it is valid also in case of traditional development techniques).

During *Requirements Specification*, a software analyst derives information about the application domain and functional requirements for the application from the business requirements and produces a detailed and formal specification for application designers. Already at this step some requirements related to semantic aspects can be highlighted and properly formalized. Based on the resulting specification, the designer can select existing domain ontologies to be imported (*Ontology Import*) and integrated in the Web application. Both ontology schemas and ontology instances maybe imported at this point. Never the less, for the Web application design, only ontology schemas are relevant. The imported ontologies can be possibly modified or merged to better suite the Web application purposes with specialized tools like Protégé (Noy et al., 2001). During the *Data Design* phase, the database structure and new ontologies can be created. Then, during the *Hypertext Design* the actual Web application structure is designed; in this phase, new primitives allows to specify how to query ontologies. Notice that the design of interactions with relational data remains unchanged with respect to the standard WebML design techniques. At this point the designer takes decisions about the *Presentation Design* of the Web application (e.g. design of graphical mock-ups and resources) and about the semantic annotations he wants to include in the rendered pages. The *Annotation Design* enriches the Hypertext Design and relies on the Presentation Design for deciding the actual position, formatting, and display style of annotations. *Architecture Design* mainly concerns the definition of hardware and software components as well as the design of the required network infrastructure. Once an overall application design has been specified, the *Implementation* phase can be carried out. It produces in output the software components (databases, ontologies, HTML templates, business logics, etc.) that build up the actual Web application, running on the selected architecture. During the following *Testing and Evaluation* phase, several tests (regarding e.g. functionality, usability, and performance) validate the application's conformance with respect to the initial business requirements, and could lead to novel derived requirements or modifications of the implementation. After testing, the application is released and undergoes ordinary maintenance and evolution activities. Notice that the waterfall representation may be adjusted for some design experiences, considering that in some cases some steps are not needed at all (e.g., if only imported ontologies are necessary, the data design step can be skipped). We did not depict all the variants for sake of clarity.

In the following paragraphs we discuss in more detail the changes introduced to the development process (yellow blocks in Figure 5).

Extending the WebML Data Model

Current existing model-driven methodologies for Semantic Web applications either evolved from existing ones by extending their data source coverage to ontologies, or have born with native support of semantic data sources.

Although ontology support is obviously necessary for Semantic Web applications design, we think that relational data sources can still provide great added value to Web applications. Relational databases are still a valuable option for modeling portion of domain data where features offered by ontologies are not needed (e.g., hierarchies, polymorphism, reasoning support, …) and where performances are a key issue (e.g., transactional data).

Therefore, allowing seamless interaction between ontologies and databases is a desideratum of current Semantic Web applications. Notice that we do not aim at extending the data model of WebML so as to model ontologies (see Noy et al. 2001), but at allowing Web applications to query imported semantic knowledge together with relational sources. By adopting a conceptual model, like WebML, interaction between ontology instances and database instances can be quite straightforward.

The designer is in charge of carefully deciding what is going to be part of an ontological data source and what is going to part of a relational data source. As data design (both in case of ontologies and databases) in general is an already consolidated discipline, we do not provide further details in this chapter. The only critical aspect is that the designer needs also to foresee if there is any need of references between the two data sources (e.g., the URI identifier of a Foaf user profile may be referenced in the relational schema to connect some relational attributes).

Extending the WebML Hypertext Model

WebML comes with a basic set of primitives for data access (e.g., *Index* unit, *Multidata* unit, the *Data* unit) that have a general purpose meaning (see Section "Background") and are perfectly fitting in the role of query and navigation primitives for both relational and ontology sources. Indeed, this is a general feature of conceptual models: if the used abstractions are generic enough, they do not need to be changed when their data source grounding is changed. They can be easily extended for supporting the additional expressive power and the different data model of the ontological sources.

For example, in WebML the *Index* unit is extended so that, besides extracting lists of relational instances, allows to produce lists of instances of a particular class within an ontology model. In previous sections we highlighted some requirements related to this kind of query: (1) the possibility to show only direct instances or also inferred instances; (2) the need for querying both instances and classes, thus mixing instances and (sub)classes in the results too. The same principles can be applied to *Multidata* unit and *Data* unit.

The *Hierarchical Index* unit, already defined in WebML, fits perfectly with the ontological data sources since it can be extended to browse and publish a portion of ontology in a hierarchical tree representation: for instance, given a class, it allows to publish the hierarchical tree underlying it, comprising subclasses and instances.

As stressed above, these basic WebML primitives remain valid, nevertheless they require some small extensions to support challenges posed by ontological data sources. Ontologies allow for queries with a wider expressive power and require some different modeling rules for the information within respect to relational data. This reflects into changes in the notations that the primitives must use for defining the conditions and the selection of the data.

Some of the challenges posed by ontologies are:

- There is no distinction between relationships and attributes within the set of properties of a class. E-R style relationships might be considered as ontological properties having an URI as value, and attributes to ontological properties having a literal as value;
- Several Semantic Web framework (e.g., OWL, RDF) assume that any instance of a class may have an arbitrary number (zero or more) of values for a particular property;
- Properties specification may include cardinality constraints and classes as range (and of course domain). In this case, it is possible to publish as values also structured objects and not only atomic attributes.

The revised units allow to model in a visual and simple way queries over ontologies. The data integration between relational data sources and different ontologies can be tackled directly in the Hypertext Model, by exploiting the data flow mechanism provided by WebML. One of the main advantages of WebML is the ability to specify business logic of applications by interconnecting smaller business logic components (units) in chains and passing parameters between them. This allows to define complex business logics composed by units that may query different ontologies or relational data sources and then can exchange information along the chain of transport links in a very easy way. Thus, parameters on the links become the actual contact point between traditional and semantic data and provide the mechanism for orthogonalizing data issues and hypertext issues.

Advanced Data Access Primitives

Many possible queries, using semantic data sources and semantic query languages, cannot be expressed using the basic data access primitives presented in the previous section. By carefully analyzing the semantic query languages presented at the beginning of the chapter, we introduce a new set of operational primitives that cover advanced queries over ontological data available in these languages. In particular these new units are largely influenced by two languages presented in the background section, namely, SPARQL (W3C, 2007) and RDF Schema syntax (W3C, 2004c). SPARQL was also selected among the various semantic query languages as the one adopted in the implementation of the runtime components (see Section *Implementation Experience*) because of its large software support and because it is probably going to become a W3C Recommendation.

The core set of new units aims to fill the gap between the basic WebML querying primitives and the increased expressive power of semantic languages. The basic primitives can not be combined to design any query that heavily exploits reasoning and mixing between schemas and instances. Thus, the basic WebML primitives miss two of the main assets of semantic languages. For example, they do not allow to infer all the classes an instance belongs to; they do not allow to infer the classes a particular property belongs to (in ontological models, a property can be associated to more than one class, since it is modeled independently from the class itself). The new units (i.e., *SubClassOf*, *InstanceOf*, *HasProperty*, *HasPropertyValue*, *SubPropertyOf*), described in Figure 6, aim at providing explicit support to advanced ontological queries and allow to extract classes, instances, properties, values; to check existence of specific concepts; and to verify whether a relationship holds between two objects. These units have a polymorphic behaviour: their business logic and the results returned change according to the configuration of their input parameters. The complete summary of the behaviour of these units is presented in Table 1.

Some other capabilities are needed within Semantic Web applications. For instance, we may need to import at runtime ontologies instances compliant with ontological schemas defined (or imported) at design time; to extract semantic descriptions of ontology portions; or to merge and compose ontologies. To this purpose, we introduce three new units, visually represented in Figure 7. The *Set Composition* operation unit is able to perform classic set operations (i.e., union, intersection, difference) over two input sets of URIs, considering the hierarchy of the URIs involved. E.g. suppose we have two set of classes: A = {ProgressiveRock, Jazz, Metal}B = {Rock, JazzFusion}.In this case, the set operation will give the following results:*A ∩ B = {ProgressiveRock, Metal, JazzFusion}A ∪ B = {Rock, Jazz}*

since Rock is superclass of ProgressiveRock and Metal, and Jazz is superclass of JazzFusion.

The *Import Ontology* unit allows to import at run time ontological data sources consistent with one or more of the ontology models imported at design time. This unit validates imported data against existing schemas and according to the designer choice, allows to store only the url of the newly imported ontology (i.e., it will be remotely queried every time, so modification on the remote source will be propagated to the application) or to import the ontology in the local OWL/RDF repository (i.e., it will be accessed locally, but modifications to the original data will not be propagated to the application). Notice however that in our methodology the navigational model of the Web application cannot be changed dynamically at runtime, thus if the imported ontology contains a new class unrelated with already existing classes queried in the hypertext, these new class will not be reachable in the navigation of the hypertext.

The *Describe* unit returns the RDF description of an URI, thus enabling data exporting and semantic annotation of pages.

The different advanced querying units are designed such that they can be combined in chains to compose complex business logic pattern that that allows to retrieve artists or albums that are correlated to artist searched by the user. The value submitted in the form is passed to an index unit that, by means of a filter, extracts all the Artists whose name contains in the string submitted by the user. The URI of the Artist picked from the index is passed to the *HasPropertyValue* unit. This unit extracts a set of URIs (instances of the class Album or Artist) that have the passed URI as

Figure 6. The new WebML units for advanced queries on ontologies

Unit name and Symbol	Input	Output	Description
SubclassOf [ClassName1=?] [ClassName2=?]	c_1, c_2 c_1, ? ?, c_2	true if c_1 is subclass of c_2 the list of superclasses of c_1 the list of subclasses of c_2	Given two classes, it returns true if the first is subclass of the second. Given a class, it returns the list of its subclasses or superclasses.
InstanceOf [ClassName=?] [Instance=?]	i, c i, ? ?, c	true if i is instance of c the list of classes for i the list of instances of the class c	Given a class and an instance, it returns true if the object is instance of the class. Given a class, it returns the list of its instances. Given an instance, it returns the list of the classes it belongs to.
HasProperty [ClassName=?] [Property=?]	c, p c, ? ?, p	true if the class c has the property p the list of properties of the class c the list of classes having p	Given a class and a property, it returns true if the class includes the property. Given a class, it returns the list of its properties. Given a property, it returns the list of the classes it belongs to.
Has PropertyValue [Property=?] [Value=?]	p, v p, ? ?, v	the list of URIs where the property p has value v the list of possible values for the property p the list of properties with value v	Given a property and a value, it returns the list of resource that has that property with that value. Given a property, it returns the list of all the possible values of the property. Given a value, it returns the list of properties with that value.
Subproperty [Property1=?] [Property2=?]	p_1, p_2 p_1, ? ?, p_2	true if p_1 is subproperty of p_2 the list of superproperties of p_1 the list of subproperties of p_2	Given two properties, it returns true if the first is subproperty of the second. Given a property, it returns the list of its subproperties or superproperties.

Table 1. Design time and runtime descriptors for two semantic units

	Design Time Descriptor	Runtime descriptor
Semantic Index unit	<SWINDEXUNIT class="mf:Track" id="swinu1" name="Tracks" ontology="onto1"> <DisplayedProperties property="mf:title"/> <DisplayedProperties property="mf:descriptor"/> <SortProperties order="ascending" property="mf:title"/> <Filter boolean="or"> <FilterCondition id="fselector1" property="mf:playedBy" predicate="eq" name="Artist"/> </Filter> </SWINDEXUNIT>	<descriptor service="org.webml.onto.SWIndexUnitService"> <onto>onto1</onto> ... <input-params> <input-param type="mm:Artist" name="swdau2.rdf:ID" /> </input-params> ... <query type="SELECT"> SELECT DISTINCT ?instance ?p1 ?p2 WHERE {?instance rdf:type mm:Track . ?instance mm:title ?p1 . ?instance mm:descriptor ?p2 . ?instance mm:playedBy ?fs1 . FILTER (?fs1 = $swdau2.rdf:ID$)} ORDER BY DESC(?p1) </query> </descriptor>
Subclas-sOf unit	<SUBCLASSOFUNIT id="iof1" name="SubClassOf" ontology="onto1"> </SUBCLASSOFUNIT>	<descriptor service="org.webml.onto.SubClassOfUnitService"> <onto>onto2</onto> ... <input-params> <input-param type="mz:Genre" name="swinu3.rdf:ID" /> </input-params> <query type="SELECT"> SELECT ?id WHERE {?id rdfs:subClassOf $swinu3.rdf:ID$} </query> </descriptor>

is then passed to the *InstanceOf* unit that checks if they are instances of the class Artist. In this case, the URIs are passed over through the OK link to an index unit showing a list of Artist, otherwise the URIs are passed on the KO link to publish a list of Album (not shown in the figure).

Extending the Presentation Model to Support Semantic Annotations

Each WebML semantic unit has been designed as able to automatically extract a RDF description of its contents. The designer has to specify in the HTML templates how he wants to use the RDF annotations. This RDF fragments can be used to

Figure 7. Symbols of the new WebML semantics management units

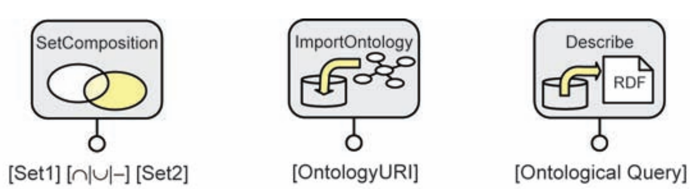
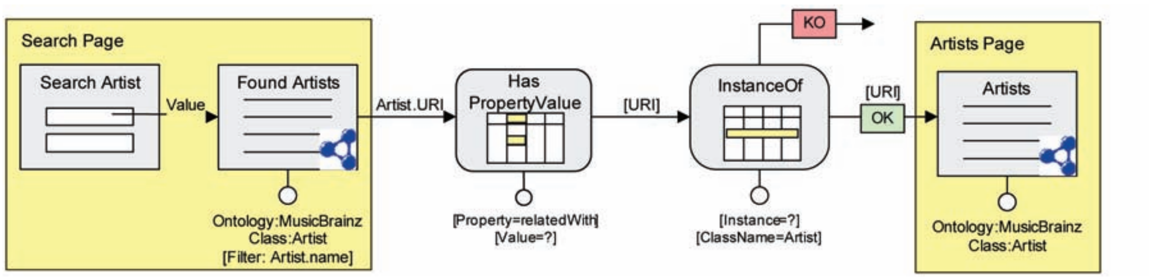


Figure 8. A portion of a Semantic Web application described by the new WebML units



annotate a single portion of the page (the one where the unit is rendered) or they can be aggregated and published together in a single predefined location, as a global semantic annotation of the whole page itself. Annotations can be made visible in the page for user reference or kept hidden in the HTML code if they are meant to be used only by machine readers. The WebML presentation model (that allows to place objects in a page grid, according to appropriate styles) can be exploited for defining how and where semantic annotations are rendered within the generated pages. If advanced annotation is needed, the solution is the *Describe* unit, which allows to query the ontologies to extract complex RDF fragments to be used as annotations. In this way, it is possible to extract semantic fragments also of instances or classes not published in any content unit of the displayed page. Different templates may be applied to different portions of the application according to the requirements of the developed Web application.

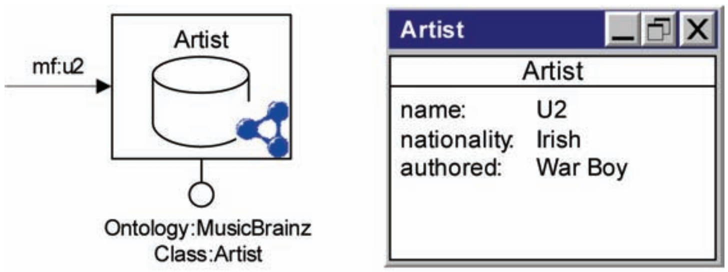
For example, the designer may decide to adopt a template that generates RDFa (Adida & Birbeck, 2007) annotations. RDFa is a syntax that expresses the structure underlying the published data using a set of elements and attributes that embed RDF in HTML. An important goal of RDFa is to achieve this RDF embedding without repeating existing HTML content when that content is the structured data. Within our framework RDFa is particularly easy to adopt since published data reflects an ontology schema; each displayed property value can be simply published with the proper reference to its property URI (setting the property attribute in the surrounding HTML tag), and reference resource (setting the about attribute in the surrounding HTML tag). Once the designer has defined the generic template, correct annotations are extracted in an automated way at runtime according to the underlying ontological model. Figure 9 presents an example of data unit with the relative generated RDFa annotation.

CASE STUDY: A MUSIC SEMANTIC PORTAL

In this section we show a simplified example of a Semantic Web portal modelled with WebML leveraging on the extensions we introduced in the previous section. For the design of the application we adopt the ontologies cited in the Introduction, namely the MusicBrainz ontology, the ontologization of the MusicMoz hierarchy; the RDF Site Summary for music news; and the Friend Of A Friend (Foaf) ontology. The main ontology adopted in the case study application is the MusicBrainz ontology; Figure 10 visually represents a fragment of it. The application includes also a small relational schema that contains registration information for the users, described by the User entity with its login data and the URI of its corresponding Foaf profile, if he has imported one in the Web application.

Figure 11 reports a fragment of the WebML model for the proposed application: content

Figure 9. An example of data unit with the relative generated RDFa annotation (highlighted in black).

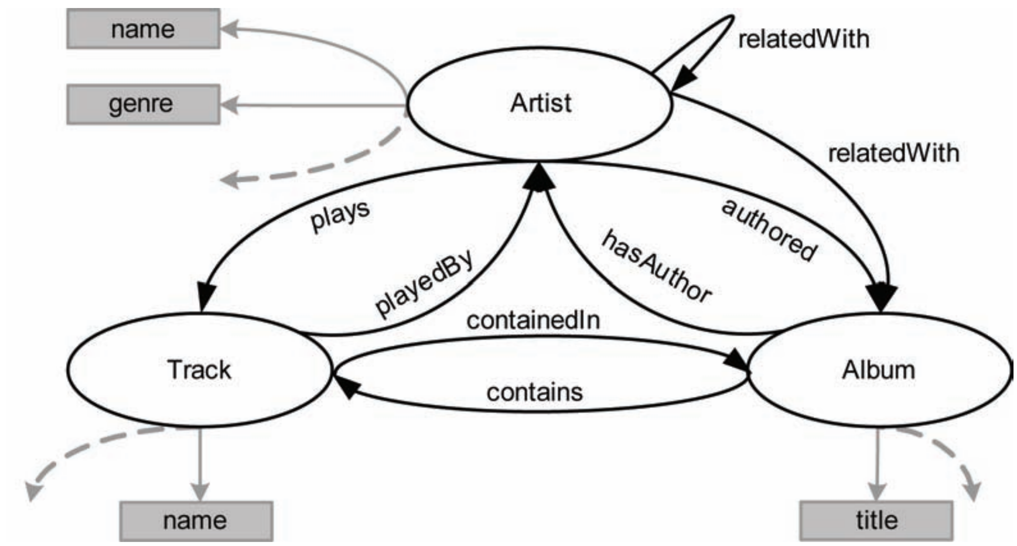



```
<TABLE xmlns:mf="http://example.org/mf#">
  <TR>
    <TD>
      <SPAN property="mf:name" about="mf:u2">U2</SPAN>
    </TD>
  </TR>
  <TR>
    <TD>
      <SPAN property="mf:nationality" about="mf:u2">Irish</SPAN>
    </TD>
  </TR>
  <TR>
    <TD>
      <SPAN property="mf:authored" content="mf:War"
       about="mf:u2">War</SPAN>
      <SPAN property="mf:authored" content="mf:Boy"
       about="mf:u2">Boy</SPAN>
    </TD>
  </TR>
</TABLE>
```

units with the RDF symbol use ontological data sources (e.g., *Artists* index unit), while the other units publish data from the a relational database (e.g., *User Data* data unit). The user starts his navigation from the *User Home Page*, where the *Foaf Profile* data unit is published; the user also import a profile if it is not available yet. This part of the application actually shows how integration between ontological data sources and relational data can be achieved using parameters transported over links: when the user imports the ontology that represents his Foaf profile, he actually stores the URI of the profile in the *User* relational entity; this URI is later used to publish his Foaf profile from the ontology repository according to the database schema.

The user can navigate from *Foaf Profile* data unit to the *Suggestion* page that presents an index of Artist corresponding to his preferences. From here (and from any other page presenting a visual query over the Artist class) the user can browse the *Artist details* page, where detailed information about the selected Artist and his Album are presented. The user can ask for the exporting of the RDF description of the Artist he is currently browsing. From the *Albums* index unit is possible to reach the *Album Details* page that reports information on the navigated instance of Album and its Track instances. Then accessing the landmark page *Search by genre* (a landmark page is a page accessible from any point in the hypertext also without an explicit link), the user can navigate a hierarchical representation of the class Genre, and then accesses all the Artist instances that are related to the selected genre. The *SubClassOf* unit extracts indirect sub-genres of the chosen one, thus

Figure 10. A fragment of the MusicBrainz ontology representing Artist, Album, and Track, with the respective relationships



allowing to display associated artists. Finally the *News* page reports an index of RSS Item and, by selecting an item from the list, the user can display the Artist associated with it (thanks to the filter that allows to retrieve Artist instances whose name or alias is included in the title or in the description property of the item). New sources of RSS Item may be imported adopting the *Import Ontology* unit, such as in the case of the Foaf profile.

IMPLEMENTATION EXPERIENCE

In this section we discuss the architecture design we adopted to implement the presented extensions to the WebML metamodel. These design choices are discussed according to the reference implementation of WebML, the Webratio toolsuite (WebModels s.r.l., 2007). As API to integrate ontologies both in the design environment and in the runtime framework we adopted the Jena framework (HP, 2007), while the reasoning support is obtained by means of the Jena integrated reasoner, or by means of the integration of Pellet (Sirin et al., 2007) with the Jena framework (the choice is made by the designer from the CASE tool). The runtime environment offered by Webratio has been extended exploiting the plug-in mechanism of the toolsuite:

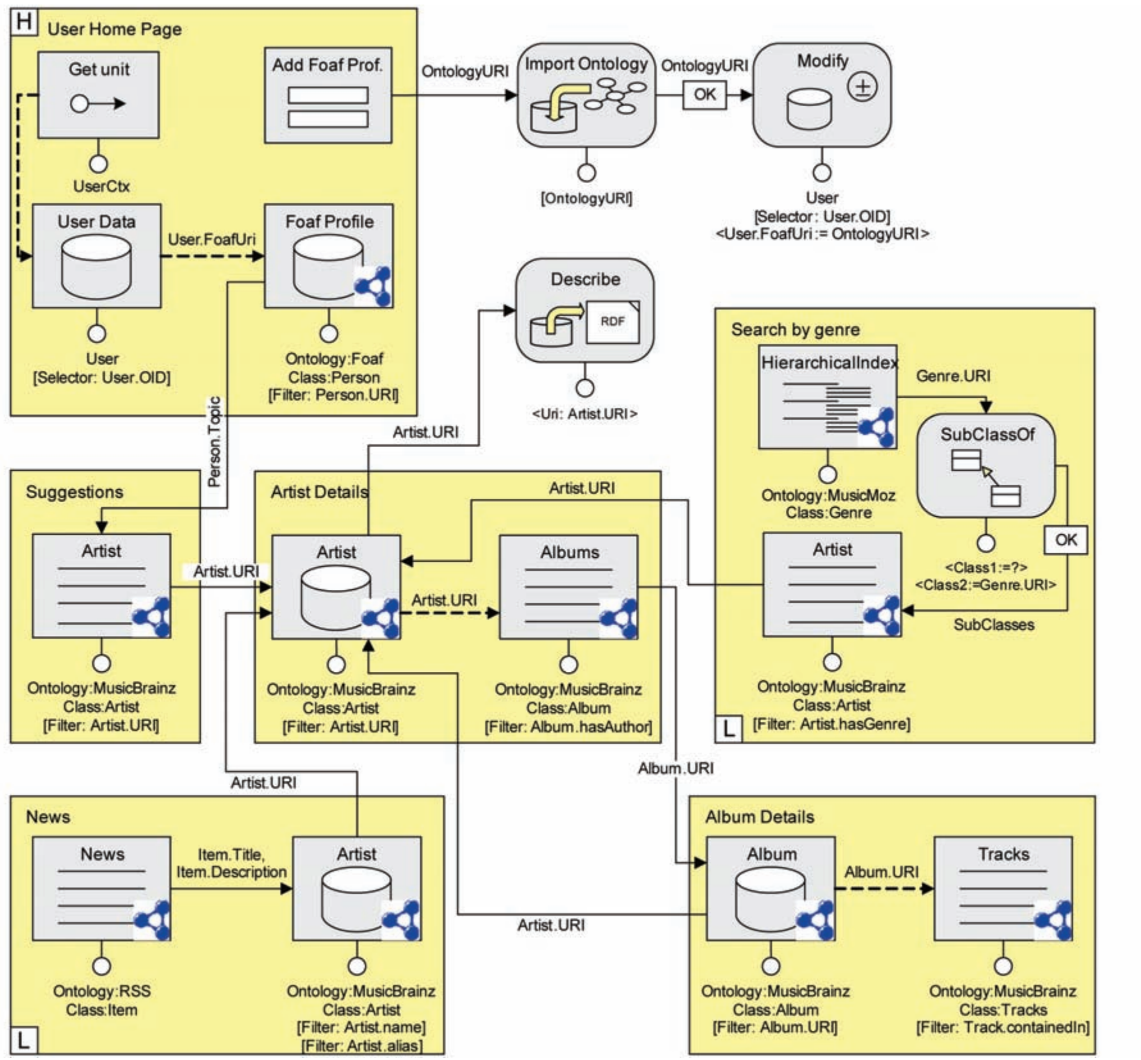
- We devised a general purpose ontology data access layer to be exploited by every unit;
- Then, we developed a runtime Java component and an XML descriptor for each unit.

Some extensions were needed on the design time interface too, in order to provide the proper management of ontological sources and units. Figure 12 reports a screenshot of the extended design environment showing the use case Semantic Web application presented in the previous section.

Implementing the Ontological Units

In the WebRatio framework, each unit is implemented by means of a generic class representing the runtime component that is executed for every instance of that kind of unit. Then, for each new

Figure 11. A portion of a WebML diagram for a Semantic Music Portal

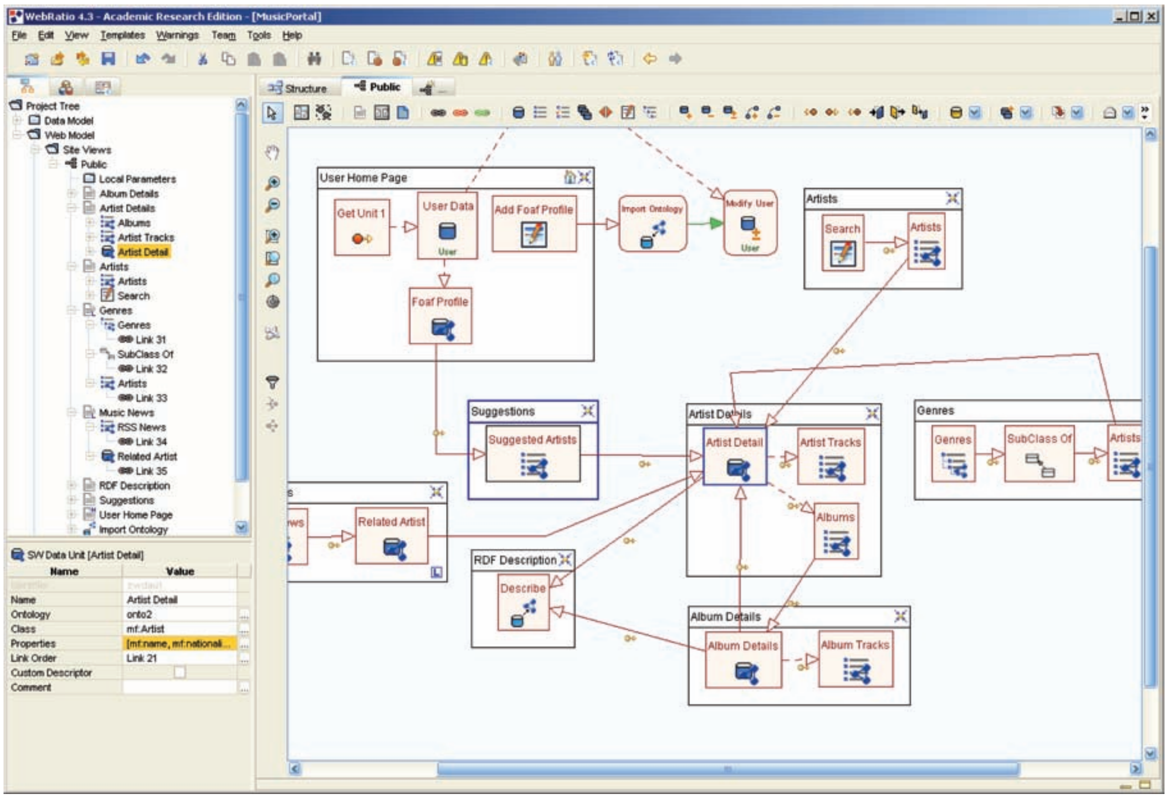


unit (including the revisited traditional units that access ontologies) we developed an XML descriptor specifying its parameters, its properties, the binding to the implementation classes, and so on. To better clarify the structure of the descriptor, we show two examples of an ontological unit descriptors (see left column of Table 1). By means of an associated XSLT transformation, design time descriptors are translated to runtime descriptors that include automatically generated template of SPARQL queries (right column of Table 1). Units are implemented by Java components that behave according to the logics specified in the runtime descriptors, defined for each instance of the unit.

Ontological Data Source Layer

We defined a new data access layer to allow the interaction with ontologies, comprising a set of general purpose Java classes to be reused by all the new units for querying the ontology repositories. These classes provide facilities to import ontologies and to select OWL/RDF classes, properties,

Figure 12. The Semantic Web Portal for the Music domain modelled with the WebRatio CASE tool



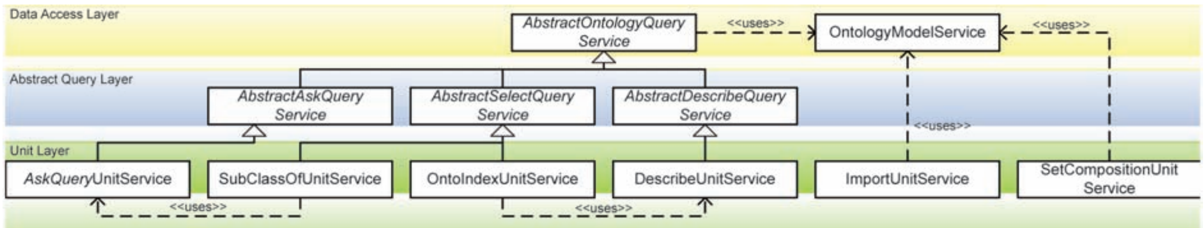
and instances (possibly filtered by one or more conditions). The main aspects of the class structure are represented in Figure 13. The *OntologyModelService* enables connections to local and remote ontologies specified at design time or imported at runtime by means of the Import Ontological Source unit. Three abstract classes offer the query services corresponding to the query methods offered by SPARQL on the ontology contents: the *AbstractSelectQueryService* class perform selection over data (SPARQL SELECT query); the *AbstractDescribeQueryService* retrieves the RDF describing a given URI (DESCRIBE query), the *AbstractAskQueryService* verifies simple predicates (ASK query). The AbstractAskQueryService is extended by the *AskQueryService* that is used by some of the advanced querying units to verify predicates (e.g., to check whether a class is subclass of another). In general, ontological unit services use or implement these general purpose services.

CONCLUSION

In this chapter we presented our research on the design of Semantic Web applications; we introduced briefly all the used background technologies and languages that are at the core of our proposal or that are adopted by other well-known frameworks. To ease the comparison of our work with other solutions, we described in detail Hera and SHDM, the most complete existing solutions.

The core of this chapter is the presentation of our extension to the WebML methodology and models for supporting the design and the specification of Semantic Web applications. The described solution provides a full coverage of the

Figure 13. UML class diagram of the class hierarchy of the new implemented units



development process: it allows the designer to specify basic and advanced queries on ontological data sources, to import existing sources, and to annotate Web pages with semantic descriptions of the contents and of the models. Our approach provides substantial added value with respect to the existing frameworks for Semantic Web application design, although some of them are more advanced on some aspects (e.g., seamless integration of different ontologies).

Table 2 reports a summary that compares the features of the previously cited models for Semantic Web Portals and the WebML extensions presented in this chapter. All the methodologies, except for XWMF, have a complete development methodology that covers all the needed aspects to create a Semantic Web application. They also offer a wide support for ontology languages: basically all the models support both RDF and OWL (except for XWMF). However, our extension is the only one that leverages on Semantic Web query languages to offer advanced query primitives that allow both query on schema and instances, together with simple reasoning patterns over data. Other models (e.g., Hera) offer query on data schema and instances. Hera and OntoWebber offer direct support to data integration by means of an integration model that can be used to query different data schemas using the same query. WebML offers the chance to integrate relational, XML and ontology data sources, while other methodologies seem to support explicitly only ontologies (of course, this issue can be solved by adopting extraction techniques to import other data sources within ontologies).

SHDM does not allow to import ontologies but only to create them from UML diagrams. Then, it offers a tricky way to link these ontologies to the external ones. Only a specific WSDM extension (Casteleyn et al., 2006) provides an approach to annotate pages so as to make them machine readable. Most of the new methodologies offer runtime frameworks that include or allow integration of reasoners, while some of them do not clarify if the reasoning is supported also at design time.

An important factor to assure the success of a design methodology is the existence of CASE tool support, since a powerful methodology that is not accompanied by adequate tools will make the designer tasks very difficult to fulfil. While most of the traditional design methodologies have powerful CASE tools, no established tool support is provided for Semantic Web design, although all the cited methodologies offer some basic tools. We support our proposal with a prototype implementation within the CASE tool WebRatio.

FUTURE RESEARCH DIRECTIONS

Future research will span on several directions: one of the main aspect we want to address is to generalize the way for extending the querying expressive power of the WebML units. Indeed, the provided extensions address some specific additional query power, but if further query ex-

Table 3. Comparison of methodologies for modeling Semantic Web portals

Requirement	XWMF	OntoWebber	SHDM	Hera	WebML+Sem.
Methodology	Partial	Yes	Yes	Yes	Yes
Semantic Model Description	Yes	Yes	Yes	Yes	Partial
Advanced query support	No	Partial	Partial	Partial	Yes
Flexible integration	No	Partial	Yes	Yes	Partial
Heterogeneous data sources	No	No	Partial	Partial	Yes
Distributed data sources	No	No	No	Yes	Yes
Reuse of ontologies	Yes	Yes	Partial	Yes	Yes
(Automatic) Annotation	No	No	No	No	Yes
Reasoning Support	No	No	Yes	Yes	Yes

pressive power is needed, researchers will need to devise new appropriate units. With a general representation framework, we plan to avoid this burden and to provide a quicker and more compact way for obtaining the results.

Another interesting aspect that was highlighted by this extension experience is related to the traditional way in which WebML provides querying facilities: standard WebML units (index unit, multi-data unit, data unit, and so on) provide both query service and publishing features in one single component (the *unit* specified within the page). Our extensions instead provide components that offer only the query services, without dealing with publishing, which is left to traditional WebML units. This separation of concerns could probably be applied to traditional querying too, so that the two aspects are always kept separate. This would mean a deep rethinking of the WebML hypertext models. We plan to consider this hint and to study possible solutions to the problem in the immediate future.

Besides pure research topics, we plan to extend and refine our implementation. Future implementation work includes:

- Providing a content integration layer to allow for seamless integration of different ontologies;
- Extensive testing of the new framework, including the application of the approach to real industrial scenarios;
- Integration of existing Eclipse based solutions for ontology editing with in the CASE tool.

REFERENCES

W3C (2004a, February). *OWL Web ontology language overview: W3C recommendation*. Retrieved June 16, 2007, from http://www.w3.org/TR/owl-features/.

W3C (2004b, February). *RDF Primer: W3C recommendation*. Retrieved June 16, 2007, from http://www.w3.org/TR/owl-features/

W3C (2004c, February). *RDF Vocabulary Description Language 1.0: RDF Schema W3C Recommendation.* Retrieved June 16, 2007, from http://www.w3.org/TR/rdf-primer/

W3C (2007, June). SPARQL Query Language for RDF: W3C Candidate Recommendation. Retrieved June 26, 2007, from http://www.w3.org/TR/rdf-sparql-query.

Adida, B., & Birbeck, M. (2007). *Rdfa primer 1.0: Embedding RDF in XHTML*, W3C Working Draft 12 March 2007. Retrieved June 16, 2007, from http://www.w3.org/TR/xhtml-rdfa-primer/

Aduna (2007). *Sesame*. Retrieved June 26, 2007, http://www.openrdf.org/.

Bailey, J., Bry, F., Furche, T., & Schaffert, S. (2005). Web and Semantic Web query anguages: A survey. In J. Maluszinsky and N. Eisinger (Eds.), *Reasoning Web Summer School 2005*, (pp. 35–133). Springer-Verlag.

Beckett, D. (2004). *Turtle - Terse RDF triple language*. Retrieved June 16, 2007, from http://www.dajobe.org/2004/01/turtle/.

Brambilla, M., Celino, I., Ceri, S., Cerizza, D., Della Valle, E., & Facca, F. M. (2006, November 5-9). A software engineering approach to design and development of Semantic Web service applications. In *The Semantic Web - ISWC 2006, 5th International Semantic Web Conference, ISWC 2006, Athens, GA, USA, Proceedings* (pp. 172-186). Springer.

Casteleyn, S. Plessers, & P., Troyer, O.D. (2006, *July 11-14*). Generating semantic annotations during the Web design process. In *Proceedings of the 6th international Conference on Web Engineering (Palo Alto, California, USA). ICWE '06* (pp. 91-92). ACM Press.

Ceri, S., Fraternali, P., Bongio, A., Brambilla, M., Comai, S., & Matera, M. (2002). *Designing data-intensive Web applications*. Morgan Kauffmann

Fensel, D., Lausen, H., Polleres, A., de Bruijn, J., Stollberg, M., Roman, D., & Domingue, J. (2006). *Enabling Semantic Web services: The Web service modeling ontology*. New York: Springer-Verlag.

Furche, T., Linse, B., Bry, F., Plexousakis, D., & Gottlob, G. (2006). RDF querying: Language constructs and evaluation methods compared. In *Proceedings of Summer School Reasoning Web 2006*, Lisbon (pp. 1–52). Springer.

Haase, P., Broekstra, J., Eberhart, A., & Volz, R. (2004, November 7-11). A comparison of RDF query languages. In *Proceedings of the third International Semantic Web Conference, Hiroshima, Japan*, (pp. 502–517). Springer.

HP. (2007). *Jena a Semantic Web framework for Java*. Retrieved June 26, 2007, http://jena.sourceforge.net.

Jin, Y., Decker, S., & Wiederhold, G. (2001, July 30-August 1). OntoWebber: Model-Driven ontology-based Web site management. In I.F. Cruz, S. Decker, J. Euzenat, & D.L. McGuinness (Eds.), *Proceedings of SWWS'01, The First Semantic Web Working Symposium*, Stanford University, California, USA. (pp. 529–547)

Karvounarakis, G., Magkanaraki, A., Alexaki, S., Christophides, V., Plexousakis, D., Scholl, M., & Tolle, K. (2004). RQL: A functional query language for RDF. In Gray, P., King, P., & Poulovassilis, A. (Eds.), *The Functional Approach to Data Management* (pp. 435–465). Springer-Verlag.

Klapsing, R., Neumann, G., & Conen, W. (2001, November). Semantics in Web engineering: Applying the resource description framework. [IEEE Computer Society.]. *IEEE MultiMedia*, *8*(2), 62–68. doi:10.1109/93.917972

Lima, F., & Schwabe, D. (2003). Application modeling for the Semantic Web. In *1st Latin American Web Congress (LA-WEB 2003), Empowering Our Web, Sanitago, Chile, Proceedings* (pp. 93–102). IEEE Computer Society.

Miller, L., & Brickley, D. (2000). *Foaf project*. Retrieved June 16, 2007, from http://www.Foaf-project.org.

Noy, N. F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R. W., & Musen, M. A. (2001). Creating Semantic Web contents with Protègè 2000. [IEEE Computer Society.]. *IEEE Intelligent Systems*, *16*(2), 60–71. doi:10.1109/5254.920601

Prud'hommeaux, E. (2004). *Algae RDF query language*. Retrieved June 16, 2007, from http://www.w3.org/2004/05/06-Algae/.

Racer Systems. (2007). *RacerPro*. Retrieved June 16, 2007, from http://www.racer-systems.com/.

RSS-DEV Working Group. (2000). RDF Site Summary (RSS) 1.0. Retrieved June 16, 2007, from http://web.resource.org/rss/1.0/spec.

Schwabe, D., & Rossi, G. (1998, October). An object-oriented approach to Web-based application design. In *Theory and Practice of Object Systems (TAPOS)*, 207–225. John Wiley & Sons

Sintek, M., & Decker, S. (2002). TRIPLE - A Query, inference, and transformation language for the Semantic Web. In *1st International Semantic Web Conference, ISWC 2002, Chia, Sardinai, Italy, Proceedings* (pp. 364 - 378). Springer.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2007). Pellet: A Practical OWL-DL Reasoner. [Elsevier]. *Journal of Web Semantics*, *5*(2), 51–53. doi:10.1016/j.websem.2007.03.004

Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., & Wenke, D. (2002). OntoEdit: collaborative ontology development for the Semantic Web. In *1st International Semantic Web Conference, ISWC 2002, Chia, Sardinai, Italy, Proceedings* (pp. 221 - 235). Springer.

Vdovjak, R., Frasincar, F., Houben, G. J., & Barna, P. (2003). Engineering Semantic Web information systems in Hera. [Rinton Press]. *Journal of Web Engineering*, *2*(1-2), 3–26.

WebModels s.r.l. (2007). *Webratio tool*. Retrieved June 16, 2007, from http://www.webratio.com

ADDITIONAL READING

Berners-Lee, T. (2003). *Web services - Semantic Web talk*. Retrieved June 16, 2007, from http://www.w3.org/2003/Talks/08-mitre-tbl/

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 5*(284), 34–43.

Brambilla, M., Ceri, S., Fraternali, P., & Manolescu, I. (2006c). Process modeling in Web applications. [ACM Press.]. *ACM Transactions on Software Engineering and Methodology*, *15*(4), 360–409. doi:10.1145/1178625.1178627

Broekstra, J., Kampman, A., & van Harmelen, F. (2002, June 9-12). Sesame: An architecture for storing and querying RDF and RDF schema. In *Proceedings of the First International Semantic Web Conference (ISWC 2002)*, Sardinia, Italy, (pp. 54-68). Springer.

De Troyer, O., Casteleyn, S., & Plessers, P. (2007). WSDM: Web semantics design method. In Rossi, G., Pastor, O., Schwabe, D., & Olsina, L. (Eds.), *Web Engineering: Modelling and Implementing Web Applications, Human-Computer Interaction Series, 12,* 303-351. Springer.

Fernandez, M. F., Florescu, D., Levy, A. Y., & Suciu, D. (2000). Declarative specification of Web sites with Strudel. *The VLDB Journal*, *9*(1), 38–55. doi:10.1007/s007780050082

Fons, J., Pelechano, V., Albert, M., & Pastor, O. (2003, October 13-16). Development of Web applications from Web-enhanced conceptual schemas. In *Conceptual Modeling - ER 2003, 22nd International Conference on Conceptual Modeling*, Chicago, IL, USA, *Proceedings* (pp. 232–245). Springer.

Karvounarakis, G., Christophides, V., Plexousakis, D., & Alexaki, S. (2001, *October 29-November 2*). Querying RDF descriptions for community Web portals. In *17iemes Journees Bases de Donnees Avancees (BDA'01) Agadir, Maroc Proceedings* (pp. 133–144).

Kifer, M., Lausen, G., & Wu, J. (1995). Logical foundations of object-oriented and frame-based languages. [ACM Press.]. *Journal of the ACM*, *42*, 741–843. doi:10.1145/210332.210335

Maedche, A., Staab, S., Stojanovic, N., Studer, R., & Sure, Y. (2003). Semantic portal - The seal approach. In Fensel, D., Hendler, J., & Lieberman, H., (Eds.), *Spinning the Semantic Web* (pp. 317-359). MIT Press, Cambridge, MA.

Murugesan, S., Deshpande, Y., Hansen, S., & Ginige, A. (2001). Web engineering: A new discipline for development of Web-based systems. [*Lecture Notes in Computer Science*. Springer.]. *Web Engineering*, *2016*, 3–13. doi:10.1007/3-540-45144-7_2

OMG. (2003). *Model driven architecture guide v1.0.1. Tech. rep., Object Management Group*. Retrieved June 16, 2007, from http://www.omg.org/docs/omg/03-06-01.pdf.

OMG. (2006). *Ontology definition metamodel. tech. rep., object management group*. Retrieved June 16, 2007, from http://www.omg.org/cgi-bin/doc?ad/06-05-01.pdf

Rossi, G., Schwabe, D., & Lyardet, F. (1999, November). Web application models are more than conceptual models. In *18th International Conference on Conceptual Modeling (ER 99), Paris, France, Proceedings* (pp. 239–252). Springer.

Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Maedche, A., Schnurr, H.-P., Studer, R., & Sure, Y. (2000). Semantic Community Web Portals. *Computer Networks*, 33(1-6) (pp. 473–491). Elsevier.

Staab, S., & Maedche, A. (2001). Knowledge Portals - Ontologies at Work. [AAAI.]. *AI Magazine*, *21*(2), 63–75.

Tzitzikas, Y., Spyratos, N., Constantopoulos, P., & Analyti, A. (2002). Extended Faceted Taxonomies for Web Catalogs. In *Third International Conference on Web Information Systems Engineering, WISE 2002, Singapore, December, 2002* (pp. 192–204). IEEE Computer Society

van der Sluijs, K., Houben, G. J., Broekstra, J., & Casteleyn, S. (2006). Hera-S - Web Design Using Sesame. In *Proceedings of the 6th international Conference on Web Engineering (Palo Alto, California, USA, July 11 - 14, 2006). ICWE '06* (pp. 337–344). ACM Press.

van Harmelen, F., Horrocks, I., & Patel-Schneider, P. (2001). *Reference Description of the DAML+OIL (March 2001) Ontology Markup Language*. Retrieved June 16, 2007, from http://www.daml.org/2001/03/reference.html

Vdovjak, R., Barna, P., & Houben, G. J. (2003). EROS: Explorer for RDFS-based Ontologies. In *ACM International Conference on Intelligent User Interfaces, IUI 2003, Proceedings* (pp. 330). ACM.

ENDNOTES

1. The serialization reported is the TURTLE (Beckett 2004) RDF serialization format.
2. Frame based languages are metalanguages that apply the frame concept to the structuring of language properties. They are rather focused on the recognition and description of objects and classes, and relations and interactions are considered as "secondary". They are widely adopted for ontology modeling since inference based on them it's easier than using other paradigms.

APPENDIX: QUESTIONS FOR DISCUSSION

Beginner

Q1: Are model-driven techniques influenced by technology changes?

A: Model-driven techniques aim at abstracting the modeling layer of applications from the actual implementation layer. I.e., model-driven approaches should capture abstract characteristics of the application field so that their model is resilient to technology changes and to the different implementation platforms that can be chosen. For example, the WebML methodology proposed in this chapter showed how abstraction introduced to handle relational data are still valid for semantic data sources, proving that the abstraction level of the adopted primitives doesn't depend on the underlying implementation and the underlying technologies (SQL or SPARQL). The same discussion applies to SHDM that evolved from OOHDM.

Q2: Why the import and reuse of ontologies is a key requirement for Semantic Web applications?

A: Reuse is one of the key goals of Software Engineering and is usually widely applied to software components and libraries. This aspect is even more emphasized at the level of semantic content specification and usage. Thanks to the introduction of Semantic Web languages, like OWL, also data model can be widely reused and shared. Indeed, one of the key factors for the success of the Semantic Web initiative is the spread of a set of domain ontologies with a wide consensus on their definition. This will enable different applications to be transparently integrated thanks to the fact that they are sharing the same data model.

Intermediate

Q3: Why the adoption of a standard query language is important in the definition of metamodels for Semantic Web applications?

A: The use of a standard query language contributes to abstract furthermore the modeling layer from the actual implementation. Indeed, if the implementation of the conceptual primitives relies on a specific API, this implies the change of the whole querying mechanism (comprising the generation of queries) if a slightly different implementation is chosen. On the other hand, the use of a query language (like SPARQL) enables to change the adopted implementation at the lower level of the adopted query engine. This is particularly evident with relational database technologies, where the SQL language (and possibly some standardized APIs like JDBC/ODBC) allows for (theoretical) total independence from the actual database engine adopted. Semantic Web technologies are still evolving very quickly, therefore a similar decoupling is still missing, but probably the wide adoption of the SPARQL language will provide a similar benefit.

Advance

Q4: The WebML extension for Semantic Web achieves different source integration thanks to parameter passing on the links between WebML components. Is there any other valuable option that still maintains valid the primitives proposed and offer the opportunities to integrated different data sources?

A: This issue can be easily solved by adopting one of the currently available semantic bridges that allows to use a single domain ontology to which heterogeneous data sources can be mapped (e.g., relational databases, ontologies, web services, xml documents). The mapped data sources can then be accessed through queries (e.g. SPARQL queries) to the domain ontology. The adoption of such a technology will affect only the implementation layer, maintaining valid the design primitives proposed and the generated SPARQL queries from the ontology model. The main difference stands on the level where integration occurs: instead of integrating the sources at the application modelling level, sources are integrated at the data design level.

Practical Exercises

Exercise 1. Define a navigation model for allowing users to navigate the popular Wine ontology using the WebML notation (http://www.w3.org/TR/owl-guide/wine.rdf)

Solution:

The simplest solution is to provide access to the wines directly through the wine categorization provided in the ontology or through flavors. A hierarchical index can be used for browsing the wine categories and selecting a wine. The Wine Details page shows the information of the selected wine.

Two simple index units can be used to display the flavors and hence to show the wines of the selected flavour. Once a wine is chosen, the user is sent to the Wine Details page again.

Exercise 2. Define a navigation model using the SHDM model-driven methodology for the Wine ontology that displays a result similar to the navigation model created in WebML.

Solution:

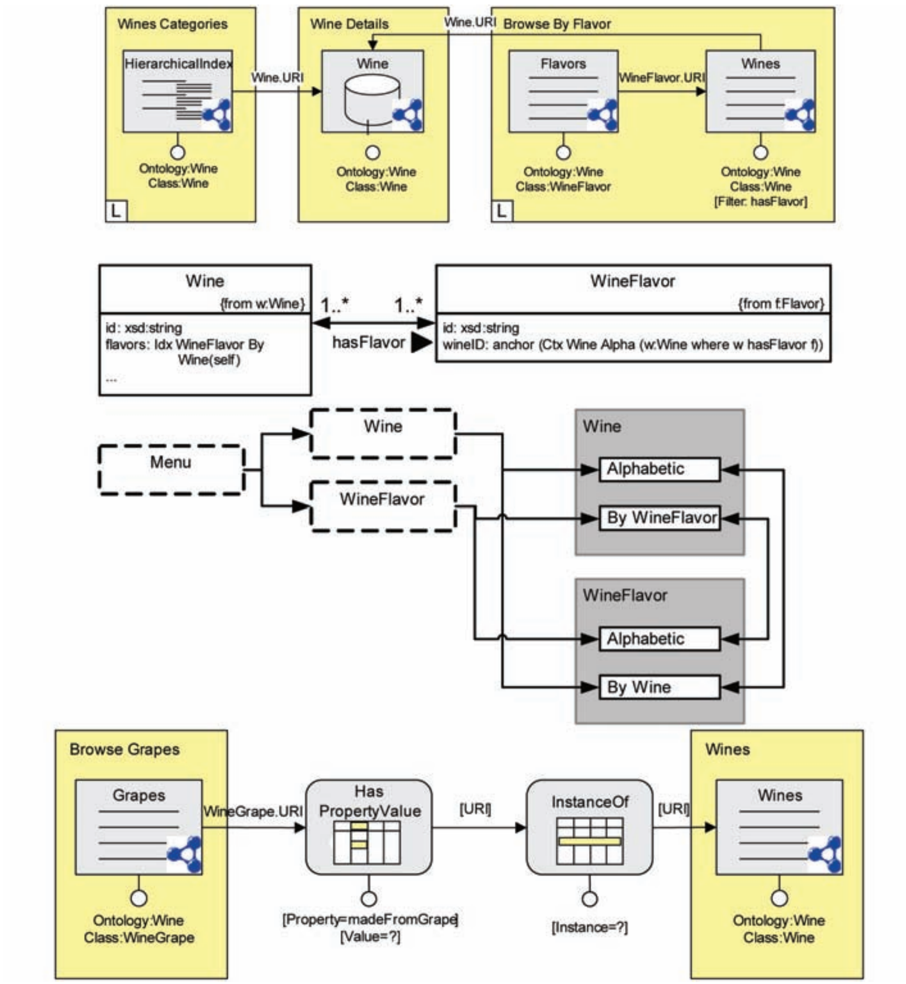
The content model shows the fragment of ontology that has been used. The navigation model represents a similar navigation to the one shown for WebML

Exercise 3. Model a WebML diagram that allows to select a Grape instance and then retrieves all the classes of Wine that use the selected Grape.

Solution:

The solution consists of an index unit that allows to select instances of class WineGrape, and of a chain of semantic query units that select all the objects made from the chosen grape (the HasPropertyValue unit) and then, among the returned objects, extract all the classes of the elements of the returned objects (InstanceOf unit). Finally, the list of classes is shown in an index. Notice that the parameter *[URI]* represents the set of URIs resulting from the respective queries.

Figure 14.



This work was previously published in Semantic Web Engineering in the Knowledge Society, edited by J. Cardoso and M. Lytras, pp. 46-106, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 2.17
Enabling Distributed Cognitive Collaborations on the Semantic Web

Amna Basharat
National University of Computer and Emerging Sciences, Pakistan

Gabriella Spinelli
Brunel University, UK

ABSTRACT

To date research on improving the state of multi-agent collaboration has only focused on the provision of grounding tools, technologies, protocols, standards and infrastructures that drive the Semantic Web and agent architectures. The basic cognitive and interactional requirements of agents have been neglected leading to the current state-of-the-art development of the Semantic Web whereby its full potential is constrained by the rigid state of multi-agent collaboration. This chapter illustrates and discusses an alternative approach to the development of the agent mediated Semantic Web. The fundamental premise of our approach is that enhancing agents cognitive and interactional abilities is the key to make the digital world of agents more flexible and adaptive in its role to facilitate distributed collaboration. The novelty of this research is that it adapts cognitive models from HCI to develop a heuristic framework called Cognitive Modelling of Multi-Agent Action (COMMAA) for modeling agents' actions in an attempt to provide an architecture that improves the flexibility of Multi-agent interaction by promoting cognitive awareness. The results of the evaluation show an improved flexibility, interoperability and reusability of agents' collective behaviours and goals.

INTRODUCTION

Agents may be autonomous and intelligent entities which typically operate in distributed collaborative environments called Multi-Agent Systems (MAS) which allows multiple heterogenous agents to collaborate by engaging in flexible, high-level interactions (Wooldridge, 2002; Jennings 2000). Presently, the usability of agent-based applications in a Semantic Web environment is limited due to lack of flexibility in agent's collaboration with multiple agents including humans. This imposes constraints on the interoperability and reusability of agents' behaviour that operate in MAS environment. In addition, the inflexibility of the agents' behaviour does not provide direct mapping to the end user since the end user cannot predict how the agent will behave, thus generating cognitive overload on humans. To date, research on improving the state of multi-agent collaboration has only focused on the provision of grounding tools, technologies, protocols, standards and infrastructures that drive the Semantic Web and agent architectures. Neglect of basic cognitive and interactional requirements are discovered to be the basic reasons for the rigid state of multi-agent collaboration constraining its full potential.

This research presented in this chapter adapts a distributed cognitive view of the agent mediated Semantic Web and argues that enhancing cognition is the key to make the digital world of agents more flexible and adaptive in its role to facilitate distributed collaboration. To this end, work on imparting cognition to improve interaction between multiple agents has been limited. The novelty of this research is that it adapts cognitive models from HCI to develop a heuristic modelling framework for COgnitive Modelling of Multi-Agent Actions (COMMAA) in an attempt to provide an architecture that improves the flexibility of Multi-agent interaction by promoting cognitive awareness. The highlight of the framework is that it identifies architectural and knowledge-based requirements for agents to structure ontological models for cognitive profiling in order to increase cognitive awareness between themselves, which in turn promotes flexibility, reusability and predictability of agent behaviour. The ultimate aim is towards applications which advocate user-centeredness such that as little cognitive overload is incurred on humans. The Semantic Web is used as an action mediating space, where shared knowledge base in the form of ontological models provides affordances for improving cognitive awareness.

Based on the rationale and concerns described above, the objectives and a brief outline of the chapter presented in the next section.

OBJECTIVES OF THE CHAPTER

The following chapter will serve the following aims and objectives:

- Delineate upon the current limitations in the state of multi-agent collaboration in order to elaborate the rationale, need and the synergistic role of cognitive dimension to the Semantic Web with particular regard to distributed collaborations amongst agents
- Describe the conceptual constituents of a theoretical framework called Cognitive Model of Multi-Agent Action (COMMAA) derived from cognitive models in HCI to improve the state of multi-agent collaboration
- Detail upon the Design and Implementation of Semantic Representational and Ontological Models based on the theoretical principles of COMMAA that allow cognitive processing of an agents action using state of the art Semantic Web technologies
- Describe heuristic reasoning mechanisms that can be derived from cognitive models to enhance the cognition of Semantic Web agents
- Analyze and discuss the impact of using COMMAA to model multi-agent collaborative applications on the Semantic Web

BACKGROUND

The Semantic Web vision of Berners Lee (2001) has enabled the Web applications to move from a purely human user community towards a mixed user community consisting of humans as well as of software agents. This imposes certain challenges and brings new requirements towards models for modelling Semantic Web-based systems (Scott et al. 2005; Klein et al. 2004; Neuhold 2003). The foremost issue is that of adaptive collaborative coordination and cooperation for utilizing services and Web information and imposes challenges from the perspective of interaction as well as interoperability amongst both agents and humans (Arai & Ishida 2004). As software agents become more capable and more prevalent, they must be able to interact with a heterogenous collection of both humans and software agents, which can play diverse roles in a system, with varying degrees of autonomy, initiative, and authority across different tasks (Schreckenghost et al. 2002). However, research supporting such interaction with these types of agents has received relatively little attention (Martin et al. 2003a; 2003b).

Limitations in the State of Multi-Agent Interaction

While there has been a significant proliferation of agent architectures and applications in the Semantic Web domain, there is significant separation of concerns from the principles that ensure flexibility of distributed interaction between heterogenous agents. Recently, much effort has been expended on making agents interoperate in the emerging open environments and standards. The Foundation for Intelligent Physical Agents (FIPA), an IEEE Computer Society standards organization, has attempted to facilitate the interoperation and inter-working between agents across multiple, heterogenous agent systems (FIPA, 2007). A variety of FIPA-Compliant platforms have emerged (Luck et. al (2005) provide a review). Despite this effort, this goal has still not yet been achieved as Louis and Martinez (2005a) point out. In addition, the Agentcities European project (Willmott, 2003) which resulted in the deployment of a worldwide open testbed environment, underlined the lack of 'spontaneous' exchanges between agents running in this environment. In almost all cases, agents can only interact with agents they have been designed to interact with. One reason is that agents are implemented using mechanisms such that they conform to only a limited set of interaction protocols generally resulting in inflexible or rigid agents. It is reported in research and learnt from previous experience (Ahmad et al. 2005, Shafiq et al. 2005, Tariq et al. 2005a, Tariq et al. 2005b) that the agents show unyielding behaviour to messages not specified by the protocol. Efforts such as Louis and Martinez (2005a;2005b) have attempted to address the issues but the focus has been largely to provide semantic handling of messages.

It is therefore believed that there is still a long way to go before true homogenisation of agent communities can be achieved. This is because the variance of agent communication and functional pragmatics introduces a certain level of mismatch and the need of flexible and adaptive interactions that promote interoperation becomes imperative. This is particularly essential when the agents from diverse platforms intend to co-exist and cooperate in mutual. In addition, most agent-based applications assume pre-defined knowledge of agents' capabilities and/or neglect basic cognitive and interactional requirements in multi-agent collaboration. Thus the research community is faced with the challenges of improving the limited visibility of agent's processing ability and behaviours that may be a result of possible mismatches between the agents' mental and implementation models. In addition, it is claimed that inadequate adjustability of the agent's autonomy and a basic lack of compatibility between the required capabilities and those provided by an agent impose further research challenges (Martin et al. 2003a; 2003b).

Potential of Cognitive Models to Improve the State of Multi-Agent Interaction

Some interdisciplinary research has stressed the potential of cognitive models studied in cognitive science as substantial means of better probing multi-agent issues, by taking into account essential characteristics of cognitive agents and their various capacities (Sun 2001). The term cognitive models has traditionally been associated with humans as cognition is essentially a human characteristic. Cognition can be thought of as a modelling process which creates a model from which deductions can be drawn (Meredith 1970). Humans are known to continually create and access internal representations of their current situation - referred to as their cognitive model (Saja 1985). It is said that the state of interaction with a system can be greatly improved by design activities that account for and support the emergence of a user's cognitive model. These models are referred to as mental models in human-computer interaction (Norman 1986; 1988).

Inspired by the potential of cognitive models, the research presented in this chapter investigates the possibility of meeting the above mentioned challenges by bringing cognitive models and theories in the Semantic Web to achieve more robust and effective architectures for agents that facilitate distributed collaboration– be it amongst agents or between agents and humans or vice versa. The premise of the research is based on the hypothesis that if artificial (or software) agents were to be designed to emulate the interactional and cognitive properties of humans in a complementary way, such that they interact with each other and their environment in the manner that humans do, it would increase their functional capability to serve humans. Additionally it would also reduce the cognitive load that humans require in distributed collaborations while viewing it from a distributed cognitive perspective.

RESEARCH MOTIVATION AND CONTEXT: DISTRIBUTED COGNITIVE VIEW OF THE AGENT MEDIATED SEMANTIC WEB

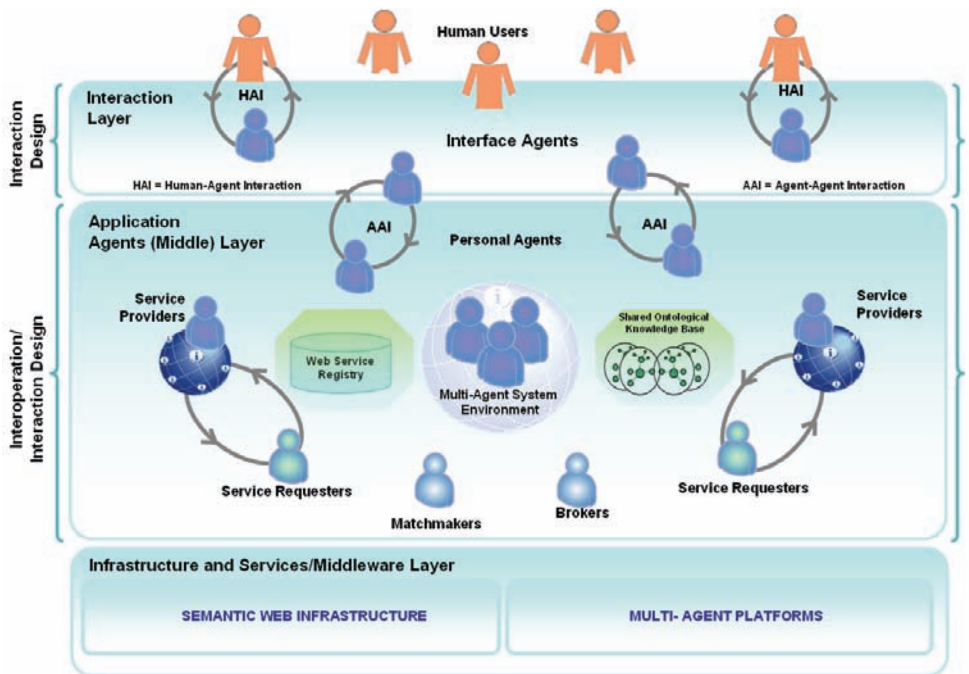
The view adapted for Semantic Web is that of a world-mediating system as it mediates between users and a part of the world, often by manipulating machine representations of the world (Clark, 2001). At the basis of this research is the idea to view the Semantic Web as a distributed cognitive system, a basic unit of analysis, composed of human and machine agents in a work domain that is delineated by roles, work and communication norms, artefacts, and procedures (Zhang et al, 2002).

According to (Lu, Dong & Fotouhi 2002) the Semantic Web uses ontologies to describe various Web resources, hence, knowledge on the Web is represented in a structured, logical, and semantic way allowing agents to navigate, harvest and utilise this information (Payne, et al., 2002a;2002b). Agents can also read and reason about published knowledge with the guidance of ontologies. Also the collection of Web-services described by ontologies like OWL-S (Ankolekar et al. 2001; Martin et al. 2005;2007) will facilitate dynamic matchmaking among heterogenous agents: service provider agents can advertise their capabilities to middle agents; middle agents store these advertisements; a service requester agent can ask a middle agent whether it knows of some provider agents with desired capabilities; and the middle agent matches the request against the stored advertisements and returns the result, a subset of the stored advertisements (Sycara et al. 2002;1999).

Based on the analysis of literature and state of the art, agent-mediated distributed computing paradigm for the Semantic Web is viewed as a layered abstract architecture shown in Figure 1 -a lens through which multi-agent collaboration can be viewed. Applying the distributed cognitive view of Semantic Web, the abstract layered model revises the traditional view of the Semantic Web by adding cognitively modelled interactions.

The above abstract model of the Semantic Web and agent characteristics requires multi-agent interaction which can consist of three interaction levels: human-human, human-agent, and agent-agent. In each interaction level, both interaction design and interoperability are necessary for

Figure 1. Layered Abstract Model for illustrating various levels of Cognitive Collaborations in Agent Mediated Semantic Web



mutual accessibility and understanding among them as has also been highlighted by Arai & Ishida (2004). The chapter extends the vision for the need and requirements for modelling these interactions from a distributed cognitive perspective (Basharat and Spinelli, 2008a). By applying the Distributed Cognition perspective Chandrasekharan (2004) this model considers the importance of studying interaction and interoperation amongst multi-agents not in isolation but within the environment agents inhabit. The combination of the Semantic Web inspired by cognitive model can generate a framework where agents and application can better cooperate.

After a review of the available cognitive approaches that model human activities in interaction with any system (artefacts and the environment), the Action Cycle (AC) (Norman 1986; 1992) has been selected as the most promising approach for this research. The aim is to unfold the potential of the AC in an attempt to identify the design and interactional gaps between heterogenous agents on the Semantic Web. The resulting contribution lies in the adoption of the AC to develop a heuristic modelling framework called COgnitive Model of Multi-Agent Actions (COMMAA).

The framework is proposed to model multi-agent actions in a collaborative MAS environment from a cognitive perspective. The framework is intended to aid the designer in modelling the agent behaviour and action through a cognitive cycle. Using the principles of the proposed framework designers can define both functional and non functional aspects of the design of an agent's interactive role in a collaborative scenario, especially focusing on the concepts of semantic and articulatory distances, as derived from AC, as mismatch between agents' goals and its functional capabilities.

Not only this research proposes the theoretical guidelines for cognitive modelling of agents, it also provides design illustration for the architectural elements necessary for realizing these principles in the Semantic Web context. Thus the framework identifies architectural and knowledge-based requirements for agents to structure ontological models for cognitive profiling in order to increase cognitive awareness amongst agents. By cognitive awareness, a term coined within this research, it is intended the ability of the Web agents to diagnose their processing limitations and to establish interactions with the external environment (in the form of other agents including humans and software agents) using the principles derived from the framework for COMMAA. This is with the aim to support users' goals in a more direct manner by providing agents that can share, discover and access each other's capabilities in a collaborative manner and are able to function dynamically and adaptively without continuous human intervention.

This brings about a more effective MAS environment; where agents may delegate each other tasks and goals based on each other's awareness of abilities, behaviours and affordances. The ultimate aim is towards applications which advocate user-centeredness such that as little cognitive overload is incurred on humans. The strength of this framework lies in its robust theoretical foundation that has found validation in a developmental infrastructure that helps realise the theoretical principles.

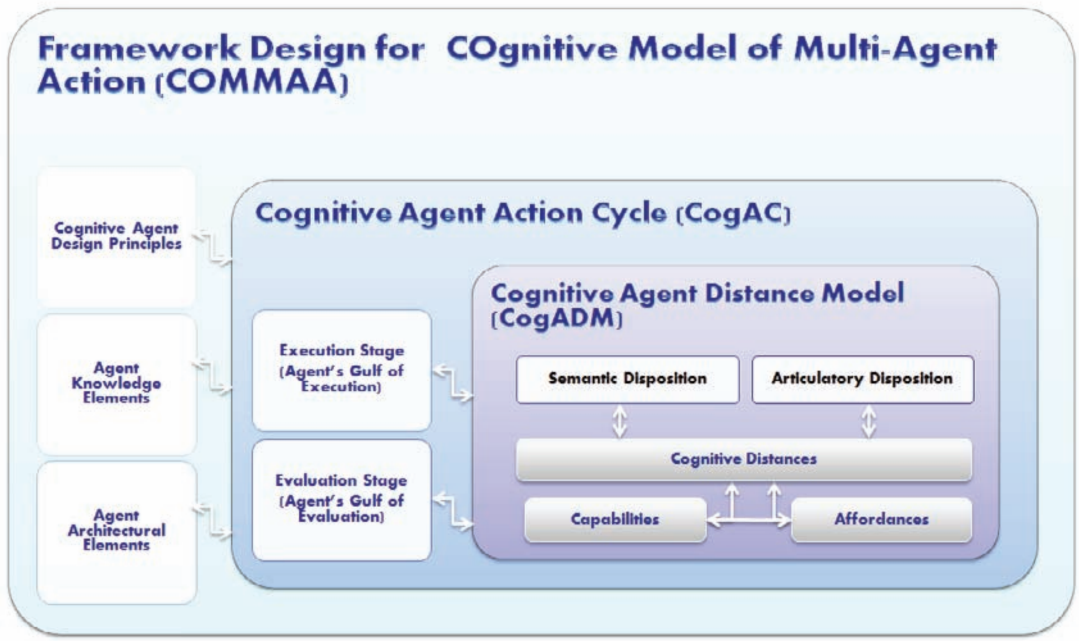
THEORETICAL FRAMEWORK AND CONCEPTUAL CONSTITUENTS FOR COGNITIVE MODEL OF MULTI-AGENT ACTIONS (COMMAA)

The framework developed in this work and presented below is based on these fundamental principles of Agents' action, borrowed and modified from the Human Action Cycle (Norman 1986) and its elaboration for Direct Manipulation Interfaces by Hutchins et al. (1986). The framework of COgnitive Model of Multi-Agent Actions (COMMAA) is shown in the Figure 2. The primary aim of the framework is to cognitively model the agent's action in a collaborative MAS environment situated in the Semantic Web such that the limitations in the state of Multi-Agent Interaction can be overcome.

The conceptual constituents of the framework of COMMAA include:

- **Cognitive agent Action Cycle (CogAC):** COMMAA is based on Cognitive agent Action Cycle (CogAC) which serves as the fundamental core of the framework, and is designed to aid the designer elaborate the agent's functional behaviour using two stages namely Execution and Evaluation, each with its respective steps. The CogAC views agent as the primary entity that interacts and functions in a MAS environment. The stages of an agent interacting with a MAS environment are described such that in order to accomplish a goal, which is in turn delegated to it by a human user, the following steps are traversed by an agent: Goal Formation, Intention formation, Action specification, Execution, Perception, Interpretation and Evaluation.
- **Cognitive Distance Model (CogDM):** The further elaboration of the steps of CogAC leads to the formulation of agent's semantic and articulatory disposition in each stage of execution and evaluation, described and illustrated by the Cognitive Agent Distance Model CogADM. These dispositions also help to identify the agent's Gulf of Execution and Evaluation. These two are discussed in the subsections to follow.
- **Cognitive Agent Design Principles:** The principles are derived using the Action Cycle mapped for agents. The principles may serve as heuristic for evaluating the design of agent-based applications. They help to identify the significant mismatches, constraints and affordances. The framework

Figure 2. Conceptual design of Theoretical Framework / Conceptual Constructs of COMMAA – Cognitive Model of Multi-Agent Action



defines CogADPs only at abstract level. These serve as blueprints which may be specialised to a domain specific context by the designer to derive a context specific cognitive profile for an agent. These may be specialised according to the domain knowledge and the specific contexts of application the agent may be operating in. The design principles aid in Cognitive Mismatch (Distance) Analysis, that allow at design time to be made known the possible stages where distance of execution or evaluation may occur. The designer by analyzing whether provision for these principles in made in the agent's infrastructure can help develop the cognitive profile of the agents, alternatively agents may have the dynamic capability of identifying these distances at runtime and may change, update their profile dynamically.

- **Agent Architectural Elements:** As each sub-stage of agent's execution and evaluation stages are elaborated, and as distances are identified, it helps in identifying the architectural needs and components required for successful completion of agent action cycle. As these are identified by the designer, they may be cross checked against the environment that the agent in being built using. In other words, it may also be said that provision of these capabilities ensure that the agent's Cognitive Design Principles will be met to some extent. The availability of these elements would ensure that the agent is Cognitively Directed, or Cognitive Directness is exhibited in agent's action with respect to its interaction with the environment. The Architectural elements identified ensure the minimal design requirements that must be met in order to bridge the agent's gulfs of execution and evaluation.
- **Agent Knowledge Elements:** The knowledge requirements are identified by the designer such that at each stage of agent's execution and evaluation, these are the elements that the agent must possess or must

be provided with in order to achieve the successful realisation of the respective stage.

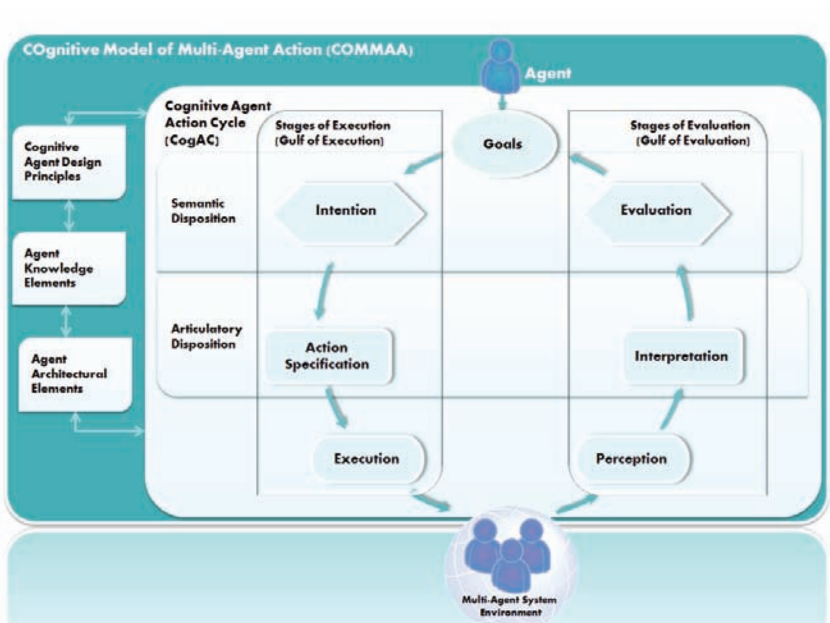
Cognitive Agent Action Cycle (CogAC)

Agents are designed to continually act upon and monitoring the MAS Environment, interacting with it and collaborating with other agents and entities, evaluating its state, and executing actions. The system is a closed loop: when agents act, it is usually done so in response to some prior evaluation of its perceptions or as a result of some goal delegated to it by a human. After an agent acts, it evaluates the impact of the executed act, often modifying the action as it carries it out. In MAS environment the fundamental unit of agent's social ability is its interaction with other agents using messages in Agent Communication Language (ACL). Agent's Interaction is modelled to have two stages as shown in Figure 3. The Interaction will have some Goal i.e. the objective that needs to be achieved using the interaction. In order to achieve this goal, agent will need to go through the two stages namely Execution and Evaluation in order to successfully achieve the goals of the interaction respectively.

The stages of an agent interacting with a MAS environment are described such that in order to accomplish a goal, which is in turn delegated to it by a human user, the following steps are traversed by an agent: Goal Formation, Intention formation, Action specification, Execution, Perception, Interpretation and Evaluation (as shown in Figure 3.). The essential concepts are the Gulfs of Evaluation and Execution, each arising as a result of semantic and articulatory distances (cognitive distances in general). The relevance of these concepts to the agent domain is formally described in the next sections in the form of Cognitive Agent Distance Model, which are detailed further.

The inter-relationship between the CogAC, its stages of execution and evaluation, the semantic and articulatory dispositions and other elements of COMMAA is schematically shown in Figure 3. In Table 1 and Table 2 these stages are elaborated. In addition, a generic view of the corresponding CADPs, Knowledge Requirements and archi-

Figure 3. High-level conceptual design of CogAC in relation to other constructs of COMMAA



tectural needs are also identified. Together with these main components, the three additional components are identified to help the designer model agents' behaviour in a more robust manner. Cognitive Agent Design Principles, Agent Architectural Elements and Agent Knowledge Elements are all identified, as each stage of the agent's execution and evaluation are elaborated upon.

Cognitive Agent Distance Model (CogADM)

On the Semantic Web, Knowledge is invisible and intangible. While meanings are essential to knowledge, they cannot get across to an agent without some kind of representational form. Knowledge representation has two aspects: the meaning of the information, named semantics, and the physical form or appearance, named syntax. When an agent interacts with a knowledge representation, it interacts with both the semantics and the syntax.

There is often, however, a gap, known as Cognitive Distance, between the knowledge an agent needs and the manner in which this is represented in its environment, as shown in Figure 4. This manner of representation also includes the mechanisms with which the knowledge is accessed and reasoned about.

The prospective relevance of the cognitive distances is highly relevant to Communication, collaboration and interactions taking place between agent application residing on the Semantic Web - since the basis of interaction is the communication language and its vocabulary represented as knowledge on the Semantic Web.

Cognitive Distance in this context is taken to be a measure of the gulfs of execution and evaluation

—the conceptual gap, or mismatch between the agent's goals and intentions, and the way in which they are in cohesion to, or represented by, the multi-agent system environment. A large distance is representative of a large gulf in the execution or evaluation stages, signifying that a lot of cognitive load is incurred in translating the

Table 1. Steps of CogAC in Execution Stage (Agent's Gulf of Execution) cross referenced with design principles, Knowledge elements and Architectural requirements

Steps of CogAC	Description	Cognitive Agent Design Principle (CADP)	Agent Knowledge Elements	Agent Architectural Elements
Execution Stage (Agent's Gulf of Execution)				
Step 1: **Form Goal**	•An Agent's Goal is the state the agent wishes to achieve	•Compatible goal formation •Intention formation •Agent must have the capability to express goals to desired intentions •Semantic Distance Occurs if the capability is not provided to the agent. •Agent is being required to work at lower level of detail, resulting in greater semantic distance	•Task Knowledge Domain knowledge: Appropriate Situation •Affordance for Goal Representation and Formation •Affordance for Expressive Power and Communication	•Goal Representation •Goal to Intention Formulator •Task Representation •Domain Knowledge Representation •Intention Representation
Step 2: Form/ Specify Intentions	•An intention is the decision to act so as to achieve the goal	•Exhibits Semantic Directness of Interaction with the Environment (Disposition of Agent is Semantically Direct wrt its execution of its interactions in the given environment) •Functional Capability for Intention Formation	•Task Knowledge : intention details	•Content Representing the intentions, as rules or some construct •Heuristic Reasoning needed for Analysis of Functional Capabilities and Semantic Directedness
Step 3: **Action Specification**	•The process (mental) of determining the logical representation of the actions that are to be executed by the agent on the mechanisms of the system (MAS environment) •Action must be taken with respect to the environment the agent is operating in	•Constraints imposed by the environment or the world of the agent must be taken into account •Articulatory Directedness (Disposition of Agent is does not posses any Articulatory distance wrt to execution of its actions in the given environment) •Affordance for Action	•Task Knowledge : Action Details •Domain •Environment Knowledge: Constraints (e.g. Representational)	•Actions Represented as some protocol e.g. Conversation Protocol or Interaction/Negotiation Mechanism •Heuristic Reasoning needed for Analysis of Affordances available and determining the Articulatory directedness
Step 4: **Execution**	•Agent executes the tasks formulated as a result of action specification	• Using the MAS Environment and the middleware to execute its actions	• Knowledge of Execution Mechanisms	•MAS Middleware, Execution Framework, Control Mechanism for Agent's Execution

Table 2. Steps of CogAC in Evaluation Stage (Agent's Gulf of Evaluation) cross referenced with design principles, Knowledge elements and Architectural requirements

Steps of CogAC	Description	Cognitive Agent Design Principle (CADP)	Agent Knowledge Elements	Agent Architectural Elements
Evaluation Stage (Agent's Gulf of Evaluation)				
Step 5: **Perception**	•Perceive the output, response of an event, a trigger or a message	• Affordance for perceiving the state of the environment	•MAS/Platform Knowledge: •Constructs, Content, Protocols	•Means of Feedback Specification •Means of Perception
Step 6: **Interpretation**	•The relationship between the state of the MAS environment and the goals of the agent can only be determined by first translating the environment state into psychological states for the agent then interpreting the perceived environment state in terms of agent's variables of interest.	•Articulatory Directedness (Disposition of Agent does not exhibit Articulatory distance wrt to evaluation of its actions and the response of the environment) •Affordance for Reasoning •Affordance for extracting the required meaning from interaction output •Affordance for Interpretation of Feedback of the Interaction	•Domain Knowledge : Expected effects	•Means of Interpretation •Reasoning Mechanism for Interpretation •Logic Mechanism •Feedback Means •Means to Recognize and Interpret Feedback •Interpretable effects
Step 7: **Evaluation**	•Evaluation of the environment state requires comparing the interpretation of the environment state with the desired goals. This often leads to a set of new goals and intentions.	•Agent's Functional Capability for Evaluating the Outcome to determine if the goal has been achieved and to determine if it possesses the means to predict the outcome •Semantic Directedness (Disposition of Agent is Semantically Direct wrt to evaluation of its actions and the response of the environment	•Task Knowledge : expected effects •Expected Feedback	•Logics based on which Evaluation is to be carried out •Heuristic Logics •Representational Logics •Expectation Means •Predictability

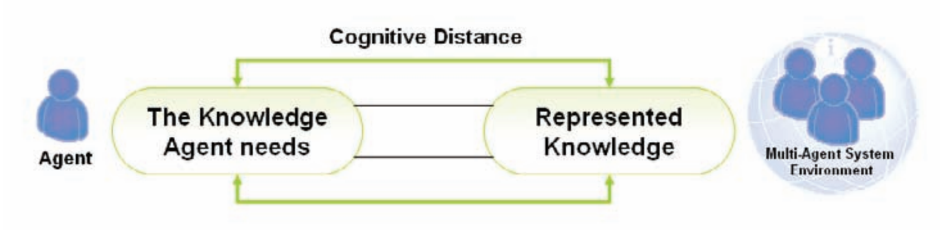
agent's intentions into the system's representations, or vice versa.

That is, a large distance of execution means it is relatively difficult or not possible for the agents to express their query or desires to the system, and a large distance of evaluation indicates mismatch of some form within the agent's infrastructure to interpret or evaluate the system's output or response as a result of some interaction with it.

Directness of Agent's Interaction with MAS Environment

Using the directness of interactions as usability measures would involve understanding the agent's

Figure 4. Cognitive Knowledge Gap between agent and multi-agent system environment



problem solving strategies, approaches and intentions as now well the MAS environment supports the agent's functional needs. A design and evaluation methodology that assesses directness requires cognitive basis because understanding the user's mental processes is key to assessment. The concept of directness as suggested by (Hutchins et.al, 1986), and later by (Cuomo 1993) is adapted here to refer to the degree of capacity of an agent to bridge the Gulfs of Execution and Evaluation.

The concept of multi-agent interaction is virtually an unexplored area in terms of operationally defining and assessing the directness of interactions an agent engages in with its MAS environment to a degree that can be applied in practice and measured using qualitative or quantitative tools. Directness of an Agent's Interaction with its environment may be used as a qualitative indicator of the amount of cognitive processing needed to carry out a successful interaction. Directness is inversely proportional to the amount of cognitive processing it takes to manipulate and evaluate the results of agents' interaction with its MAS environment. Moreover the cognitive processing required for an agent is a direct result of the gulfs of execution and evaluation the agent has to deal with. The better the agent's architecture, the less cognitive processing needed and the more direct the resulting interaction between agents.

Thus Distances are complementary to Directness; the lower the distances, the greater the directness, and vice-versa. The two terms distance and directness may be alternatively used, depending on the nature of mismatch or the extent of gulf in the stages of agent action. It would be subsequently shown in the later sections how these concepts are encoded into the profile of agents and how the concepts are important to derive heuristic mechanisms for agents' reasoning.

Applying the Gulfs of Execution and Evaluation in Multi-Agent Environment

In order to identify gaps that separate agents mental states from execution ones, the agent's gulfs of execution and evaluation must be detailed as below:

Agent's Gulfs of Execution: The gulf of execution arises as a result of cognitive distances of execution between the agent and its interaction with the environment (multi-agent, open). The gulf can be identified by elaborating the cognitive distances of execution – which result due to the difference or mismatch between the intentions and the allowable, available actions, capabilities and affordances in its environment. One indicator representing this gulf is to determine how well the agent is able to do the intended actions directly, without extra effort: is the agent able to fulfil its goals delegated to it by humans? Do the actions provided by the agent match those intended by the person? Does the environment/infrastructure (internal infrastructure and/or external environment) provide affordances that allow for the intentions of the agent? If there is a limitation, then there is a gulf of execution in the state of agent's interaction which must be bridged, either with collaborative effort with other agents in the environment or eventually by the human, who would incur cognitive overload or processing. The ultimate aim is to design agents that are able to readily identify and bridge/overcome such a gulf to efficiently achieve the goals set up for them.

Agent's Gulfs of Evaluation: The gulf of evaluation arises as a result of cognitive distances of evaluation between the agent and its interaction with the environment (multi-agent, open); the possible mismatches between agents reasoning capabilities and its representation mechanisms. The Gulf of Evaluation reflects the amount of effort that the agent must exert to interpret the state of its interaction with the system and to determine how well the expectations and intentions have been met. The gulf is smaller when the system provides information about its state in a form that is easy to get, is easy to interpret and matches the manner in which agents' affordances allows these to be perceived.

Bridging the Gulfs and the Cognitive Distances

Directness is inversely proportional to the amount of cognitive effort/processing it takes to manipulate and evaluate the state of interaction with MAS Environment and, moreover, the cognitive effort is a direct result of the gulfs of execution and evaluation. The better the architecture and the environment of the agent helps the agent bridge the gulfs, the less cognitive processing needed and the more directed the resulting interaction. For this, the architecture should facilitate some means of identify, through some indicators, the possible distances and also identify the capabilities and affordances available and not-available corresponding to these indicators.

Agents' Capabilities and Affordances

The concept of affordance is here explored and exploited. It is believed that an affordance inspired agent architecture and environment will help to bridge the gulfs of execution and evaluation. By interfacing perception and action in terms of capabilities and affordances for agents, the aim is to provide a new way for reasoning about agent's capacities and bring about cognitive awareness amongst the agents about each others capacities and constraints, when interacting in a collaborative environment. In Cognitive Science, an affordance is a resource or support that the environment offers an agent for action, and that the agent can directly perceive and employ (Gibson, 1979). Although, this concept has only rarely been used in Semantic Web agent architectures, it offers an original perspective on coupling perception, action and reasoning, differing notably from standard reactive and hybrid architectures. Taking it literally as a means or a metaphor for coupling perception and action directly, the potential that affordances offer for designing new powerful and intuitive agent-based Semantic Web architectures is obvious (Vugt et al. 2006).

The term affordance is adapted for COMMAA to refer to the agent's capacity of action. At each stage of the agent action cycle the agent's corresponding affordance is determined and accounted for. Any constraint in any of the agent's capacities, being the ability to achieve the goal, to formulate the intention or to interpret the consequences of its action, may result in a cognitive distance being introduced. Usually the term affordance is linked to a machine or an application (Norman 1992). In the context of this research affordance is the ability of one agent (the sender) to behave in a way that the other agents in the MAS environment(the receiver) can understand, such that they both have a shared mental model and can trigger, complement or facilitate each other's action/behaviour; therefore these affordances need to be ensured. The affordances are provided by the cognitive artefacts that form part of the agent's internal infrastructure/architecture and its external environment.

Affordances are the opposite concept to distances, so they are complimentary. For instance, an agent's affordance for intention formation will ensure there is no semantic distance of execution and a semantic distance of execution will mean that agent has no affordance for forming an intention. The psychologist Gibson was the first to frame affordances as unified relations between the environment and an actor (Gibson, 1979, p. 127). Affordances can be explained as action possibilities that actors have in the environment. That is, an affordance exists relative to (1) properties of the environment and (2) the action capabilities of an actor (McGrenere and Ho, 2000). For example, a chair has the affordance of 'sitting', because of its shape, height and carrying capacity and because of the humans' ability to sit, the length of their legs, and their weight. The concept of affordances is of particular interest in the field of HCI, which primarily concerned is studying how properties of computers (the environment) and humans (actors) influence their interaction with each other.

Extending the affordance concept to the Semantic Web, it is believed that by semantically identifying and encoding the affordances of an agent will help achieve cognitive directness of interactions amongst the agents. This will contribute to making agents more interoperable in open and heterogenous environments. Some considerations that must be taken into account include the relationship of goals with respect to agents. Goals are central in affordance evaluations. It is important to understand that an affordance does not change as the needs and goals of the person change (McGrenere & Ho, 2000 interpreting Gibson, 1979). Similarly, for an agent, an affordance must be identified irrespective of what the agent's eventual goals are. E.g. ff an agent affords Message translation from FIPA-ACL into FIPA-SL, it is independent of whether another agent will eventually participate in such interaction where this translation is required. However, agent's actions do depend on the goal context. Agents will typically need to act within the environment (they use an affordance) because of a goal they want to achieve for example, performing a task (Vugt et al. 2006, Kakazu & Hakura 1996).

Semantic and Articulatory Distances in Multi-Agent Interaction Model

Following is a detailed account of how may the components of cognitive distance namely: the Semantic and the Articulatory distances occur in the execution and the evaluation cycles. It is also important to illustrate how the cognitive distances are identified with the help of agents' capabilities and affordances. A conceptual high-level view is shown in Figure 5. Identifying these distances, along with agents' capabilities and affordances is of primary importance in highlighting the possible gulfs of evaluation and execution.

- **Semantic distance:** The degree to which the semantic concepts used by the agent are (1) compatible with those of the other agents (including humans if the interaction is being carried out with a human) and, (2) can be used to easily accomplish the agent's goals
- **Articulatory distance:** The degree to which the form of communication between an agent and its environment reflects the application objects and tasks involved

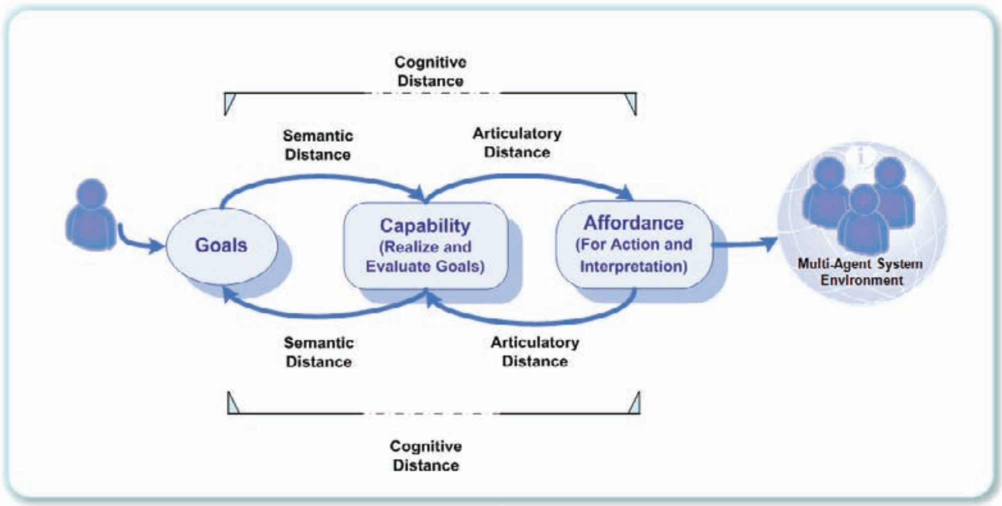
Articulatory distance concerns the actual form that communication takes between say two agents; for example, the choice of Message Encoding, Message Content, Content Type, Interaction Protocol used/employed for the communication. A small articulatory distance results when the input techniques and output representations used are well suited to conveying the required information. Articulatory distance is also decreased when the form of inputs and outputs relate to the semantic concepts used of the underlying conceptual model.

Semantic distance involves the capability of agents required to express desired actions within the concepts of the system. It deals with the possibility to express the concepts of interest concisely using the available capabilities. Semantic distance also involves a measure of how closely the agent's conception of the task domain matches that of the environment. The two distances that compose the Cognitive Distance in Agent's Behaviour are the Semantic Distance and Articulatory Distance. These can be considered as subsets of each of the gulfs, but in reference to input behaviour (initiator's behaviour in communication/interaction) and output behaviour (Respondent's behaviour in communication/Interaction). These are illustrated and discussed in depth below.

An Illustration of Semantic Distance of Execution

Semantic Distance relates to the relationship between an agents intentions and the meaning of expressions, required to convey the agents'

Figure 5. Components of cognitive distances in multi-agent collaboration



intention such that its meaning is interpretable by the intended recipients in the environment. Semantic distance is related to the 'nouns' and 'verbs' or 'objects' and 'actions' provided by an agent's infrastructure and its environment. For execution, forming an intention is the activity that spans semantic distance. The intention specifies the meaning of the input expression that is to satisfy or reach the' agents goal or sub-goal.

If *semantic* indirectness of *execution* existed, agents would not be able to express their intentions directly, or at all in order to achieve the goal delegated to them. Lack of capabilities or affordances would be indicators of this condition. Agent is programmed at lower level of functionality then desired by the human user. The agent or the human may need to carry out more actions then would be expected to accomplish the same goal or intention. E.g. Agent may need to collaborate or request another agent to achieve the goal on its behalf. This requires agent to have desired affordances and capabilities defined. The semantic distance of execution is illustrated in more detail in Figure 6.

An Illustration of Articulatory Distance of Execution

Whereas semantic distance relates to relationships between Agents' formulated intentions and the meanings of expressions/functional capabilities available, articulatory distance in an agent context is defined to relate to relationships between the meanings of expressions/capability and the agent has the affordance to realise the capability in action, or whether agent affords the actions that are essential in achieving the desired intentions. A mismatch of such a form will cause an articulatory distance of execution to exist. The articulatory distance of execution is illustrated in more detail in Figure 7.

An Illustration of Semantic Distance of Evaluation

Semantic distance also occurs on the evaluation side of the interaction cycle. Semantic distance of evaluation is proportional to the amount of processing required by the agent to determine whether the goal has been achieved. The semantic

Figure 6. An illustration of agent's semantic distance of execution

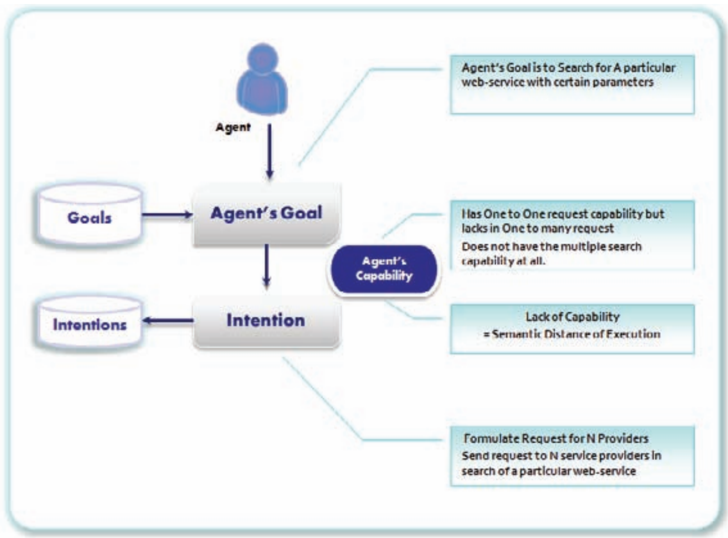


Figure 7. An illustration of agent's articulatory distance of execution

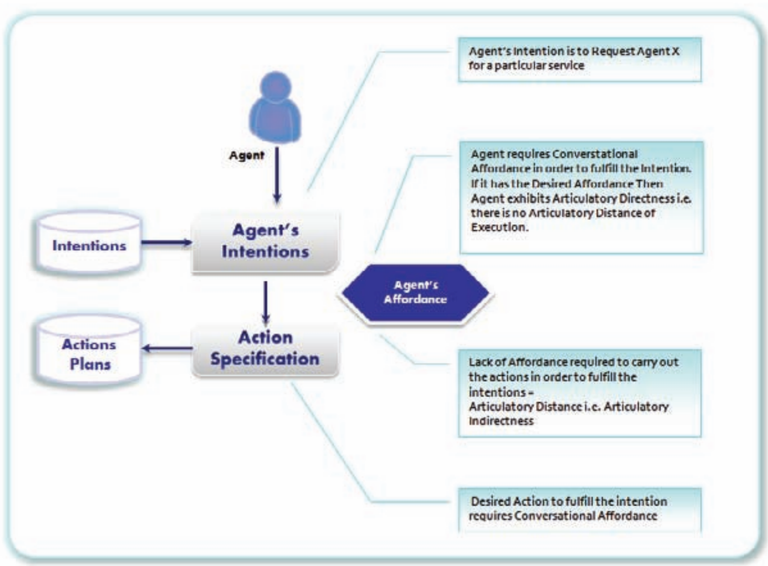
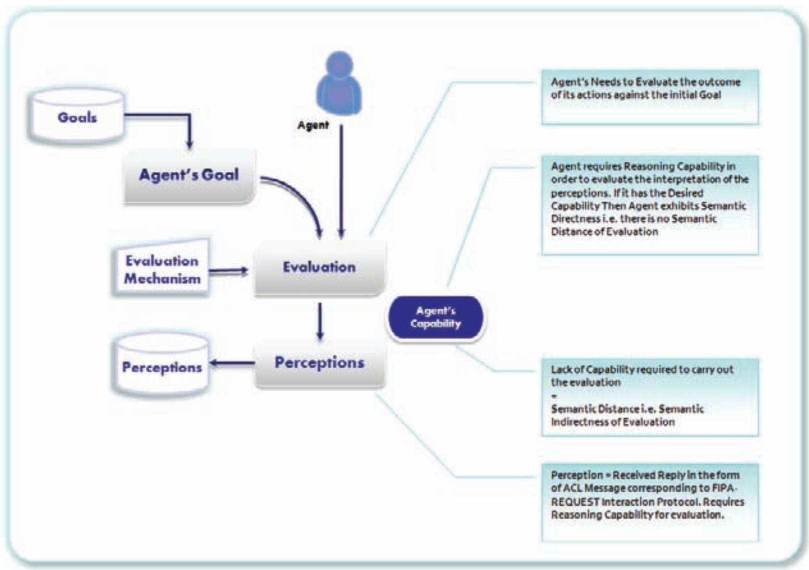


Figure 8. An illustration of agent's semantic distance of evaluation



distance of evaluation is illustrated in more detail in Figure 8.

An Illustration of Articulatory Distance of Evaluation

Articulatory distance or indirectness of evaluation would be indicated by errors in interpretation and having to take extra actions to correctly interpret the state of communication or the result of interaction. An agent's inherent articulateness is closely tied to its level of technology. The articulatory distance of evaluation is illustrated in more detail in Figure 9.

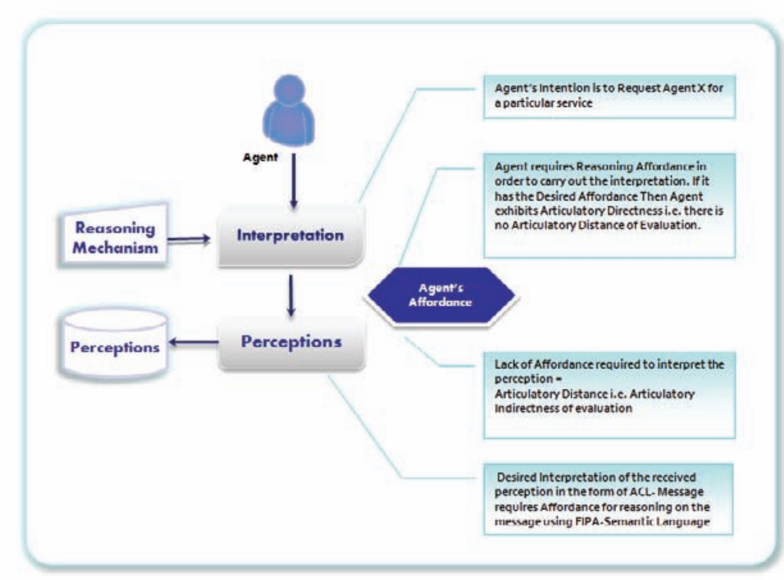
DESIGN AND IMPLEMENTATION OF COMMAA INSPIRED AGENT ARCHITECTURE

Design Goals for Cognitive Profiling of Agents

The theoretical foundations of Cognitive Models such as the Action Cycle (AC) bring about important implications in the current Semantic Web architectures. Traditional Web-service agent architectures only allow agents to discover about each others services. However this research claims that an architecture that is inspired from cognitive models would allow the agents to develop a cognitive awareness about each other which could bring to a more effective MAS environment. To validate this claim, a lower-level classification of design goals which provide the basis upon which the Cognitive Profiling Architecture is devised upon includes: (a) Enhanced Negotiation and Collaboration based on Cognitive Awareness (b) Flexibility and Reusability (c)Adaptive Interaction and Interoperability (d)Discovery based on Heuristic reasoning and (e) Minimization of cognitive load on humans.

In an attempt to realise the above design goals, which are direct implications of COMMAA, some important considerations are taken into account. Firstly, some mechanism is needed that enables the agents to discover and find out about each others cognitive distances, semantic and articulatory dispositions, capabilities, affordances and constraints. Secondly, dynamic and built in

Figure 9. An illustration of agent's articulatory distance of evaluation



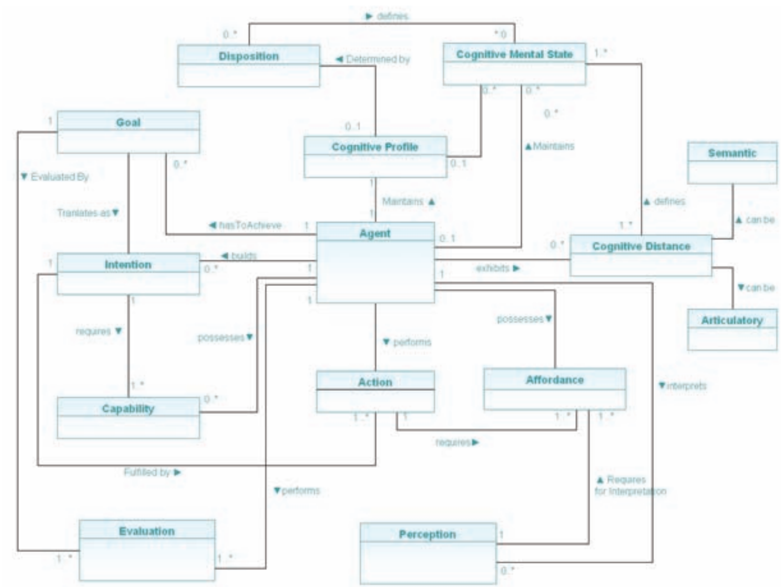
mechanisms are needed for heuristic reasoning and invoking/developing learning, adaptive measures for these constraints and distances to be bridged. These considerations are considered rudimentary to enable agents to be aware of these limitations and Gulfs that may limit their functionality or the extent of services they can provide. As minimal architectural consequences, the elements necessary in COMMAA inspired semantic-Web agent architecture are implemented in a cognitive profiling architecture described in Basharat and Spinelli, (2008b.). The Cognitive Profile of Agents is detailed next.

A Semantic Representation Model for Cognitive Profile of Agents

A possible way of implementing the cognitive profile is by associating an ontology with an agent i.e. agent is given knowledge about its constraints and affordances, its semantic and articulatory disposition. At the heart of the framework is the adoption of ontology to drive the cognitive profile of agents. From a philosophical perspective ontology can be defined as a set of things whose existence is acknowledged by a particular theory or system (Honderich, 1995 cited in Bell et al. 2007). Such 'things' include both types (such as the class of Agents) and individual elements (such as the agent TravelAgent). The adoption of such a definition is important because, when compared with more computationally orientated definitions of ontology (for example, (Gruber 1993); p.1) states that "an ontology is a specification of a conceptualisation"), there is an explicit reference to a system's ontic commitment (i.e., things whose existence is acknowledged or recognised). This leads to representations that are more closely mapped to real world objects.

The use of ontology is also prospective in the Semantic Web action space, since the emerging standards enable reasoning to be carried out effectively on such models. An ontology based reasoner could be invoked on the ontology to carryout reasoning using the heuristic reasoning rules defined; agent could then be made to reason about its state of processing according to CogAC. Architectures that facilitate the sharing of ontologies would enable the agents to discover each other cognitive dispositions thus improv-

Figure 10. Conceptual Model of Ontological Knowledge Model for Cognitive Profile of Agents



ing the manner in which they interact. The issue at stake is to be able to represent the cognitive profile adequately and in a manner that can be shared among agents. The Cognitive profile is therefore implemented as an Ontology in OWL, with OWL-DL (Description Logics) as basis of representation of the profile parameters and properties. Being an emergent standard, OWL-DL ensures that the model caters for a more open community. Ontologies have been recognised by the research community as a – model of expressing the knowledge model for agents e.g. by the recent research of (Laclavík et al. 2006).

Conceptual Model for Agents' Cognitive Profile Ontological Model

Figure 10 shows a generic conceptual model, with objects and properties (shown by labels on associations between concepts) of the ontological model to be implemented in order to enable shared cognitive profiling of agents.

Cognitive Profile of Agent

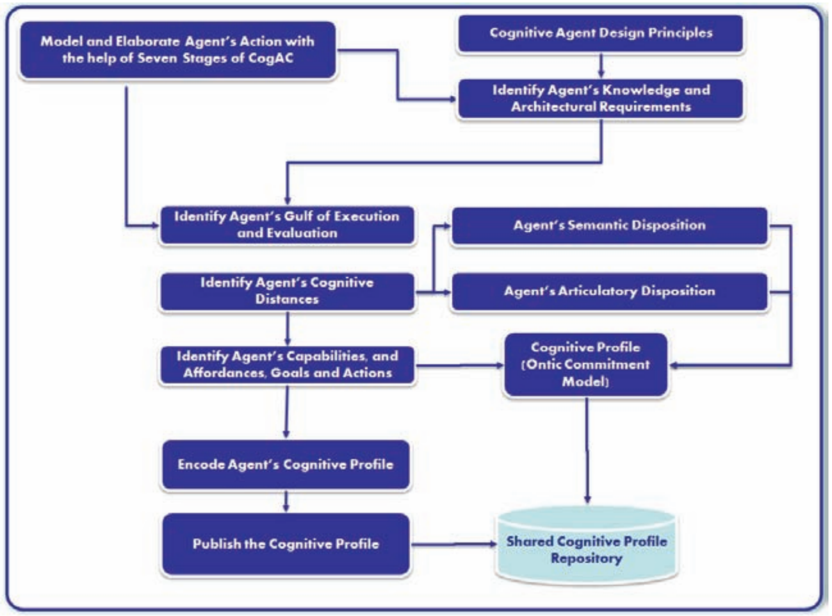
The framework's correct implementation calls for maintaining a cognitive profile of Semantic Web agent. Following are the elements to be maintained in the cognitive profile:

- Agents Cognitive Mental States
- Capabilities
- Affordances
- Semantic Disposition (Semantic Distances of Evaluation and Execution)
- Articulatory Disposition (Articulatory Distances of Evaluation and Execution)
- Goals, Intentions, Perceptions, Evaluations

Cognitive Mental State of an Agent

An agent modelled along the lines of the CogAC requires various cognitive mental states corresponding to the various steps in the execution and

Figure 11. Conceptual Flow of Activities to develop an application according to the guidelines of COMMAA



evaluation stages of the action cycle. These mental states of the agent facilitate the agent's behaviour in both its execution and evaluation cycles. It provides provision for state modelling, representation and tracking an agent's state of execution. In addition representing these states agents may have the dynamic capability of identifying these distances at runtime and may change and update their profile dynamically. The conceptual model is developed using the Protégé Ontology editor to generate OWL Ontology.

Modelling Agents Using COMMAA

In order to utilise COMMAA in practice to process and analyse the interactions and tasks of an agent, the process shown in Figure 11 is used during the design and development process of an agent-based application. The figure also summarises the components and principles that contribute to the agents state modelling, based on the principles of COMMAA. The process is described in detail in Table 3.

Abstract Heuristic Reasoning Mechanism for Agents

To bridge the gulfs of execution and evaluation, agent must have some heuristic reasoning mechanism built into its architecture, such that given a shared cognitive profile is available, it should be able to reason on the knowledge present in it to aid the agent's processing and help bridge the gulfs of execution and evaluation through collaboration or other means. The important considerations are with regards to the representation, discovery and reasoning of the cognitive distances. Rules generate advice by defining the combination of agent knowledge, action stages, distances, abilities/capabilities, and affordances, typically with the generic format shown in Figure 12.

Heuristic Reasoning Mechanism for Agents Execution Stage

The encoding of agents cognitive distances requires a reasoning mechanism. The reasoning is carried out based on the Semantic and Articulatory

Table 3 . Description of Activities in processing agents' action through COMMAA with input and output artefacts

Activities	Description	Input Artefact	Output Artefacts
Cognitive Modeling of Agent Behaviour	▪ Model Agent's Action using the CogAC ▪ The agent's behaviour is programmed according to the stages described in the CogAC ▪ Elaborate Agent's Action with the help of Seven Stages of CogAC ▪ Identify agent's knowledge requirements and elements ▪ As the knowledge elements needed by the agent are identified, it is ensured that the architecture of the agent provides for these knowledge elements to ensure a successful completion of the CogAC ▪ Program agent's behaviour according to available architecture	▪ Agents Functional Requirements ▪ Agents Knowledge and Representational Constraints ▪ Cognitive Agent Design Principles. The CADPs are used as heuristics to evaluate the design at each stage	▪ Knowledge Requirements
Scoping the Agent's Cognitive Profile	▪ Identify Agents capabilities capacities and affordances ▪ Identify agents cognitive distances ▪ Identify known possible agent's gulf of execution and evaluation ▪ Each of the components of the Agent Profile are interpreted such that they represent its Ontic Committment	▪ Knowledge Requirements ▪ Agents Functional Requirements ▪ Agents Knowledge and Representational Constraints	▪ Individual Agent Cognitive Profile (Ontic Commitment Model)
Encoding of Cognitive Profile	▪ Encode Agent's Cognitive Profile parameters in the ontology ▪ The Disposition of Agent (Semantic and Articulatory) in both stages of execution and evaluation is encoded at design time	▪ Profile Parameters ▪ Agents Semantic Disposition ▪ Agent's Articulatory Disposition	▪ Cognitive Profile ▪ (Ontic Commitment Model)
Sharing the Cognitive Profile	▪ Registering or Publishing the profile	▪ Cognitive Profile	▪ Shared Cognitive Profile Repository populated with Cognitive Profile of Agent

disposition of agents encoded in their Cognitive Profile. A reasoning mechanism for the execution stage in the form of pseudo-code is given in Figure 13. This is generic given that the action cycle is applied in a generic context. It may be specialised according to the agent application being developed.

Heuristic Reasoning Mechanism for Agent Evaluation Stage

A reasoning mechanism for the evaluation stage in the form of pseudo-code is given in Figure 14.

Figure 12. Generic Format for Rule-based reasoning of agents' cognitive mental states

```
At Interaction Stage (Execution of Evaluation)
Manipulate Cognitive Profile
If Agent capability and Affordance leads to Directness of Execution
            {
            Action will proceed //i.e. No Gulf Exists
            }
Else Cognitive Distance //i.e. Gulf Exists which needs to be bridged
            {
            Goal Revised
            New Interaction needs to begin
            }
```

Figure 13. Abstract Mechanism for Heuristic Reasoning of the Cognitive Profile Model (Execution Stages)

```
startCogAC(Goal goal){
While (!goalFulfilled)
          {
           - Assert Cognitive Profile
          - Reason
          //Analyze semantic profile
          if (hasSemanticDistance)
          {
          reviseGoal(g);
          startCogAC(g);
          }
          else {
                    IntentionSet  is = getIntentionFromGoal(Goal g);
                    //Analyze Articulatory Dispostion of Execution
                    if(hasArticulatoryDistance)
                    {
                    attempt to overcome/bridge using available affordances
                              else
                    search Shared CognitiveModel
                    }
          }
}
```

Recursive nature of Cognitive Agent Action Cycle

CogAC, when applied in practice is essentially recursive in nature. A high level Goal can be functionally decomposed into sub-goals. Or, an action specification as a result of intention formation to achieve a certain goal may in turn specify other sub-goals to be achieved, for which the CogAC may be recursively invoked.The same recursive behaviour is also true in the evaluation stages. The desired interpretation of the received perception or evaluation may in turn require some internal goals to be achieve.

DEMONSTRATION OF AGENTS' ENHANCED COGNITIVE CAPABILITIES

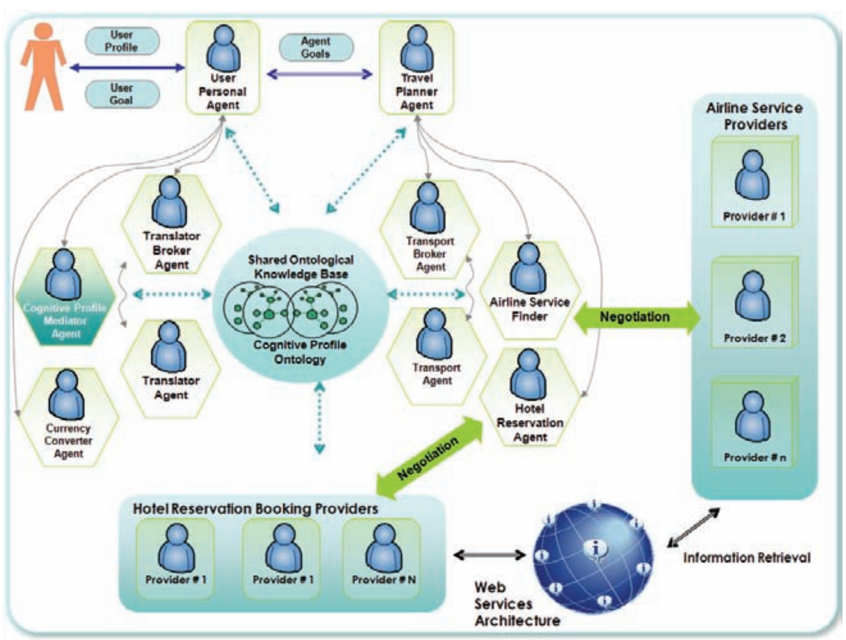
In order to give a flavour of how the framework presented above enhances the agents' cognitive capabilities by imparting improved cognition, the framework was applied to a simulated multi-agent based distributed collaborative application with the aim of testing, improving and evaluating the framework. The purpose of this distributed collaborative multi-agent application is as follows:

- To show how the framework is applied to design of MAS based applications operable on the Semantic Web
- To show how application of COMMAA helps build the cognitive profile of the Agent
- Show how sharing the cognitive Profile improves the collaboration between Agents

Figure 14. Abstract Mechanism for Heuristic Reasoning of the Cognitive Profile Model (Evaluation-Sub Stages)

```
Response r = getSystemResponse();
PerceptionSet P = convertResponseToPerceptions();
//assert Perceptions
//Perform Cognitive Profile Check;
If(hasArticulatoryDistance)
            {
            attempBridgeGulfEvaluation (P);
            }
            Else{
                        interpretPerception(P);
                        if(hasSemanticDistance){
                        Goal g = reviseGoal(P);
                        g.start();
                        } else
                        evaluateGoal(G);
            }
```

Figure 15. Cognitively Modelled Agent-based Travel Planning Scenario Modelled according to COM-MAA



High-Level Architecture of Distributed Collaborative Multi-Agent Application

Travel Planning Scenario

The high-level architecture of the travel planning scenario developed to demonstrate the enhanced cognitive abilities of agents is shown in Figure 15) is a customised adaptation from the vision of travel planning agents presented by Hendler (1999).

The top level functional goals of the demonstration application are as follows: The Multi-Agent based application is aimed to use cognitively modelled agents to solve travel problems given by a user. The user can propose to the user Agent his desired travel, and it will obtain a complete plan that inclu des information about transport, lodging, etc. The agents will Extract, filter and store information automatically from the Semantic Web using other agents. The system aims to use the same information that the user could find if he wish planning the travel himself. Cognitive Sharing of different kinds of abilities is to be demonstrated to gain efficiency in the problem solving task. The agents are simulated to reuse each others capacities, behaviours and offer affordances to each other. Agents closely work according to the user's characteristics, and functions based on the ultimate goals obtained from the user profile and adapt their functional behaviour according to the learned user preferences.

Roles Defined for Agents

Table 4 shows the Roles and Responsibilities defined for Agents involved in Travel Planning Collaboration Scenario. The Collaborative scenario aims integrate the abilities of a set of heterogenous agents. The system is made by a set of agents that can communicate and cooperate among them to reach the problem solution. All the agents in the application use FIPA Based Agent Communication Language for standardization purposes.

Cognitive Profiles of Agents in TPA

The cognitive profile ontology is central to representing the knowledge for agents. The elements of the cognitive profile ontology serve to represent the shared knowledge base of agents through which agents' cognitive awareness will be enhanced. The cognitive profiles are designed to simulate an environment such that some agents are limited in certain capabilities, while others are equipped with them in a complementary manner to facilitate interoperability, resuse and adaptive collaboration based on enhanced cognitive awareness. The cognitive profiles of the agents in the prototype application are shown in Table 5.

Reuse Mechanisms Employed

The Behaviour API and Interaction Protocol API of JADE (Bellifemine et. Al, 2001; 1999; JADE 2004) are used to model the Agent's Actions, or an action Plan. ACL Messages are used as representations for Intentions and perception. An agent is simulated in such a way that it is given runtime capability to change its behaviour, and dynamically change profile so as to demonstrate the power and potential of the Ontological Model.

Cognitive Modelling of Agent Interactions and Communication Scenarios Through CogAC

Table 6 shows an example of how the agents collaboration is modelled through the CogAC.

Illustration of Improved Cognitive Awareness

The provision of cognitive profile as shared knowledge base serves as means to increase the cognitive awareness for agents since they can not only reason about their own cognitive distances, they can also access and query other agents' cog-

Table 4. Roles and Responsibilities defined for Agents involved in Travel Planning Collaboration Scenario

Agent	Functional Role of Agent in the Application
UserAgent	• This agent handles a user query and shows him the solution. To do so, it analyzes the problem and obtains an abstract representation. Subsequently it requests a Travel Planner Agent solutions to that problem. The User Agent has different skills like communication with Travel Planner Agents and users, or learning the user's profiles necessary to customize the system answer. The User Agent has a set of interfaces to allow input and output information and the user evaluation of the solutions found.
Travel Planner Agent	• The main Travel Planner Agent's goal is reason about User Agents and other Travel Planner Agents problems, and find out a set of possible solutions. Travel Planner Agents have different skills like communication (with different agents in the system), planning (its main reasoning module) and learning.
Transport Broker Agent	• This is a mediator agent, who has the ability of directly communicating with a transport agent and may provide services to some other agent which may not be able to directly communicate with a transport agent thus acting as an intermediary.
Transport Agent	• This agent has the capability of finding desired transport options from the web service providers and provides this information to other agents.
Translator Broker Agent	• This is a mediator agent, who has the ability of directly communicating with a translator agent and may provide services to some other agent which may not be able to directly communicate with a translator agent thus acting as an intermediary.
Translator Agent	• This agent simulates a functionally computational agent on the semantic web which has the capability of performing translation services upon request from one language to another e.g. English to Italian.
Currency Converter Agent	• The role of this agent is to retrieve up to date currency conversion rates from the web and provide currency conversion services to other agents upon request.
Airline Service Providers	•Airline Service Provider is responsible for finding information relation to airline reservations.
Accommodation Service Providers	• These are agents belonging to organization such as Hotels etc. which maintain up to date information about the accommodation availability, rates etc, and can carry out automated bookings on behalf of their owners.
Cognitive Broker Agent	• This agent also plays are central role in the collaboration scenario by acting as an indexer of facilitator for other agents to access other agent's cognitive profile.

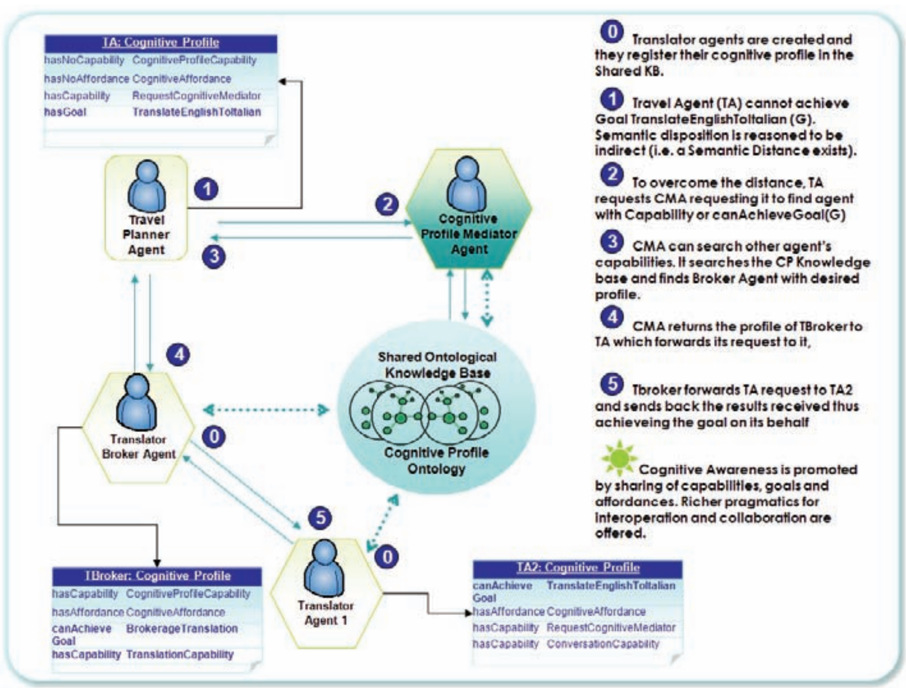
Table 5. Capabilities and affordances defined for agents in the Travel Planner Application

Agent	Goals and Capabilities	Behaviours (Actions)	Affordances
UserAgent/ TravelPlanner Agent	•CommunicationCapability •CollaborationCapability •ConversationalCapability **•canAchieveGoal (GetTravelDetails)**	•SendMessage/ReceiveMessage •RequestInteractionProtocol •InteractionProtocol-FIPA-Request •InteractionProtocol-FIPA-ContractNet	**•hasNO OntologicalAffordance** **•hasNO ComputationalAffordance** **•hasNO CognitiveProfileAffordance**
TranslatorBroker Agent	•ConversationalCapability **•canAchieveGoal (BrokerageTranslation)**	•InteractionProtocol - BrokerageProtocol	•ComputationalAffordance •CognitiveProfileAffordance
CognitiveMediator Agent	• ConversationalCapability **•canAchieveGoal** **•(SearchCognitiveProfile)**	•InteractionProtocol-FIPA-Request(Responder)	•ComputationalAffordance •CognitiveProfileAffordance
TransportBroker Agent	•BrokerageCapability **•canAchieveGoal (BrokerageTransport)**	•InteractionProtocol - BrokerageProtocol	•ComputationalAffordance •CognitiveProfileAffordance
TranslatorAgent	•ComputationalCapability **•canAchieveGoal (TranslateEnglishToFrench)**	•InteractionProtocol-FIPA-Request(Responder)	•ComputationalAffordance •CognitiveProfileAffordance
Currency ConvertorAgent	•ComputationalCapability •ConversationalCapability **• canAchieveGoal (ConvertPoundToDollar)**	•InteractionProtocol-FIPA-Request(Responder)	•ComputationalAffordance •CognitiveProfileAffordance
TransportAgent	•ComputationalCapability •ConversationalCapability **•canAchieveGoal (TranslateEnglishToFrench)**	•InteractionProtocol-FIPA-Request(Responder)	•ComputationalAffordance •CognitiveProfileAffordance
Airline/Hotel Service Provider Agent	•ConversationalCapability **•canAchieveGoal(ProvideService)**	•InteractionProtocol-FIPA-ContractNet (Initiator)	•ComputationalAffordance •CognitiveProfileAffordance
Airline/Hotel Service Finder Agent	•ComputationalCapability •ConversationalCapability **•canAchieveGoal(FindService)**	•InteractionProtocol-FIPA-ContractNet (Responder)	•ComputationalAffordance •CognitiveProfileAffordance

Table 6. Cognitively Modelled Agent Communication Scenario for Travel Planner Agent

User Profile and Travel Planner Agent			
Stage	**Description**	**Capabilities/ Affordances/Behaviours**	**Inferred Distances Rules for Inference**
Goal	• **hasGoal** **Convert Currency Pounds to Dollars (G)**	• G requires Capability(C) • C1=ConversationalCapability • C2=ComputationalCapability • G requires Affordance (A)	If (canAchieveGoal (G)) - No Cognitive Distance
Agent's Gulf of Execution			
Intention	• If (hasNoSemanticDistance Stage(Execution)) Proceed (Go to Next Step) • Else ReviseGoal **FindAgent (canAchieveGoal, G)**	• **hasNoCapability(C)**	• **Semantic Distance** G requiresCapability C Agent hasNoCapability C Therefore hasSemanticDistance
Action Specification	• If (hasNoArticulatoryDistance & Stage(Execution)) Proceed (Go to Next Step) • Else ReviseGoal **FindAgent (canAchieveGoal, G)**	• **hasNoAffordance(A)**	• **Articulatory Distance** G requiresAffordance A Agent hasNoAffordance A Therefore hasArticulatoryDistance
Execution	• Performing the actions • Send the Request , Wait for Response		•
Agent's Gulf of Evaluation			
Perception	• Noticing that feedback occurs, if any e.g. Reply Message		
Interpretation	• If (hasNoArticulatoryDistance & Stage(Execution)) Proceed (Go to Next Step) • Else ReviseGoal **FindAgent (canAchieveGoal, G)**	• **hasNoAffordance(A)**	• **Articulatory Distance** G requiresAffordance A Agent hasNoAffordance A Therefore hasArticulatoryDistance
Evaluation	• If (hasNoSemanticDistance & Stage(Execution)) Proceed (Go to Next Step) • Else ReviseGoal **FindAgent (canAchieveGoal, G)**	• **hasNoCapability(C)**	• **Semantic Distance** G requiresCapability C Agent hasNoCapability C Therefore hasSemanticDistance

Figure 16. Agent Collaboration Scenario to Illustrate Improved Cognitive Awareness



nitive profiles allowing them to adaptively refine their interactions in attempt to achieve their goals in a collaborative manner. An illustration of how this proves so is shown in Figure 16.

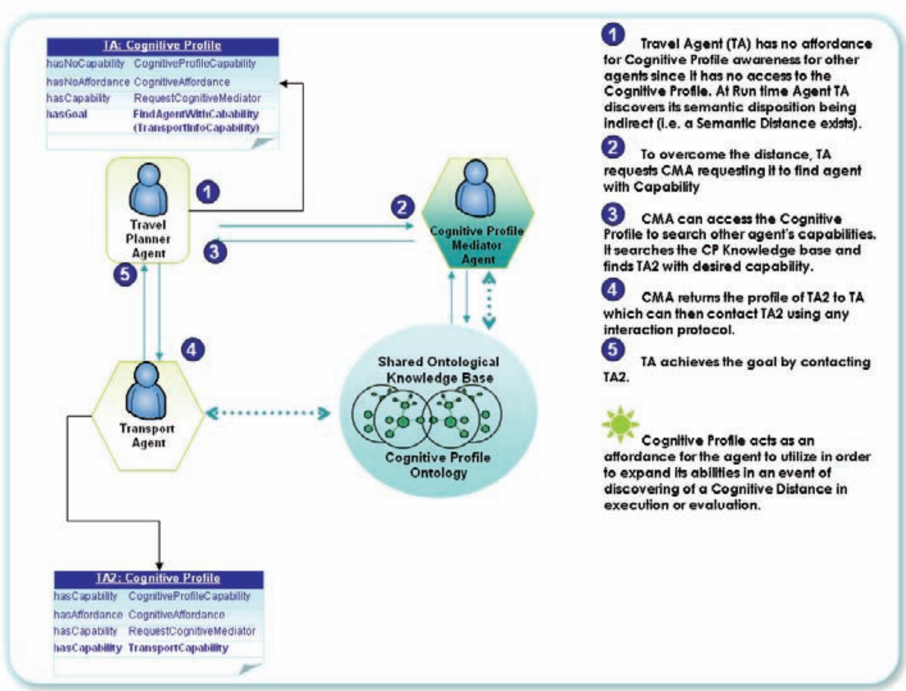
Cognitive Profile as an Affordance

The cognitive profile in the form of ontology serves as an affordance for the agent to be cognitively aware of its environment and make adaptive decisions about it. Following is a scenario in Figure 17 that illustrates how this proves so.

EVALUATION AND FUTURE TRENDS

The application of the COMMAA to the application was found very useful in structuring the agents' interaction and collaborations in an MAS environment. An attempt to assess the directness of engagement using the distances in the multi-agent interaction scenario offered useful results. As shown by the demonstration that the Cognitive Modelling of agents give a powerful boost and shows great potential in the manner that agent utilise each other's capabilities to support each other functional execution and in return facilitate and reduce cognitive overload of humans. The increased cognitive awareness promotes interoperation amongst agents and results in behaviour sharing and reuse. The high visibility into each other's mental models increases the mapping of agent's functional and execution models. The 7 stages of CogAC ensure that enough feedback is received to ensure successful fulfilment of goals through multi-agent collaboration. The Cognitive Profile Ontological Model serves as an essential artefact for agents which, continually

Figure 17. Multi-Agent Collaboration Scenario to illustrate the use of Cognitive Profile as an affordance



**Further implementation details are beyond the scope of this chapter* provides for both sides of the CogAC: execution and evaluation.

Based on the above evaluation of the application, The CogAC can come to be considered as a fundamental part of the functioning of agent's actions in interaction with its MAS environment. Thus agent modelled along lines of COMMAA exhibit strong principles of user-centred design and advocate ease of use, efficiency and reuse and interoperation. Although, the model is used to model software agents, similar model can be used for classifying user roles and capabilities and for maintaining and sharing user profiles and roles. The proof-of-concept application has also demonstrated the feasibility of implementing the constructs of COMMAA using the combination of Multi-Agent Platform and the Semantic Web middleware.

In addition, the COMMAA takes a holistic view of agent's action and its processing in the stages of execution and evaluation. It increases the cognitive awareness amongst agents by elabo-

Table 7. Improvements achieved as a result of application of COMMAA

Design Goals	Evaluation
Cognitive Awareness:	▪ Agents profile has shown to be encoded using OWL ontology and published. Thus Agents are able to cognitively describe, publish and access each others capabilities, affordances and distances/constraints by Quering the ontology using SPARQL based mechanism. More dynamic and adaptive behaviour of agents has been shown. Thus the implementation shows cognitively aware modelling agents interaction. ▪ Agents can carry out cognitively aware communication and collaboration with each other ▪ Through Cognitive awareness, agents help each other identify and bridge the gulfs of execution and evaluation
Enhanced Negotiation and Collaboration	▪ As shown by the collaboration scenarios, agents negotiation abilities are enhanced as a result of improved cognitive awareness
Flexibility and Reusability:	▪ As shown, through cognitive profiling, JADE Behaviour API and Interaction protocols have been used with much more flexibility and their reuse is promoted by agents sharing cognitively their goals and abilities. ▪ Agents are equipped with reasoning mechanism to dynamically reason about their semantic and articulatory disposition. This allows them to adapt to the required interaction scenario, thus providing intrinsic support for more flexible interaction.
Adaptive Interaction and Interoperability:	▪ Heterogenous agents in different roles including Transport Agents, Transport and Translator Brokers, Currency Convertors etc. have been shown to participate in adaptive collaborative scenarios through dynamic sharing of their cognitive profiles in order to help each other bridge cognitive distances of execution and evaluation. In doing so, they help each other achieve their goals. ▪ Interoperability is promoted through the sharing of cognitive profiles, dynamic publish and access mechanisms. ▪ Reasoning through SWRLJess based rules enhances cognitive awareness, thus improving interoperability.
Discovery based on Heuristic reasoning:	▪ Agents dynamically reason about their profiles using SWRL Rules and DIG reasoners that classify ontology. This allows dynamic discovery of cognitive distances and other agents' profiles improving agents cognitive awareness of itself and the environment.
Minimisation of cognitive load:	▪ User has been shown to be relieved of much cognitive overload since agent through all of the above functionalities is able to perform much more. It is able to meet all the preferences in the user profile. If the case was otherwise, the distribution of tasks would shift from the agent to the user. E.g. if the agent was unable to meet the goal of converting dollar to pound, given it was not able to find any other agent to achieve the goal, the user would have the added cognitive overload of meeting the desired goal on its own. ▪ Thus human information processing has shown to be reduced as a result of increased cognitive awareness and improvement in the flexibility and adaptability of agents' collaborative abilities.

rating the action infrastructure, its limitations, constraints (Distances) and its capabilities and affordances. In this way it facilitates the architecture for the agents that are designed to work or be programmed to work more adaptively on the Semantic Web. This was made sufficiently evident using the implementation of the Travel Planning scenario. Similarly, agents can use the semantic information, share it, yet can communicate using ACL languages, programmed according to the Agent principles and do not need to rely on the Web service interfaces and profiles.

The results of the application are validated against the design goals presented earlier in Table 7.

Future Directions

The abstract architecture opens new doors of research. With the core framework in place the natural next step in its expansion is the specialised enhancement towards a more rigorous definition of different levels and variations of cognitive profile and its parameters. Identifying the best level of detail for functionally decomposing each task or intention and applying it consistently is difficult. The application highlights that in order to make this model fully implementable or workable in the real world, there needs to be taxonomy of distances, directness measures, capabilities and affordances defined. This study initiates this activity by identifying the rudimentary picture of the basic Cognitive profile of agents e.g. at present the capabilities were identified as high level constructs. They can be made much more elaborate e.g. that of OWL-S. However to make it reach such a state of maturity where it could be utilised in practice will take some more effort. A taxonomy could be appealing because it would allow a generic to specific discussion of cognitive distances across different agent applications. Although some subset of cognitive distances will always be generic, it can be suspected that a fairly large subset will need to be specialised across limited applications.

Ontology Learning, Alignment and Mapping: A serious issue in making this model work on a larger scale will be ensuring the standardization for the interoperability. Issues of Ontology learning (Maedche & Staab 2001), Ontology alignment and mapping (as highlighted in works by Laera et al. 2006, Mocan, Cimpian & Kerrigan 2006; Sampson & Lanzenberger 2006) also become important for standardization and homogenization purposes. Standardization of cognitive profile for agents will be another issue foreseen if this model were to work successfully, but with rich semantic model of OWL and RDF will allow for standardization to be achieved. However it provides substantial stimulus for future research.

Enhanced Learning and Reasoning Mechanisms: Furthermore, an idea that will add immense value to the further development of agents' cognitive model is enhanced Reasoning and learning mechanisms. It would also be worth investigating how the principles of COMMAA plays a useful role in Interface characteristics i.e. investigation into role of cognitive models applied to model interface agents and their activities and management of user profiles. Another issue is with respect to the extent to which agent's knowledge model or cognitive profile is to be shared. The notion of Public, Private profile could be considered. The extent of autonomy given to the agent moving towards Autonomous Semantic Web services (Paolucci & Sycara 2003) is also highly relevant.

CONCLUSION

The Semantic Web community has recognised the advantages of an agent-based approach to building deployable solutions in a number of application domains comprising complex, distributed systems. The chapter targeted some of the key challenges faced when developing autonomic and autonomous entities in the domain of Multi-Agent Collaborations. By applying cognitive models to model agents' behaviour on the Semantic Web, a reasonably successful attempt has been made at coupling cognitive science principles of user-centred design with features of agent base systems and architectures.

This research took into account the emergent standards in both agents and Semantic Web in order to render the framework principles presented. The principles and design of COMMAA have also been used to demonstrate agents' improved capabilities through an evaluation of a multi-agent collaborative scenario to illustrate the adaptive coordination of different agents acting as owners in heterogenous and dynamically changing environments. The results of the evaluation show an improved flexibility, interoperability and reusability of agents' collective behaviours and goals. Thus, it establishes COMMAA as a step forward in providing the next generation of Semantic Web, a successful framework of multi-agent collaboration, which is inevitably required for generating robustly engineered agents able to carry out spontaneous and adaptive collaboration based on cognitive awareness of their environment and infrastructure.

REFERENCES

Ahmad, H. F., Suguri, H., Ali, A., Malik, S., Mugal, M., Shafiq, M. O., Tariq, A., & Basharat, A. (2005). Scalable fault tolerant agent grooming environment – SAGE. In F. Dignum, V. Dignum, S. Koenig, et al. (Eds.), *Proceedings of the 4th International Conference on Autonomous Agents and Multi agent Systems, AAMAS 05* (pp. 3-4).

Ankolekar, A., Burstein, M., Hobbs, J., Lassila, O., Martin, D., McIlraith, S., Narayanan, S., Paolucci, M., Payne, T., Sycara, K., & Zeng, H. (2001). DAML-S: Semantic markup for Web services. In *Proceedings of the International Semantic Web Working Symposium (SWWS,* (pp. 411-430).

Arai, S., & Ishida, T. (2004). Learning for human-agent collaboration on the Semantic Web. In *Proceedings of the 12th International Conference on Informatics Research for Development of Knowledge Society Infrastructure (ICKS'04)* (pp. 132-139).

Arai, S., Murakami, Y., Sugimoto, Y., & Ishida, T. (2003). Semantic Web service architecture using multi-agent scenario description. In J. Lee & M. Barley (Eds.), *Proceedings of the 6th Pacific Rim International Workshop on Multi-Agents, PRIMA 2003* (pp. 98-109).

Basharat, A., & Spinelli, G. (2008a). Towards a distributed cognitive view of the agent-mediated Semantic Web. In *Proceedings of the IADIS Intelligent Systems and Agents (ISA 2008).*

Basharat, A., & Spinelli, G. (2008b). Towards engineering ontologies for cognitive profiling of agents on the Semantic Web. In *Proceedings of the IEEE International Workshop on Engineering Semantic Agent Systems (ESAS 2008).*

Berners-Lee,T., Hendler, J. & Lassila, O. (2001). The Semantic Web. *Scientific American*, *279*(5), 34-43.

Bell, D., De Cesare, S., Iacovelli, N., Lycett, M., & Merico, A. (2007). A framework for deriving Semantic Web services. *Information Systems Frontiers, 9*(1), 69-84.

Bellifemine, F., Poggi, A., & Rimassa, G. (2001). Developing multi-agent systems with a FIPA-compliant agent framework. *Software - Practice and Experience, 31*(2), 103-128.

Clark, K. (2001). *The politics of schemas: Part 1.* Retrieved July 20, 2007, from http://www.xml.com/pub/a/2001/01/31/politics.html

Chandrasekharan, S. (2004). Semantic Web: Knowledge representation and affordance. In B. Gorayska & J. L. Mey (Eds.), *Cognition and technology: Co-existence, convergence, co-evolution*. Amsterdam: John Benjamins.

Cuomo, D. (1993). Methodology and encoding scheme for evaluating the usability of graphical, direct manipulation style interfaces. In *Proceedings of the Human Factors and Ergonomics Society, Inc.*, Santa Monica, CA (Vol. 2) (pp.1137-1141).

FIPA. (2007). *IEEE foundation for intelligent physical agents. (FIPA).* Retrieved July 20, 2007, from http://www.fipa.org, 2006

Gibson, J. J. (1977). The theory of affordances. In R. E. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199-220.

Hendler, J. (2001). Agents and the Semantic Web. *IEEE Intelligent Systems and Their Applications, 16*(2), 30-37.

Hendler, J. (1999). Is there an intelligent agent in your future? *Nature.* Retrieved from www.nature.com/nature/Webmatters/agents/agents.html

Hutchins, E. L., Hollan, J. D., & Norman, D. A. (1986). Direct manipulation interfaces. *Human-Computer Interaction, 1*(4), 311-338.

JADE. (2004a). *JADE (java agent development framework).* Retrieved July 20, 2007, from http://jade.cselt.it/

JADE. (2004b). *JADE documentation*. Retrieved July 20, 2007, from http://jade.cselt.it/

Jennings, N. R. (2000). On agent-based software engineering. *Artificial Intelligence, 117,* 277-296.

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems, 19*(6), 91-95.

Laclavík, M., Balogh, Z., Babík, M., & Hluchý, L. (2006). Agentowl: Semantic knowledge model and agent architecture. *Computing and Informatics, 25*(5), 421-439.

Laera, L., Tamma, V., Euzenat, J., Bench-Capon, T., & Payne, T. (2006). Reaching agreement over ontology alignments. In *Proceedings of The Semantic Web – ISWC 2006* (LNCS 4273, pp. 371-384. Berlin: Springer Berlin/Heidelberg.

Louis, V., & Martinez, T. (2005a). The jade semantic agent: Towards agent communication oriented middleware. *AgentLink News Journal, 18,* 16-18.

Louis, V., & Martinez, T. (2005b). An operational model for the FIPA-ACL semantics. In *Proceedings of the AAMAS'05 Workshop on Agent Communication (AC2005),* Utrech, The Netherlands.

Lu, S., Dong, M., & Fotouhi, F. (2002). The Semantic Web: Opportunities and challenges for next-generation Web applications. *Information Research*, 7(4). Retrieved from http://InformationR.net/ir/7-4/paper134.html

Luck, M., McBurney, P., Shehory, O., & Willmott, S. (2005). *Agent technology: Computing as interaction (a roadmap for agent based computing).* AgentLink.

Maedche, A., & Staab, S. (2001). Ontology learning for the Semantic Web. *IEEE Intelligent Systems and Their Applications, 16*(2), 72-79.

Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM, 37*(7), 30-40.

Martin, C., Schreckenghost, D., Bonasso, P., Kortenkamp, D., Milam, T., & Thronesbery, C. (2003a). Aiding collaboration among humans and complex software agents. In *Proceedings of the AAAI, 2003.*

Martin, C., Schreckenghost, D., Bonasso, P., Kortenkamp, D., Milam, T., & Thronesbery, C.

(2003b). An environment for distributed collaboration among humans and software agents. In *Proceedings of the International Conference on Autonomous Agents, 2*(1), 1062-1063.

Martin, D., Burstein, M., McDermott, D., McIlraith, S., Paolucci, M., Sycara, K., McGuinness, D. L., Sirin, E., & Srinivasan, N. (2007). Bringing semantics to Web services with OWL-S. *World Wide Web, 10*(3), 243-277.

Martin, D., Paolucci, M., McIlraith, S., Burstein, M., McDermott, D., McGuinness, D., Parsia, B., Payne, T., Sabou, M., Solanki, M., Srinivasan, N., & Sycara, K. (2005). Bringing semantics to Web services: The OWL-S approach. In *Proceedings of the SWSWPC* (LNCS 3387, pp. 26-42).

Miller, C. A., & Parasuraman, R. (2007). Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. *Human Factors, 49*(1), 57-75.

McGrenere, J., & Ho, W. (2000). Affordances: Clarifying and evolving a concept. In *Proceedings of the Graphics Interface* (pp. 179-186). Mahwah, NJ: Lawrence Erlbaum Associates.

Meredith, P. (1970). Developmental models of cognition. In P. L. Garvin (Ed.), *Cognition: A multiple view.* New York: Spartan Books.

Mocan, A., Cimpian, E., & Kerrigan, M. (2006) Formal model for ontology mapping creation. In *Proceedings of The Semantic Web – ISWC 2006* (LNCS 4273, pp. 459-472).

Neuhold, E. (2003). Semantic Web application modelling. In *Proceedings of the Conceptual Modeling -ER 2003* (LNCS 2813, pp. 1-2).

Nigay, L., & Coutaz, J. (1991). Building user interfaces: Organizing software agents. In *Proceedings ESPRIT'91 Conference* (pp. 707-719).

Norman, D. A. (1994). How might people interact with agents. *Communications of the ACM, 37*(7), 68-71.

Norman, D. A. (1992). Design Principles for cognitive artefacts. *Research in Engineering: Design Theory Applications and Concurrent Engineering, 4*(1), 43-50.

Norman, D. A. (1988). *The design of everyday things.* New York: Doubleday.

Norman, D. A. (1986). Cognitive engineering. In D. A. Norman & S. Draper (Eds.), *User centered system design: New perspectives on human-computer interaction*. Mahwah, NJ: Lawrence Erlbaum Associates

Norman, D. A., & Draper, S. (1986). *User centered system design: New perspectives on human-computer interaction*. Mahwah,NJ: Lawrence Erlbaum Associates

Paolucci, M., & Sycara, K. (2003). Autonomous Semantic Web services. *IEEE Internet Computing, 7*(5), 34-41.

Payne, T. R., Singh, R., & Sycara, K. (2002a). Calendar agents on the Semantic Web. *IEEE Intelligent Systems and Their Applications, 17*(3), 84-86.

Payne, T. R., Singh, R., & Sycara, K. (2002b). RCal: A case study on Semantic Web agents. In C. Castelfranchi & W. L. Johnson (Eds.), *Proceedings of the 1st International Joint Conference on: Autonomous Agents adn Multiagent Systems* (pp. 802-803).

Saja, A. D. (1985). The cognitive model: An approach to designing the human-computer interface. *ACM SIGCHI Bulletin, 16*(3), 36-40.

Sampson, J., & Lanzenberger, M. (2006). Visual ontology alignment for Semantic Web applications. In *Advances in conceptual modeling – theory and practice* (LNCS 4231, pp. 405-414).

Schreckenghost, D., Martin, C., Bonasso, P., Kortenkamp, D., Milam, T., & Thronesbery, C. (2002). Supporting group interaction among humans and autonomous agents. *Connection Science, 14*(4), 361-369.

Scott, R., Roth, E. M., Deutsch, S. E., Malchiodi, E., Kazmierczak, T. E., Eggleston, R. G., Kuper, S. R., & Whitaker, R. D. (2005). Work-centered support systems: A human-centered approach to intelligent system design. *IEEE Intelligent Systems*, *20*(2), 73-81.

Spiekermann, S., & Paraschiv, C. (2002). Motivating human- agent interaction: Transferring insights from behavioral marketing to interface design. *Electronic Commerce Research, 2*(3), 255-285.

Shafiq, M. O., Ali, A., Ahmad, H. F., Suguri, H., Tariq, A., Basharat, A., & Nazir, F. (2005). A distributed services based conference planner application using software agents, Grid services and Web services. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, The Netherlands. New York: ACM.

Sun, R. (2001). Cognitive science meets multi-agent systems: A prolegomenon. *Philosophical Psychology, 14*(1), 1-28.

Sutcliffe, A. (2000). On the effective use and reuse of HCI knowledge. *ACM Transactions on Computer-human Interaction*, *7*(2), 197-221.

Sutcliffe, A., Ryan, M., Doubleday, A., & Springett, M. (2000). Model mismatch analysis: Towards a deeper explanation of users' usability problems. *Behaviour and Information Technology*, *19*(1), 43-55.

Sutcliffe, A. G., & Kaur, K. D. (2000). Evaluating the usability of virtual reality user interfaces. *Behaviour and Information Technology, 19*(6), 415-426.

Sycara, K. P. (1998). Multiagent systems. *AI Magazine, 19*(2), 79-92.

Sycara, K., Pannu, A., Williamson, M., Zeng, D., & Decker, K. (1996). Distributed intelligent agents. *IEEE Expert-Intelligent Systems and their Applications, 11*(6), 36-46.

Sycara, K., Klusch, M., Widoff, S., & Lu, J. (1999). Dynamic service matchmaking among agents in open information environments. *SIGMOD Record (ACM Special Interest Group on Management of Data), 28*(1), 47-53.

Tariq, A., Basharat, A., Ahmad, H. F., Suguri, H., & Ali, A. (2005a). A hybrid agent architecture for modeling autonomous agents in SAGE. In M. Gallagher, J. Hogan, & F. Maire (Eds.), *Proceedings of the 6th International Conference on Intelligent Data Engineering and Automated Learning - IDEAL 2005* (pp. 478-485).

Tariq, A., Basharat, A., Ali, A., Ahmad, H. F., & Suguri, H. (2005b). A collaborative multi-agent based conference planner application using Web and Grid services. In *Proceedings of the 2005 International Conference on Collaborative Computing: Networking, Applications and Worksharing.*

Vugt, H. C., Hoorn, J. F., Konijn, E. A., & de Bie Dimitriadou, A. (2006). Affective affordances: Improving interface character engagement through interaction. *International Journal of Human Computer Studies, 64*(9), 874-888.

Willmott, S. (Ed.). (2003). *Technical input and feedback to FIPA from Agentcities. RTD and the Agentcities initiative* (Agentcities Task Force Technical Note 00003). Retrieved from http://www.agentcities.org/note/00003/

Wooldridge, M. J. (2002). *An introduction to multi-agent systems.* New York: John Wiley & Sons.

Zhang, J., Patel, V. L., Johnson, K. A., Smith, J. W., & Malin, J. (2002). Designing human-centered distributed information systems. *IEEE Intelligent Systems*, *17*(5), 42-47.

KEY TERMS AND DEFINITIONS

Affordance: An affordance is a resource or support that the environment offers an agent for action, and that the agent can directly perceive and employ**.**

Agent: Agents are defined as autonomous, problem-solving computational entities capable of effective operation in dynamic and open environments.

Capability: The functional ability possessed by an agent to achieve some given goal or requirement.

Cognitive Awareness: It refers to the ability of the Web agents to diagnose their processing limitations and to establish interactions with the external environment (in the form of other agents including humans and software agents).

Cognitive Model: Internal representations of the current situation created by either a human and agent to assess their state with respect to the environment.

Cognitive Profile: It is a semantic representation model which includes information about the cognitive states of an agent, its functional capacities and affordances.

COMMAA (Cognitive Model of Multi-Agent Action): A framework for modeling agents' actions and interactions in its environment in an attempt to provide an architecture that improves the flexibility of Multi-agent interaction by promoting cognitive awareness .

Multi-Agent System: Multi-Agent System (MAS) is a distributed collaborative environment which allows a number of agents to cooperate and interact with other agents (including both people and software) that have possibly conflicting aims, in a complex environment.

This work was previously published in Handbook of Research on Social Dimensions of Semantic Technologies and Web Services, edited by M. M. Cruz-Cunha; E. F. Oliveira; A. J. Tavares; L. G. Ferreira, pp. 610-642, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 2.18
Utilisation of Case-Based Reasoning for Semantic Web Services Composition

Taha Osman
Nottingham Trent University, UK

Dhavalkumar Thakker
Nottingham Trent University, UK

David Al-Dabass
Nottingham Trent University, UK

ABSTRACT

With the rapid proliferation of Web services as the medium of choice to securely publish application services beyond the firewall, the importance of accurate, yet flexible matchmaking of similar services gains importance both for the human user and for dynamic composition engines. In this article, we present a novel approach that utilizes the case based reasoning methodology for modelling dynamic Web service discovery and matchmaking, and investigate the use of case adaptation for service composition. Our framework considers Web services execution experiences in the decision making process and is highly adaptable to the service requester constraints. The framework also utilizes OWL semantic descriptions extensively for implementing both the components of the CBR engine and the matchmaking profile of the Web services.

INTRODUCTION

The Internet has become the market-place for a colossal variety of information, recreational and business services. Web services are increasingly becoming the implementation platform of choice to securely expose services beyond the firewall. Moreover, multiple Web services can be integrated either to provide a new, value-added service to the end-user or to facilitate co-operation between various business partners. This integration of Web services is called "Web services composition" and is feasible to achieve because of the Web services advantages of being platform, language neutral and loosely coupled.

Automatic Web service discovery and matchmaking is the principal aspect for dynamic services composition. The accuracy of the matchmaking (selection) process enhances the possibility of successful composition, eventually satisfying the user and application requirements. The current standard for Web service discovery, the Universal Description, Discovery and Integration (UDDI) registry is syntactical and has no scope for automatic discovery of Web services. Hence, current approaches attempting to automate the discovery and matchmaking process apply semantics to the service descriptions. These semantics are interpretable by the service (software) agents and should include WSDL-based functional parameters such as the Web services input-outputs (Martin et al., 2004a)(Akkiraju et al., 2005), and non-functional parameters such as domain-specific constraints and user preferences (Aggarwal, Verma, Miller, & Milnor, 2004).

The accuracy of automatic matchmaking of web services can be further improved by taking into account the adequacy of past matchmaking experiences for the requested task, which gives us valuable information about the services behaviour that is difficult to presume prior to service execution. Hence, there is a need for a methodology that uses domain-specific knowledge representation of the required task to capture the Web services execution experiences and utilise them in the matchmaking process. Case Based Reasoning (CBR) provides such methodology as its fundamental premise is that experience formed in solving a problem situation can be applied for other similar problem situation.

The article begins with describing the motivation behind the work. In the following section we review theory of Case Based Reasoning and describe how it can be utilised for modelling Web services matchmaking. Next we discuss the design of our matchmaking algorithm, its implementation highlights, and analyze preliminary results. Finally we investigate how case adaptation can further extend our matchmaking algorithm to cater for service composition and review related work.

MOTIVATION

The most practically deployed Web services composition techniques use the theory of business workflow-management as composition process model to achieve formalization for control and data flow. Mainly based on the Business Process Execution Language (BPEL) standard (Andrews et al., 2003), these techniques also have practical capabilities that fulfil the needs of the business environment, such as fault handling and state management. However, the main shortcoming of these techniques is the static selection and composition approach, where the service selection and flow management are done a priori and manually.

A popular research direction attempts to improve BPEL composition by introducing semantics to workflow-based composition (Osman, Thakker, & Al-Dabass, 2005). However, these approaches also match the static behaviour of Web services in terms of whether the service has similar description for functional and non-functional parameters. While for the candidate Web services it is highly likely that these parameters are semantically similar, it is the execution values for such functional and non-functional parameters that provide valuable guidance for the decision-making process regarding the service adequacy for the task. This is because service behaviour is difficult to presume prior to service execution and can only be formed based on the experience with the service execution.

Hence, the problem requires a methodology, which has the domain-specific knowledge representation system for capturing the Web services execution experiences and reason based on those experiences. We adopted CBR (Case Based Reasoning) as the engine for our Web services discovery mechanism because CBR's fundamental premise that situations recur with regularity (Aamodt & Plaza, 1994), i.e. experience involved in solving a problem situation can be applied or can be used as guide to solve other contextually similar problem situations. Reasoner based on CBR hence matches the previous experiences to inspire a solution for new problems.

OVERVIEW OF CASE BASED REASONING

The Case-Based Reasoning technology was developed in 1977 based on the research effort of Schank and Abelson. They proposed that our general knowledge about situations is recorded in the brain as scripts that allow us to set up expectations and perform inferences (Watson, 1997) The processes involved in CBR can be represented by a schematic cycle comprising four phases (Aamodt et al., 1994):

- RETRIEVE the most similar case(s);
- REUSE the case(s) to attempt to solve the problem;
- REVISE the proposed solution if necessary, and
- RETAIN the new solution as a part of a new case.

There are 4 main stages in CBR reasoning:

Case Representation

A case is a contextualised piece of knowledge representing an experience (Aamodt et al., 1994). It contains the problem, a description of the state of the world when the case occurred, and the solution to this problem. The solution contains elements that address the problem and inform about the relevance of the solution. When a reasoner is created, the elements of the case are defined according to the context. For example, the city of departure or the number of passengers could be some elements to represent a travel experience as a case. Case vocabularies are thus developed for each reasoner, to define what knowledge needs to be captured.

Case Storage and Indexing

Cases are then stored in a case library or case base. It is an important aspect for the designing of CBR systems because it reflects the conceptual view of what is represented in the case. The structure of the library should permit efficient search by the reasoner. This search can be facilitated by the use of indexing. Indices are therefore assigned to cases in order to express information about the case content.

Case Retrieval

Whenever a new problem needs to be solved, the case library is searched for the cases which can be a potential solution. The first phase of this search is case retrieval, which aims to find cases that are contextually similar to the new problem. The retrieval is done according to the index of the cases.

Matchmaking

Matchmaking performs the comparison between retrieved cases and the request to verify if a past solution can be reapplied. There are several available methods for matchmaking in CBR literature. The Nearest-Neighbour Matching and Ranking is an interesting one because it involves the assessment of similarity between stored cases and the input (request) case. It assigns importance ranking to properties of cases and then computes the degree of matching by comparing the cases for these properties (Kolodner & Simpson, 1989). The matchmaking process is thus performed on each retrieved case, and the most similar case to the input case is assigned the highest ranking. If the system finds a matching case, it is possible to reuse the solution suggested by the retrieved case for the new problem.

In our CBR matchmaking approach, Web services execution experiences are modelled as cases. The cases are the functional and non-functional domain specific Web services properties described using semantics. In this modelling, the case library will be the storage place for such execution experiences and is identical to Web service registry in that it stores Web services references, but unlike registries case libraries also describe execution behaviour.

Case retrieval is similar to Web services discovery problem in that both mechanisms seek to find potential Web services for the current problem. Case matchmaking is similar to Web services matchmaking as both attempts to select acceptable Web services, from the retrieved Web services during the case retrieval or Web service discovery phase respectively.

The apparent compatibility confirms our thesis that the CBR methodology is well suited to build automatic Web service composition frameworks

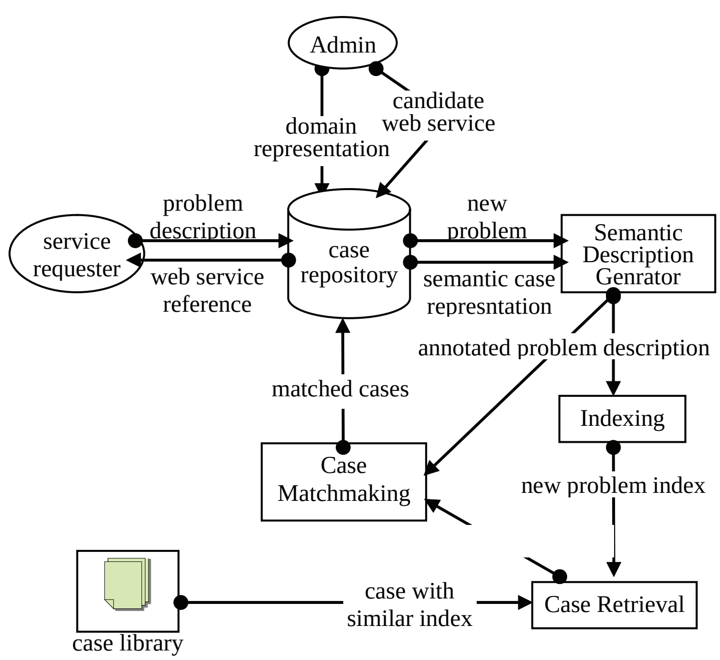
MATCHMAKING WEB SERVICES USING CASE BASED REASONING

The Framework Architecture

In our Semantic CBR matchmaking, there are two main roles: case administrator who is responsible for case library maintenance by entering or deleting cases from the library and case requestor who searches the case library to find solution for the problem. Figure 1 illustrates a schematic diagram for our framework.

The dynamics of the framework operation is as follows:

Figure 1. Architecture of the CBR matchmaking framework



- Initially, the administrator populates the repository with semantic case representation formats for specific application domain. This representation is used to semantically annotate both the user requests for suitable services and the execution experiences of Web services for the specific domain.
- The user inputs the service requirements and as a result receives Web service references via the framework interface. The same interface is used by service provides or the system administrator to subscribe a Web service as a candidate for available services for the specific domain.
- The case representation repository retrieves the appropriate semantic case representation format for the requested service and forwards it together with the problem description to the Semantic Description generator module, which semantically annotates the new problem according to the representation format.
- The annotated problem is then passed to the indexing module, which computes a suitable index for the new problem based on the domain feature and/or the functional parameters of the requested service. The index is passed for case retrieval.
- The case retrieval module queries the case library for cases with similar indexes. Output at this stage will be the cases that have similar index to the current problem, which will be candidates for matchmaking.
- The case matchmaking module takes the retrieved cases and the annotation of problem description from the semantic description generator module, runs them through a matchmaking algorithm and forwards the closest match Web service to the requester.

Although the chosen case study for this work is from the travel domain, the modular, ontology-driven design of framework makes it application-independent and allows for its seamless reuse for other applications domain. In order to enable matchmaking for the financial markets domain for instance, it would suffice to enter a new case representation format into the repository, keeping the rest of the reasoning logic intact.

Ontology Support for Case Representation and Storage

The most common use of ontologies is the reconciliation between syntactically different terms that are semantically equivalent. Applied to CBR case descriptions for Web services, ontologies can be used to provide a generic, reasoner-independent description of their functional and non-functional parameters. Moreover, ontologies can also be used to further index and structure cases with key domain features that increase the efficiency of the matchmaking process. For instance, we can add a feature to the travel domain ontology to indicate whether a trip is domestic or international. Web services QoS parameters are also indexed using ontologies to further improve the accuracy of case matchmaking.

In our framework, ontologies are also used to describe the rules of the CBR reasoning engine, which not only streamlines the intercommunication between the Web service, user request, and the case library, but promotes exploring the collaboration at the reasoning level between different composition frameworks.

Case Vocabulary

In CBR theory, the first step is to define all the elements contained in a case and the associated vocabulary that represents the knowledge associated with the context of a specific domain (our case study is the travel domain).

This vocabulary includes functional and non-functional parameters:

- Functional parameters are the service input (e.g. the travel details) and the service output or results (e.g. the travel itinerary). Input corresponds to the request of the user (e.g. date or city of departure) whereas output corresponds to the response given to the user (e.g. price, flight number).
- Non-functional parameters are constraints imposed by the user (e.g. exclusion of particular travel medium) or preferences over certain specific parameters (e.g. Price range, Quality of Service expected). In addition, execution experiences stored in the case library should also include the solution (i.e. Web services effectively used) and a notion to specify if the solution is acceptable for the end-user. Features that characterise the domain are extremely useful for top-level indexing and can also be included as non-functional parameters.

Case Representation Using Frame Structures

After deciding on the knowledge and corresponding vocabulary to be represented as a case, we need to decide how this knowledge can be represented.

In our approach, we adopt frame structures (Kolodner et al., 1989) the case representation. In frame structures, frame is the highest representation element consisting of slots and fillers. Slots have dimensions that represent lower level elements of the frame, while fillers are the value range the slot dimensions can draw from. In our implementation, slot dimensions represent case vocabulary in modular fashion while fillers describe the possible value ranges for the slot dimensions.

The frame representations are highly structured and modular which allows handling the complexity involved in representation. Moreover, frame structure has a natural mapping to the semantic OWL description language as the semantic net representations largely borrowed from the frame structures(Elaine & Kevin, 1992), which makes natural transition to the Semantic Web descriptions possible. Table 1 shows such a frame structure for our travel domain case vocabulary.

The slot Travel Request corresponds to the Input, i.e. all the travel details for a travel agent. The Travel Response slot corresponds to the Output, i.e. the answer given to the user at the end of the process. The elements of the answer are the price and the corresponding currency, the access point to the WSDL file of the corresponding Web Services and the Services Used (companies involved in the trip, e.g. an airline and a hotel).

Semantic Encoding of the Frame Structure

In the developed framework, we map the frame structures to ontologies. We derive rules for such mapping as described in Figure 2.

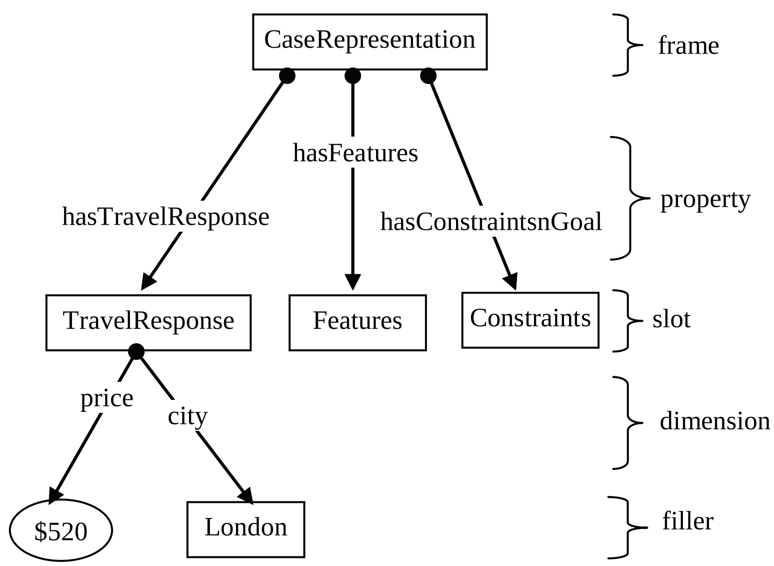
According to this mapping, frame and slot are represented as classes. The relationship between frame and slot is expressed in terms of properties of a frame, i.e. the range for these properties are the slot classes. Dimensions are the properties of the slots. Possible range for these properties is the values the respective filler can derive from.

We use Web Ontology Language (OWL), a Semantic Web standard for constructing these ontologies. OWL is the most expressive Semantic Web knowledge representation so far. The layered approach adopted by semantic web, allows reasoning and inference based on ontologies, which is the most powerful and ubiquitous feature of Semantic Web. After applying the mapping, the ontology for the travel domain case representation is created, where for instance the CaseRepresentation class has: hasTravelResponse, hasConstraintsOnGoal, and hasFeature object properties. Range for these properties are TravelResponse, Constraints, and Feature classes respectively.

Table 1. Representation of a case

Slot	Dimension	Filler
Travel Request	Name of Traveler	Any text
	Date of Arrival	Any valid date
	City of Departure	Any valid city
Travel Response	Solution	Service WSDL file
	Price Range	Any positive Double
	Currency	Any valid currency
Constraints	On Domain	Any Valid Travel Domain
	On Price range	Any positive Double
	On QoS parameter	Any possible QoS parameter(s)
Features	Travel Regions	Domestic/International

Figure 2. Mapping between frame structure and semantic case representation



globalization of semantic descriptions, we used external ontologies where appropriate. For instance, the property cityOfArrival is an object property referring to a publically available ontology(PORTAL, 2003) , where other useful information about the specific city can be found such as country, the number of inhabitants, etc.

An example of a semantically-encoded travel request is illustrated in Table 2. "Find a Trip for a single person, Mr Lee; Mr Lee wants to travel from Boston to New York, with a maximum price range in total of $220, He does not want to travel by road. The dates of Travel will be 27-02-2005 for departure and 01-03-2005 for return. He prefers to pay in USD and requires a fast response (approximately in 1.5 seconds)".

Case Storage

All the Web service execution experiences, i.e. solutions deemed valid for a particular request, are stored in the Case Library to be reused by

In order to exercise the noble objective of globalization of semantic descriptions, we used external ontologies where appropriate. For instance, the property cityOfArrival is an object property referring to a publically available ontology(PORTAL, 2003) , where other useful information about the specific city can be found such as country, the number of inhabitants, etc.

An example of a semantically-encoded travel request is illustrated in Table 2. “Find a Trip for a single person, Mr Lee; Mr Lee wants to travel from Boston to New York, with a maximum price range in total of $220, He does not want to travel by road. The dates of Travel will be 27-02-2005 for departure and 01-03-2005 for return. He prefers to pay in USD and requires a fast response (approximately in 1.5 seconds)”.

Case Storage

All the Web service execution experiences, i.e. solutions deemed valid for a particular request, are stored in the Case Library to be reused by the reasoner. The Case Library itself is also an ontology. It contains some instances of the class CaseRepresentation (e.g. a travel experience or a travel case).

DEVELOPMENT OF THE CBR MATCHMAKING FRAMEWORK

Case Indexing and Retrieval

To facilitate the search procedure, cases are indexed based on vocabularies. In our framework, we use “partitioning the case library” method, which is a variation of “flat memory indexing” technique (Kolodner et al., 1989). In this indexing method, case library is partitioned based on certain vocabularies and the new problem is recognized based on the identical vocabularies to decide which partition the problem falls into.

In our architecture, cases are stored based on vocabulary element Features as presented in Table 1, which corresponds to hasFeatures property from the CaseRepresentation ontology class. For our travel agent case study, the possible values for this property are either Domestic or International (predefined instances from the TravelRegion class), hence indexing will partition case library into two parts. In more complex examples more than one vocabulary term or a combination of terms can be used for more sophisticated indexing. As in relational databases selection, the efficiency of the retrieval process largely depends on the precision of the indexing.

Table 2. Example of a case

Name of passengers	Lee	<TravelRequest:namePassengers>Lee
City of Arrival	Boston	<TravelRequest:cityArrival rdf:resource= “http://localhost/uk/2005/City.owl#Boston”/>
Date of Arrival	1-3-05	<TravelRequest:dateArrival>2005-03-01
Constraint on domain	Road	<Constraints:OnDomain rdf:resource= “http://localhost/ntu/TravelDomain.owl#Airline”/>
Constraint on price	220	<Constraints:OnPrice>220
Constraint on currency	USD	<Constraints:OnCurrency rdf:resource= “http://ecs.soton.ac.uk/currency.daml#USD”/>
Constraint on QoS	1.5 s	<QoS:ExecutionDuration>1.5

Whenever a new Web service needs to be fetched, the problem description involving the functional parameters and non-functional parameters are encoded using the case representation frame structure, i.e. as an instance of CaseRepresentation ontology as illustrated in Table 2.

Matchmaking and Ranking

Case retrieval fetches Web services that are a potential solution to the problem. The matchmaking process narrows down the retrieved cases to present acceptable solution(s). From the available methods for matchmaking in CBR literature, we choose Nearest-Neighbour Matching and Ranking using numeric evaluation function (Remind, 1992) method for our framework. The method operates as follows:

- Compare the similarity for each property, between the new problem and the cases retrieved. The method used for comparison depends on the type of the property.
- Quantify the weight of the similarity. A ranking is assigned to each property in accordance with its importance as exemplified in Table 3.

For each case retrieved, the similarity degree is computed and the case with the highest score corresponds to the best-match. Similarity takes values between 0 and 1, which is attributed to each property for each retrieved case. Our similarity comparison method depends on the type of the dimension: data or object.

Data Property Comparison

To compare data type properties, like the price range or the value of QoS (e.g. execution time), we use the qualitative regions based measurement method (Kolodner et al., 1989). The closer the value in a retrieved case is to the value in the request, the higher the similarity coefficient is.

For each data type property, this formula used is: $|Vr - Vc| \leq X.[Vr|$, where V is the value of the property in the request r or in the retrieved case c and X the factor of tolerance. Thus, a factor of tolerance of 0.9 means the value of the retrieved case should be in ±10% region in relation to the value of the request. The optimum tolerance value is determined by the administrator and can be calculated heuristically.

Object Property Comparison

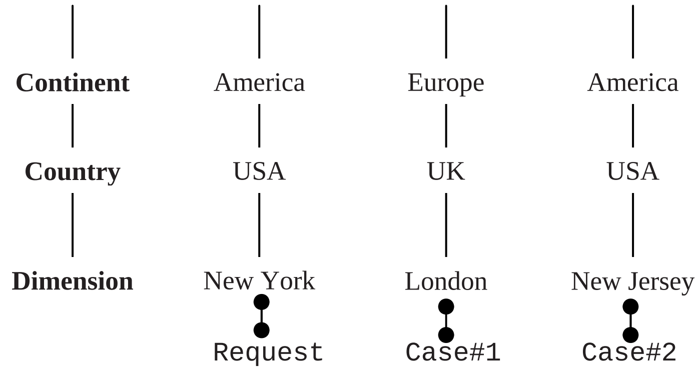
For the dimensions annotated as object properties, the possible filler values will be an instance of slot class. Hence, for semantically matching object property value of the new problem and the retrieved cases, the algorithm compares the instances. If the instances match, then the degree of match is 1. Otherwise, the algorithm traverses back to the super (upper) class that the instance is derived from and the comparison is performed at that level.

The comparison is similar to traversing a tree structure (Zhang, Arpinar, & Aleman-Meza, 2003), where the tree represents the class hierarchy for the ontology element. The procedure of traversing back to the upper class and matching

Table 3. Quantifying the travel domain case dimensions

Slot	Dimension	Importance (0-1)
Travel Request	City Departure	1.0
	City Arrival	1.0
Constraints on Goal	On Instance	0.2
	On Domain	0.8

Figure 3. Semantically matching object properties (dimensions)



instances is repeated until there are no super classes in the class hierarchy, i.e. the top node for the tree is reached, giving degree of match equal to 0. The degree of match (DoM) degree is calculated according to the following equation:

$$DoM = \frac{MN}{GN} \tag{1}$$

Where the MN is Total number of matching nodes in the selected traversal path, and GN Total number of nodes in the selected traversal path

For example, for the request in Figure 3, case#1 will return a degree of match of 0 because no matches are found while traversing the ontology tree until the leaf node is reached. However, for case#2, the degree of match will be 2/3=0.67 as the instances (New Jersey, New York) does not match but the instances of the Country super class match.

It is worth to note that Constraints on object properties are handled by omitting that path in the case ontology tree that renders the constraint invalid. For example, if the passenger is reluctant to travel by air, then the Brit Air, Flight path will not be traversed.

Computing The Overall Similarity Value

Overall similarity is evaluated by computing the aggregate degree of match (ADoM) (Remind, 1992) for each retrieved case according to the following equation:

$$ADoM = \frac{\sum_{i=1}^{n} W_i \times sim(f_i^N, f_i^R)}{\sum_{i=1}^{n} W_i} \tag{2}$$

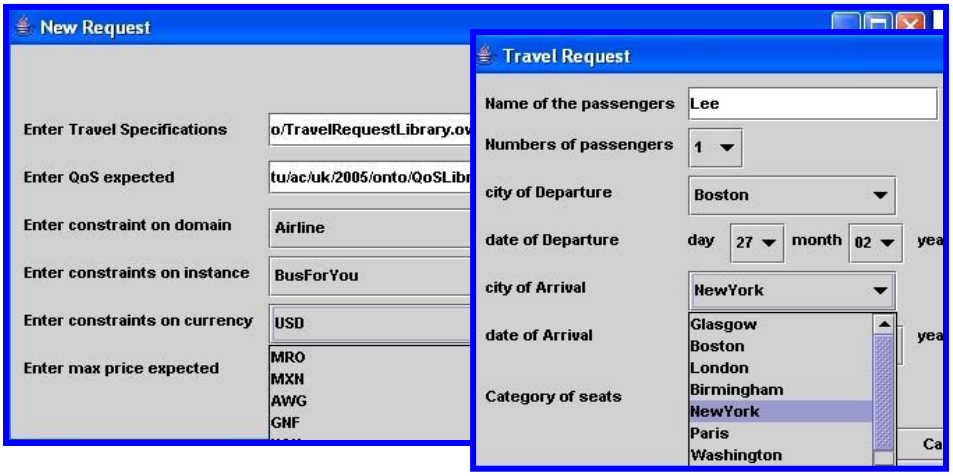
Where, n is the number of ranked dimensions, Wi is the importance of dimension i, sim is the similarity function for primitives, and fiN and fiR are the values for feature fi in the new problem and the retrieved case respectively.

The evaluation function sums the degree of match for all the dimensions computed in the previous step, and takes aggregate of this sum by considering the importance of dimensions.

IMPLEMENTATION HIGHLIGHTS

The implementation of our framework uses semantics extensively to implement both the utility ontologies describing the components of the Case-Based Reasoner (Case representation), and the domain ontologies that describe the profile of the Web services in the Case library with a semantic representation (Case Storage).

Figure 4. Admin and user interface



OWL was our ontology language of choice. We used Pellet (Parsia & Sirin, 2004) - a Java based OWL reasoner, as our ontology engine in favour of the more popular Jena (HP-JENA, 2003), because it supports user-defined simple types. Pellet was used to load and verify (type and cardinality) ontology class instances of user requests and candidate cases.

Figure 4 illustrates a snapshot of the GUI developed for the matchmaking framework. The interface allows different options to two kinds of users: The case administrator, who is responsible for maintaining the case library, and a standard client, who wants to retrieve Web services for a trip. The case administrator has admin privileges to perform case maintenance activities like case seeding, modifying the ranking system or deleting old cases. The client can also setup a ranking system, which will be applicable for a particular session.

While seeding the case library with a new case or making a new trip request, the interface assists the client in creating the required ontology instances. The value entered for a particular property is validated in relation to the range and cardinality drawn from the ontologies.

The solutions (cases) resulting from the matchmaking process are presented to the client are stored into the case library.

PRELIMINARY RESULTS

At this initial stage of development, the focus of our experiments was to validate the logic of our matchmaking framework, rather than testing a fully working prototype. Hence, we tested our framework with simple in-house developed Web services and compatible wrappers for external publicly available services.

In order to consolidate the test process, we applied different rankings against each test case and associated them with a specific profile. The profile represents a group of users that have similar requirements for the travel request. For instance, the Business profile stands for corporate users, who have to travel frequently; therefore a high standard of comfort is a significant element of choice. These users also need reliability of services. Price is not very important because firms very often have contracts with travel companies. On the other hand, for regular users, represented by the Personal profile, cost is of paramount importance.

The three other types of users are mainly based on specific comparison properties: Economic retrieves cases which price never exceeds a user-defined maximum amount; Travel Medium is specific for constraints on travel domain as well as instances; and Enterprise is useful for companies which are interested in using reliable services. The latter can be important if contracts between the company and different Web services exist so that they can restrict other services.

The rankings are currently administered centrally, but in the future we would like to give the users the opportunity to tweak some of them using a user-friendly interface. Table 4 shows the ranking of our profile system. Example of constraint on Domain is reluctance to travel using a certain transport and constraint on instance the exclusion of certain airline from the search. Quality of service is represented as a single parameter, but in this experiment it is expressed as the availability and response time of the service.

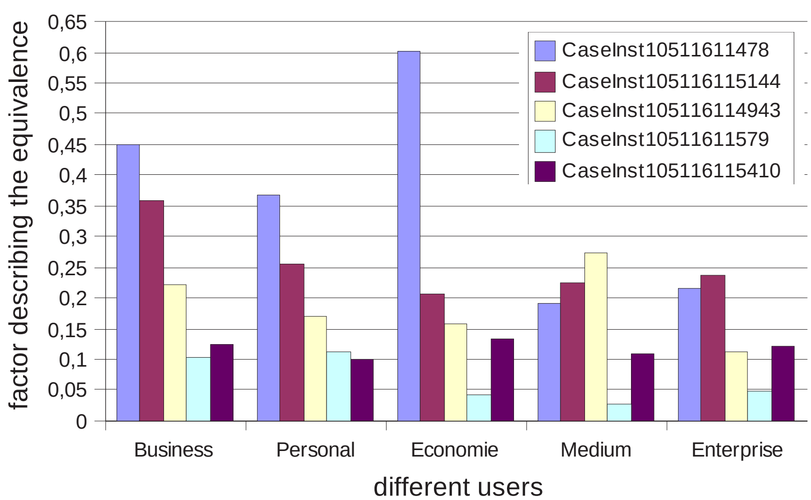
Figure 5 shows the matchmaking degree for different cases using the criteria above. Some cases (Web service execution experiences) present satisfactory results to all users (CaseInst10511611478). Another interesting highlight is that the chosen ranking systems provide different results only if the coefficients are significantly different. This is probably due to the fact that that our case library is not richly populated at the moment.

The average execution time of our matchmaking program at the time of the experiment was approximately 40 seconds, relatively slow considering we only have 30 cases stored in the library. Using semantics has the disadvantage of being more time-consuming than scanning databases. We identified the use of imported ontologies as the

Table 4. User profiles

	Property				
Profile	*Category*	*Constraint on Domain*	*Constraint. on Instance*	*Price*	*Quality of Service*
Business	0.6	0.6	0.4	0.1	0.5
Personal	0.2	0.4	0.7	0.5	0.2
Economic	0.2	0.4	0.2	1	0.1
Travel Medium	0.2	1	0.8	0.3	0.2
Enterprise	0.5	0.3	0.1	0.2	1

Figure 5. User profiles



main performance leak for our program. We plan to develop an off-line caching system to enable us to access the public ontologies locally.

EXTENDING THE MATCHMAKING FRAMEWORK TO WEB SERVICES COMPOSITION

The current framework addresses the problem of automatic Web services discovery and matchmaking by annotating Web services execution experiences and storing them into case base (Osman et al., 2006). The search considers domain-specific criteria and user preferences to find Web services execution experience that solved a similar problem in the past. However the framework assumes that the case library contains suitable cases for every possible problem. This assumption is not always satisfied considering the vast number of problems and problem parameters. Moreover, the framework also needs to deal with situations where the aggregate degree of match (ADoM) is below the *domain-specific* expected degree of match set by the domain administrator or to deal with negative user feedback, where the matched services are not acceptable to the user.

Work under progress involves exploring case adaptation, which is termed as the REVISE phase (Figure 6 - The REVISE phase in CBR) in CBR theory. Adaptation is applicable when the available cases cannot fulfil the problem requirements, so matchmaking is attempted by adapting available cases. Adaptation looks for prominent differences between the retrieved case and the current case and then applies formulae or rules that take those differences into account when suggesting a solution (Watson & Marir, 1994).

Applied to the current framework, when the existing web services experiences in their original form are not sufficient to satisfy current request, the framework should look for relaxing the case restrictions under which a solution is acceptable. If the latter fails, the framework should attempt to merge potential cases to suggest a *composite* solution.

Case adaptation can be defined by the following formula (Maher & Garza, 1997)

$$C' = \alpha(C) \tag{3}$$

Where, C' = new case, C = old case(s) and α indicates adaptation operator.

The adaptation operator indicates the process of identifying and substituting or transforming an existing solution to fit new situations and is used in *knowledge-based substitution* adaptation.

Knowledge Based Substitutions

In CBR matchmaking process previous cases cannot be always reused without making some changes. Reasoning about these changes requires general and domain specific knowledge to mould case adaptation. For example, if an existing case-solution is applicable for the current travelling problem with the exception of the travel medium

Figure 6. The REVISE phase in CBR

REVISE

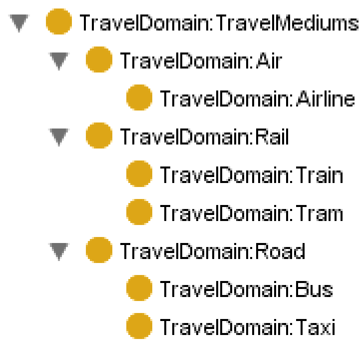
IF ADoM < $\{C_{results}, ADoM_{expected}\}$

OR

IF$\{C_{results}\} \neq U_{expected}$

$U_{expected} \rightarrow$ User expected results
$C_{results} \rightarrow$ proposed case solutions

Figure 7. Local search



– Bus (a road based travel medium which is constrained in the current problem), then the reasoner should use local search on knowledge structure (see the ontology in Figure 7) and conclude the Taxi domain as a possible substitution. Similarly, interpolating parameter of old solution to adapt for the new problem can solve the problem of parameter mismatch.

Another substitution method can use the semantic "sibling" rule for equivalent classes to enables them to replace each other in order to present an appropriate solution.

Under this circumstance, the Equation 3 can be reformulated as:

$$C' = \alpha(C, K) \qquad (4)$$

Where K indicates the influence of general or domain specific knowledge.

The role of knowledge in repairing the existing cases can be described as follows:

- Relaxing the service descriptions (functional parameters) to find a sufficiently similar description (applied at the description level).
- Relaxing the execution values of candidate cases (their non-functional parameters) in an attempt to adapt the solution.

The criteria for applying knowledge based substitution are:

- In a situation, where the reasoner cannot find an exact or reasonable match or where exact match is not possible or not desired or not required.
- To make absolutely sure that only possible solution is transformation, which is an expensive operation involving an AI planner, which is a resource expensive exercise.

Hence, if for the current problem C, the available cases in the case library are:

$$C_{libray} = \{\ C_1\ (S_1{+}S_2,\ F_1),\ C_2\ (S_2{+}S_3,\ F_2),\ C_3\ (S_1,F_3),\ C_4\ (S_2,F_4)\ ,\ C_5\ (S_1{+}S_2,\ F_5)\ \ \}$$

Where, C (S, F) indicates cases with Web services as solution S applied under circumstances defined by F The circumstances can be characterized by service description, problem description, constraints and preferences applied while solving the problem.

Then the solution for new problem C, C_6 (S_1 + S_2, F_6) must be reached by exploring the matching cases C_1 and C_5 first, before transforming C_3 and C_4 to find a solution from a scratch.

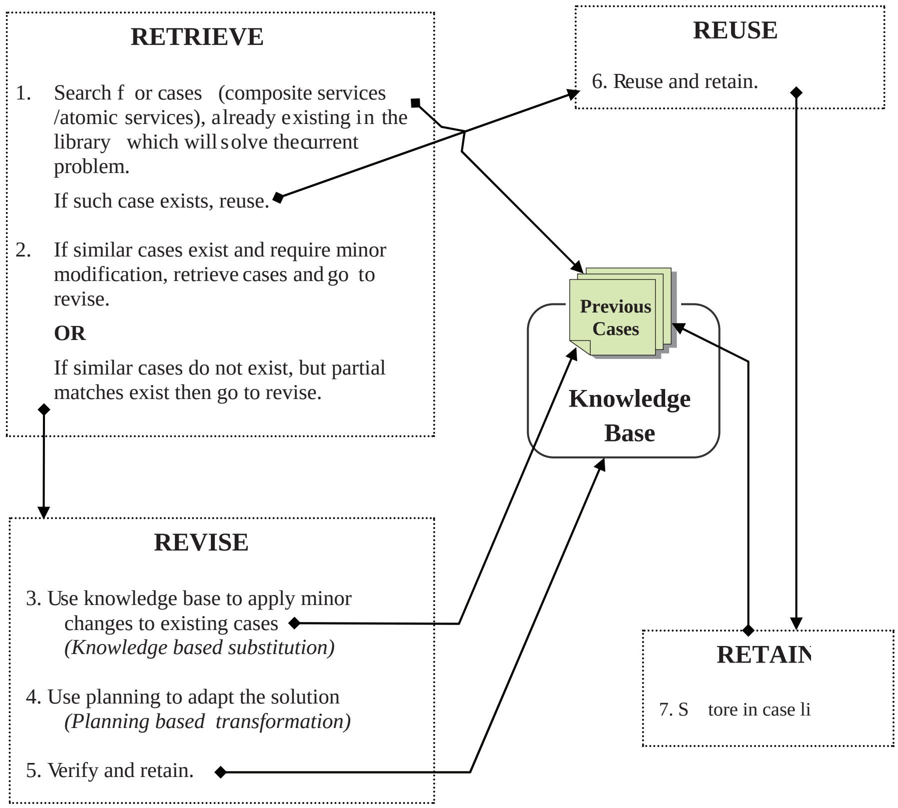
Planning Based Transformations

The planning based transformations can be applicable when the available solutions can not fulfil the problem requirements with normal matchmaking and discovery mechanism or by applying minor modifications using substitution based transformation. Under these circumstances, the equation 3 can be reformulated as:

$$C' = \alpha(C,\ \rho) \qquad (5)$$

ρ indicates the application of planner for transformation, where classical planner handles the task of coming up with a sequence of actions that will achieve a goal (Russell & Norvig, 2003). The planning phase is a resource-intensive and computational expensive (Selman, 2000; Long &

Figure 8. CBR methodology for Web services composition (modified from [(Aamodt et al., 1994)])



Fox, 2002) yet inevitable option, hence the two previous phases should narrow down the number of possible services planner can successfully use to generate a composed service.

Figure 8 on the shows the holistic CBR methodology to achieve Web services composition using the REVISE cycle.

RELATED WORK

Semantic descriptions are increasingly being used for exploring the automation features related to Web services discovery, matchmaking and composition. In (Zhang et al., 2003) such semantic-based approach is described. They use ontology to describe Web services templates and select Web services for composition by comparing the Web service output parameters with the input parameters of other available Web services. A constraint driven composition framework in (Aggarwal et al., 2004) also uses functional and data semantics with QoS specifications for selecting Web services. DARPA's OWL-S (Ontology Web Language for Web services) is the leading semantic composition research effort. OWL-S (Martin et al., 2004b; Ankolekar et al., 2001) ontologies provide a mechanism to describe the Web services functionality in machine-understandable form, making it possible to discover, and integrate Web services automatically. An OWL-based dynamic composition approach is described in (Sirin, Hendler, & Parisa, 2003), where semantic description of the services are used to find matching services to the user requirements at each step of composition, and the generated composition is then directly executable through the grounding of the services. Other Approaches use Artificial

Intelligence planning techniques to build a task list to achieve composition objectives: selection of services and flow management for performing composition of services to match user preferences. (Mcllraith & Son, 2002) uses Golog – AI planning Reasoner for automatic composition, while in a similar spirit some other approaches (Wu, Parisa, Hendler, Nau, & Sirin, 2006; Nau, Cao, & Lotem, 1999) have used the paradigm of Hierarchical Task Network (HTN) planning to perform automated Web service composition. These approaches use semantics for automatic Web services discovery, but they overlook the Web service execution behaviour in the decision-making process.

Use of CBR, Semantic Web and Web services are common technologies in our effort and the efforts in (Nern et al., 2006) with different objectives, their's being to consume these technologies to assist the procedure of Semantic Web services creation using Case-Based reasoning approach, while our main concern is services composition.. Their INFRAWEBS project has Semantic Web Unit (SWU) – a collaboration platform and interoperable middleware for ontology-based handling and maintaining of Semantic Web services. The framework provides knowledge about a specific domain and relies on ontologies to structure and exchange this knowledge to Semantic Web services development process.

There is also a number of existing approaches which applies CBR for workflow modelling. (Madhusudan, Zhao, & Marshall, 2004) proposes an approach to support workflow modelling and design by adapting workflow cases from a repository of process models where workflow schemas are represented as cases and are stored in case repositories. The cases are retrieved for a problem which requires similar business process to solve the problem. The description and implementation language of framework is based on XML and main focus is on assisting workflow designer in creating business process flows. In similar line, (Cardoso & Sheth, 2005) represents adaptive workflow management system based on CBR and targets highly adaptive systems that can react themselves to different business and organization settings. The adaptation is achieved through the CBR based exception handling, where the CBR system is used to derive an acceptable exception handler. The system has the ability to adapt itself over time, based on knowledge acquired about past execution experiences that will help solve new problems. Our approach concentrates on Web services as a unit of computation to take advantage of highly accessible and loosely coupled nature of Web services technologies. We focus on utilising service execution experiences to best serve user requirements and encode the framework with semantics.

Experience based learning using CBR is a relatively old branch of Artificial Intelligence and Cognitive Science and is being used (Hammond, 1986; Ashley & Rissland, 1988) as an alternative to rule-based expert system for the problem domains, which have knowledge captured in terms of experiences rather than rules. However, Case based reasoning for Web services was initially documented in (Limthanmaphon & Zhang, 2003), where the developed framework uses CBR for Web services composition. In their approach, the algorithm for Web services discovery and matchmaking is keyword based and has no notion for semantics. This affects the automation aspects for Web services search and later for composition. Similar approach described in (Diaz, Salgado, Moreno, & Ortiz, 2006) proposes an extension of UDDI model for web services discovery using category-exemplar type of CBR, where web services are categorized in domains and stored as exemplar (Porter & Bareiss, 1986) of particular domain. Their implementation of CBR reasoner facilitates UDDI registry by indexing the cases based on the functional characteristics of Web services. However, the approach does not take into consideration the importance of non-functional parameters in service selection and the use of semantics at CBR level is peripheral as they primarily use the UDDI based component for service discovery. UDDI is text-based leaving little scope for automation. Our framework consumes semantics extensively and achieves the automation required for Web service discovery and matchmaking. Use of ontologies also makes our framework extensible and reusable.

CONCLUSION

Semantic description of Web service profile paves the way for automating the discovery and matchmaking of services since it allows intelligent agents to reason about the service parameters and capabilities. However, the accuracy of such automatic search mechanism largely relies on how soundly formal methods working on such semantic descriptions consume them.

In this article, we argued for the importance of considering the execution values for semantically described functional and non-functional Web services parameters in decision making regarding Web service adequacy for the task. This is because the service behaviour is impossible to presume prior to execution and can only be generalized if such execution values are stored and reasoned for deciding service capability. AI planning and Intelligent Agent based reasoning methods offer rule-based reasoning methodology rather than experience-based. Hence, we used Case Based Reasoning method that allows capturing experiences and reasoning based on them.

We implemented a Semantic Case based Reasoner, which captures Web service execution experiences as cases and uses these cases for finding a solution for new problems. The implemented framework extensively uses ontologies, as semantics are used both for describing the problem parameters and for implementing components of the CBR system: representation, indexing, storage, matching and retrieval. Our approach for modelling CBR as ontology-based reasoner achieves developer transparency and makes the framework extensible and reusable.

A problem that research in semantic-based matchmaking and composition has not addressed sufficiently is the interoperation between independently developed reasoning engines. Without this interoperation, the reasoning engines remain imprisoned within their own framework, which is a drawback, especially that most engines usually specialise in servicing a particular domain, hence interoperation can facilitate inter-domain orchestration. We believe that in this work we took a small step towards standardization at the reasoner level by describing the CBR reasoning model semantically

In this article we also presented the preliminary experimental results of our framework, which informally proved the correctness of our approach despite the relatively slow response time of the matchmaking process. The latter is primarily attributed to exporting external ontologies, which can be countered by utilising off-line caching of public ontologies. The experimental results also demonstrated the advantages of classifying user groups into profiles that have standard set of constraint rankings.

The final contribution of the article was documenting our investigation into extending the discovery and matchmaking algorithm to cater for web services composition. We discuss how we envisage exploiting the REVISE stage of the CBR cycle, i.e. case adaptation, to facilitate service composition. The article advocates an exhaustive *knowledge-based substitution* approach to adapt the functional and non-functional attributes of the candidate case to the requested solution before suggesting more complex and computationally taxing AI-based planning-based transformations that integrate the service profile of a number of cases to deliver candidate solutions.

The next stage of this research will involve the formal validation and implementation of our adaptation-based composition model.

REFERENCES

Aamodt, A. & Plaza, E. (1994). Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Communications*, *7*, 39-59.

Aggarwal, R., Verma, K., Miller, J., & Milnor, W. (2004). Constraint Driven Web Service Composition in METEOR-S. In *2004 IEEE International Conference on Services Computing (SCC 2004)* (pp. 23-30).

Akkiraju, R., Farrell, J., Miller, J., Nagarajan, M., Schmidt, M.-T., Sheth, A. et al. (2005). Web Service Semantics - WSDL-S. In *A joint UGA-IBM Technical Note.*

Andrews, T., Curbera, F., Dholakia, H., Goland, Y., Klein, J., Leymann, F. et al. (2003). Business Process Execution Language for Web Services version 1.0. http://www-128.ibm.com/developerworks/library/specification/ws-bpel/ [On-line].

Ankolekar, A., Burstein, M., Lassila, O., Martin, D., Mcllraith, S., Narayanan, S. et al. (2001). DAML-S: Semantic Markup for Web Services. In *International Semantic Web Working Symposium (SWWS)* (pp. 411-430).

Ashley, K. D. & Rissland, E. L. (1988). A case-based approach to modelling legal expertise. *IEEE Expert: Intelligent Systems and Their Applications, 3,* 70-77.

Cardoso, J. & Sheth, A. (2005). Adaptation and Workflow Management Systems. In *International Conference WWW/Internet 2005* (pp. 356-364). Lisbon, Portugal.

Diaz, O. G. F., Salgado, R. S., Moreno, I. S., & Ortiz, G. R. (2006). Searching and Selecting Web Services Using Case Based Reasoning. In *Computational Science and Its Applications (ICCSA 2006)* (pp. 50-57).

Elaine, R. & Kevin, K. (1992). *Artificial Intelligence*. McGraw-Hill.

Hammond, K. J. (1986). Learning to anticipate and avoid planning problems through the explanation of failures. In *AAAI-86 conference* (pp. 556-560). Cambridge, MA: AAAI press/MIT press.

HP-JENA (2003). Jena- A semantic Web Framework for Java. http://jena.sourceforge.net/ [On-line]. Available: http://jena.sourceforge.net/

Kolodner, J. & Simpson, R. (1989). The MEDIATOR: Analysis of an early case based problem-solver. *Cognitive Science, 13,* 507-549.

Limthanmaphon, B. & Zhang, Y. (2003). Web service composition with case-based reasoning. In *Fourteenth Australasian database conference on Database technologies* (pp. 201-208). Adelaide, Australia.

Long, D. & Fox, M. (2002). Progress in AI Planning Research and Applications. *Upgrade/ Novatica, 3,* 10-25.

Madhusudan, T., Zhao, L. J., & Marshall, B. (2004). A case-based reasoning framework for workflow model management. *Data & Knowledge Engineering archive, 50,* 87-115.

Maher, M. L. & Garza, A. G. d. S. (1997). Case-Based Reasoning in Design. *IEEE Expert: Intelligent Systems and Their Applications, 12,* 34-41.

Martin, D., PaolucciSycara, M., McIlraith, S., Burstein, M., McDermott, D., McGuinness, D. et al. (2004a). Bringing Semantics to Web Services: The OWL-S Approach. In *First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004)* San Diego, California, USA..

Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., Mcllraith, S. et al. (2004b). Semantic Markup for Web Services. http://www.w3.org/Submission/OWL-S/ [On-line].

McIlraith, S. & Son, T. C. (2002). Adapting Golog for Composition of Semantic Web Services. In *the Eighth International Conference on Knowledge Representation and Reasoning (KR2002)* (pp. 482-493).

Nau, D., Cao, Y., & Lotem, A. (1999). SHOP: Simple Hierarchical Ordered Planner. In *International Joint Conference on Artificial Intelligence (IJCAI-99)* (pp. 968-973).

Nern, J., Agre, G., Atanasova, T., Marinova, Z., Micsik, A., László, K. et al. (2006). INFRAWEBS Semantic Web Service Development on the Base of Knowledge Management Layer. *International Journal "Information Theories and Applications, 13,* 161-168.

Osman, T., Thakker, D., & Al-Dabass, D. (2005). Bridging the Gap between Workflow and Semantic-based Web services Composition. In *the 2005 IEEE/WIC/ACM International Joint Conference* (pp. 13-23). Compiègne, France.

Osman, T., Thakker, D., & Al-Dabass, D., Lazer, D., & Deleplanque, G. (2006). Semantic-Driven Matchmaking of Web services using Case-Based Reasoning. Accepted for the 2006 IEEE International Conference on Web Services (ICWS 2006), held on September 18-22, 2006, Chicago, USA.

Parsia, B. & Sirin, E. (2004). Pellet: An OWL DL Reasoner. In *the Third International Semantic Web Conference (ISWC2004)* Hiroshima, Japan.

PORTAL (2003). Portal Ontology. http://www.aktors.org/ontology/portal [On-line].

Porter, B. W. & Bareiss, R. E. (1986). PROTOS: An experiment in knowledge acquisition for heuristic classification tasks. In *First International Meeting on Advnces in Learning (IMAL)* (pp. 159-174). Les Arcs, France.

Remind (1992). ReMind Developer's Reference Manual. *Cognitive systems.*

Russell, S. & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. (2 ed.) Prentice Hall, Pearson Education, Inc, Upper Saddle River, New Jersey.

Selman, B. (2000). Compute-Intensive Methods for Artificial Intelligence. *Annals of Mathematics and Artificial Intelligence, 28,* 35-38.

Sirin, E., Hendler, J., & Parisa, B. (2003). Semi-automatic composition of web services using semantic descriptions. In *Web Services: Modeling, Architecture and Infrastructure workshop in ICEIS 2003* Angers, France.

Watson, I. (1997). *Applying Case-Based Reasoning: Techniques for Enterprise Systems.* (1st ed.) Morgan Kaufmann Publishers.

Watson, I. & Marir, F. (1994). Case-Based Reasoning: A Review. *The Knowledge Engineering Review, 9(4), 1994, 9,* 355-381.

Wu, D., Parisa, B., Hendler, J., Nau, D., & Sirin, E. (2006). HTN Planning for Web Service Composition Using SHOP2. *Journal of Web Semantics, 1,* 1-30.

Zhang, R., Arpinar, B. I., & Aleman-Meza, B. (2003). Automatic Composition of Semantic Web Services. In *International Conference on Web Services* (pp. 38-41). Las Vegas, Nevada, USA.

This work was previously published in International Journal of Intelligent Information Technologies, Vol. 5, Issue 1, edited by J.N.D. Gupta, S.K. Sharma, and J. Hsu, pp. 24-42, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 2.19
Rule Markup Languages and Semantic Web Rule Languages

Adrian Paschke
Freie Universität Berlin, Germany

Harold Boley
National Research Council, Canada

ABSTRACT

Rule markup languages will be the vehicle for using rules on the Web and in other distributed systems. They allow publishing, deploying, executing and communicating rules in a network. They may also play the role of a lingua franca for exchanging rules between different systems and tools. In a narrow sense, a rule markup language is a concrete (XML-based) rule syntax for the Web. In a broader sense, it should have an abstract syntax as a common basis for defining various concrete languages addressing different consumers. The main purposes of a rule markup language are to permit the publication, interchange and reuse of rules. This chapter introduces important requirements and design issues for general Web rule languages to fulfill these tasks. Characteristics of several important general standardization or standards-proposing efforts for (XML-based) rule markup languages including W3C RIF, RuleML, R2ML, SWRL as well as (human-readable) Semantic Web rule languages such as TRIPLE, N3, Jena, and Prova are discussed with respect to these identified issues.

DOI: 10.4018/978-1-60566-402-6.ch001

INTRODUCTION AND MOTIVATION

Web rule languages provide the required expressiveness enabling machine-interpretation, automated processing and translation into other such Web languages, some of which also being the execution syntaxes of rule engines. One of these languages may act as a “lingua franca” to interchange rules and integrate with other markup languages, in particular with Web languages based on XML and with Semantic Web languages (e.g. W3C's RDF Schema, OWL and its new OWL 2 version) for ontologies serialized in RDF/XML or directly in XML. Web rule languages may also be used for publication purposes on the Web and for the serialization of external data sources, e.g. of native online XML databases or RDF stores. Recently, there have been several efforts aiming at rule interchange and building a general, practical, and deployable rule markup standard for the (Semantic) Web. These encompass several important general standardization or standards-proposing efforts including RuleML (www.ruleml.org), SWRL (www.w3.org/Submission/SWRL/), SWSL (http://www.w3.org/Submission/SWSF-SWSL/), R2ML (oxygen.informatik.tu-cottbus.de/rewerse-i1/?q=R2ML), RIF (www.w3.org/2005/rules/), and others such as XCL (http://www.altheim.com/specs/xcl/1.0/), designed as a concrete (serialization) syntax for ISO's Common Logic (CL) standard.

In this chapter, a system of general requirements and design choices for Web rule languages will be introduced and instantiations discussed in the context of the current prominent general Rule Markup Languages and Semantic Web rule languages. This chapter is intended to be of help to a wide audience. In particular, it is targeted to rule practitioners who want to serialize the declarative rules of their applications in a general rule markup language, and publish and interchange them on the Web. Rule practitioners will find here a discussion of general design criteria with examples from the current rule markup languages. These examples, together with a discussion of advantages and drawbacks, will offer guidance to readers when declaratively representing their own rule-based applications in a Web rule language. The structure of the rest of this chapter is as follows: Section 2 introduces current rule markup languages and rule interchange formats as well as Semantic Web rule languages. Section 3 comprises the main part of this chapter, discussing important design issues and characteristics of the introduced rule languages. Section 4 presents future research issues in Web rule language design. Section 5 concludes this chapter with a summary.

WEB RULE LANGUAGES

Rule markup (serialization) languages have been developed for the Web-based interchange of, e.g., privacy policies, business rules, and - as focused here - Semantic Web rules. Rules are central to knowledge representation for the Semantic Web (Boley, 2007), hence are increasingly considered as being side by side with ontologies, e.g. in W3C's layered Semantic Web architecture (2007 version shown in Figure 1).

Rule interchange in an open format is important for all higher Semantic Web layers, including a Web of Trust and, generally, a Pragmatic Web (Paschke et al, 2007), and is crucial for applications in eBusiness, eGovernment, eHealth, etc. This section introduces major *rule markup languages* including RuleML, R2ML, and RIF, as well as human-readable *Semantic Web rule languages* such as TRIPLE and N3, and platform-specific rule engine languages such as Jena and Prova.

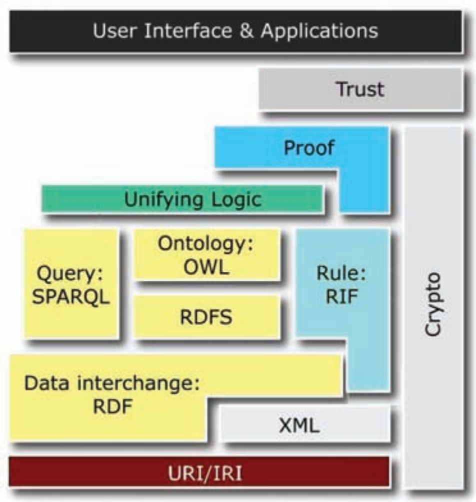
Rule Markup Languages

We characterize rule languages as rule markup languages if they are serialized in XML, employ URIs/IRIs for constants etc., and can interface with Web ontology languages.

RuleML

The Rule Markup Language (RuleML, www.ruleml.org) is a markup language developed to express a family of Web rules in XML for deduction, rewriting, and reaction, as well as further inferential, transformational, and behavioral tasks. It is defined by the Rule Markup Initiative (www.ruleml.org), an open network of individuals and

Figure 1. Semantic Web Layer Cake [adapted from (W3C, 2007)]



groups from both industry and academia that was formed to develop a canonical Web language for rules using XML markup and transformations from and to other rule standards/systems. It develops a modular, hierarchical specification for different types of rules comprising facts, queries, derivation rules, integrity constraints (consistency-maintenance rules), production rules, and reaction rules (Reaction RuleML, http://ibis.in.tum.de/research/ReactionRuleML), as well as tools and transformations from and to other rule standards/systems.

Datalog RuleML is defined over both data constants and individual constants with an optional attribute for IRI (URI) webizing . Atomic formulas have n arguments, which can be positional terms or, in Object-Oriented Datalog, slots (F-logic-like key->term pairs); OO Datalog also adds optional types and RDF-like oids/anchors, via IRIs (Boley, 2003). Inheriting all of these Datalog features, Hornlog RuleML adds positional or slotted functional expressions as terms. In Hornlog with equality, such uninterpreted (constructor-like) functions are complemented by interpreted (equation-defined) functions. This derivation rule branch is extended upward towards First Order Logic, has subbranches with Negation-As-Failure, strong-Negation, or combined languages, and is parameterized by 'pluggable' built-ins.

SWRL

The Semantic Web Rule Language (SWRL, www.w3.org/Submission/SWRL/) is defined as a language combining sublanguages of the OWL Web Ontology Language (OWL DL and Lite) with those of the Rule Markup Language (Unary/Binary Datalog).

The specification was submitted to W3C in May 2004 by the National Research Council of Canada, Network Inference (since acquired by webMethods), and Stanford University in association with the Joint US/EU ad hoc Agent Markup Language Committee.

Compared to Description Logic Programs (DLP), a slightly earlier proposal for integrating description logic and Horn rule formalisms by an overlapping authoring team, SWRL takes the opposite integration approach: DLP can be seen as the 'intersection' of description logic and Horn logic; SWRL, as roughly their 'union'. For DLP, the resulting rather inexpressive language corresponds to a peculiar looking description logic imitating special rules. It is hard to see the DLP restrictions, which stem from Lloyd-Topor transformations, being either natural or satisfying. On the other hand, SWRL retains the full power of OWL DL, but adds rules at the price of undecidability and a lack of complete implementations, although the SWRL Tab of Protégé has become quite popular (http://protege.cim3.net/cgi-bin/wiki.pl?SWRLTab).

Rules in SWRL are of the form of an implication between an antecedent (body) conjunction and a consequent (head) conjunction, where description logic expressions can occur on both sides. The intended interpretation is as in classical first-order logic: whenever the conditions specified in the antecedent hold, then the conditions specified in the consequent must also hold.

R2ML

R2ML (http://oxygen.informatik.tu-cottbus.de/rewerse-i1/?q=R2ML) was developed as a subproject in the EU Network of Excellence REWERSE (http://oxygen.informatik.tu-cottbus.de/rewerse-i1/). The R2ML project is about the design of integrity and derivation rules on the basis of the Rule Markup Language (RuleML) and the Semantic Web Rule Language (SWRL). R2ML defines a general markup framework for integrity rules, derivation rules, production rules and reaction rules. Rule concepts are defined with the help of MOF/UML, a subset of the UML class modeling language proposed by the Object Management Group (OMG) for the purpose of 'meta-modeling', i.e. for defining languages conceptually on the level of an abstract (semi-visual) syntax. From these MOF/UML language models concrete markup syntax is obtained by applying a mapping procedure for generating corresponding languages from parameterized schemas.

W3C RIF

The W3C Rule Interchange Format (RIF) Working Group (http://www.w3.org/2005/rules/wiki/RIF_Working_Group) is an effort, influenced by RuleML, to define a standard *Rule Interchange Format* for facilitating the exchange of rule sets among different systems and to facilitate the development of intelligent rule-based application for the Semantic Web. For these purposes, *RIF Use Cases and Requirements* (RIF-UCR) have been developed. The RIF architecture is conceived as a family of languages, called *dialects*. A *RIF dialect* is a rule-based language with an XML syntax and a well-defined semantics.

So far, the RIF working group has defined the *Basic Logic Dialect* (RIF-BLD), which semantically corresponds to a Horn rule language with equality. RIF-BLD has a number of syntactic extensions with respect to 'regular' Horn rules, including F-logic-like frames, and a standard system of built-ins drawn from *Datatypes and Built-Ins* (RIF-DTB). The connection to other W3C Semantic Web languages is established via *RDF* and *OWL Compatibility* (RIF-SWC). Moreover, RIF-BLD is a general Web language in that it supports the use of IRIs (Internationalized Resource Identifiers) and XML Schema data types. The RIF Working Group has also defined the *Framework for Logic Dialects* (RIF-FLD), of which RIF-BLD was shown to be the first instantiation. RIF-FLD uses a uniform notion of terms for both expressions and atoms in a Hilog-like manner.

Current efforts of the RIF Working Group are expected to introduce a *Core* (RIF-Core) in the intersection of RIF-BLD and a new *Production Rule Dialect* (RIF-PRD) influenced by OMG's PRR, which can then be further extended or supplemented by reaction rules.

Semantic Web Rule Languages

In contrast to the XML-based rule markup languages in the previous section, the Semantic Web rule languages described in this section are human-readable rule languages, using an ASCII syntax based, e.g., on the ISO Prolog syntax standard. Typically, they are designed as compact presentation languages for human consumption. While they may be serialized in an XML-based rule markup language such as RuleML or RIF, e.g. for interchange purposes, they can also be employed directly: dynamically interpreted by platform-specific rule engines (at runtime) or statically translated into executable code (at compile time).

TRIPLE

TRIPLE (http://triple.semanticweb.org/) was designed as a practical rule language for linked-data applications. It is an RDF query, inference, and transformation language for the Semantic Web extending F-logic with modules. TRIPLE rules have been used to implement RDFS and other schema languages.

N3 / Turtle

Notation3 (w3.org/TeamSubmission/n3/), more commonly known as N3, is a shorthand non-XML serialization of Resource Description Framework (RDF) models, designed with human readability in mind: N3 is much more compact and readable than RDF/XML serializations. N3 has several features that go beyond the serialization of RDF models, such as support for RDF-based rules. Supporting the triple pattern syntax of SPARQL, Turtle (w3.org/TeamSubmission/turtle/) is a simplified, RDF-only subset of N3.

Jena Rules

The default representation format in Jena (jena.sourceforge.net/) for a rule in the rule-based reasoner is a Java Rule object with a list of body terms (premises), a list of head terms (conclusions) and an optional name and an optional direction. However, in Jena2 a rather simple parser is included which allows rules to be specified in reasonably compact form in text source files.

Prova

Prova (http://www.prova.ws/) is both a (Semantic) Web rule language and a highly expressive distributed (Semantic) Web rule engine which supports complex reaction rule-based workflows, rule-based complex event processing, distributed inference services, rule interchange, rule-based decision logic, dynamic access to external data sources, Web Services, and Java APIs. Prova follows the spirit and design principles of the W3C Semantic Web initiative and combines declarative rules, ontologies and inference with dynamic object-oriented programming and access to external data sources via query languages such as SQL, SPARQL, and XQuery. One of the key advantages of Prova is its separation of logic, data access, and computation as well as its tight integration of Java, Semantic Web technologies and enterprise service-oriented computing and complex event processing technologies.

DESIGN AND CHARACTERISTICS OF WEB RULE LANGUAGES

General requirements that need to be addressed by a rule markup language include semantic expressiveness and clarity, computational efficiency and Web scalability, machine-readable and machine-interpretable syntaxes, usability by both human users and automated agents, compact representation, interchangeability with other formats, means for serialization and persistence, as well as tool support in authoring, parsing/generating, and verifying rules. An important property that refers to development-time software engineering quality is the extensibility of the language and its interoperability with other representation formats. In this section, general language design principles, together with a selection of four important issues and criteria for rule markup language design, are identified and characteristics of the current rule markup languages RIF, RuleML, R2ML, SWRL (DAML Rules) as well as specific Semantic Web rule languages are exemplified with respect to them. Further design issues and requirements for Web rule languages have been elaborated in, e.g., (Wagner et al, 2005), (Bry and Marchiori, 2005), (Boley, 2007), and (Paschke, 2007).

Language Design Principles

Given the large design space of rule languages and rule concepts, the specification of a rule markup language is a difficult integration and conceptualization challenge that calls for balancing many (interrelated) design choices with respect to semantics, syntax, and pragmatics. In this subsection, we will raise four (markup) language design principles and will illustrate the actual design choices of the current rule markup languages with examples.

1. Criteria of Good Language Design

Rule markup language should be clear, compact, precise and easily adaptable. They should strive to fulfill typical criteria for good language design (Codd, 1971) - as known from logic, databases and programming - such as minimality, referential transparency and orthogonality:

- *Minimality* requires that the language provides only a small set of needed language constructs, i.e., the same meaning cannot be expressed by different language constructs
- *Referential transparency* is fulfilled if the same language construct always expresses the same semantics regardless of the context in which it is used
- *Orthogonality* asks for pairwise independent language constructs, thus permitting their meaningful systematic combination

The RuleML family follows these design principles as far as possible and provides only a set of needed language constructs which can be applied in every meaningful combination. This leads to a compact homogeneous syntax which is easier to maintain, learn, read and understand by end users, as well as easy to process automatically by machines (e.g. translators).

SWRL and RIF, which build on RuleML, basically follow this compact minimalistic design approach. However, SWRL introduces a more fine-grained distinction of constructs than RuleML, e.g. of Atoms into various types of specialized atoms such as *classAtom*, *datarangeAtom*, and *invidiualPropertyAtom*, which can be formed from unary predicates (classes), binary predicates (properties), and equalities or inequalities.

R2ML introduces further differentiated types of terms and atoms. This leads to a rich structure-preserving syntax with many highly specialized constructs. For instance, variables in R2ML are provided in the form of *ObjectVariable* (i.e. variables that stand for objects), *DataVariable* (i.e. variables that stand for data literals), and *GenericVariable* (i.e. variables that do not have a type), whereas RuleML (as well as SWRL and RIF) only provide a generic Var construct. Like RuleML, R2ML defines the notion of an individual (constant) and distinguishes between objects and data with the notions of an *object name* and *data value*.

The main design goal of the specific Semantic Web rule languages such as TRIPLE, Jena, and (following ISO Prolog syntax) Prova is to provide a terse scripting syntax with a minimal set of needed constructs. RuleML's POSL syntax combines and extends the terse ISO Prolog and F-logic syntaxes.

2. Different Syntactic and Semantic Layers

A complete specification of Web rule languages consists of a formalization of their syntax, semantics and, often left implicit, pragmatics. As implied by their name, the syntax of markup languages always includes the concrete syntax of (XML) markup, perhaps indirectly through other languages such as via RDF/XML. Often, there is another more or less concrete syntax such as a compact shorthand or presentation syntax, which may be parsed into the XML markup. While a presentation syntax can already disregard certain details, an abstract syntax systematically replaces character sequences with abstract constructors, often in a (UML) diagram form or as an abstract syntax tree (AST). Together with different token dictionaries, it can be used to generate corresponding concrete syntaxes. The semantics is formalized in a model-theoretic, proof-theoretic, or procedural manner, sometimes in more than one. When rules and speech-act-like performatives, such as queries and answers, are transmitted between different systems, their pragmatic interpretation, including their pragmatic context, becomes relevant, e.g. in order to explain the effects of performatives - such as the assertion or retraction of facts - on the internal knowledge base (Paschke et al, 2007).

A general distinction of three modeling layers can be adopted from OMG's model driven architecture (MDA) engineering approach (http://www.omg.org/mda/):

- A platform specific model (PSM) which encodes the rule statements in the language of a specific execution environment
- A platform independent model (PIM) which represents the rules in a common (standardized) interchange format, a rule markup language
- A computational independent model (CIM) with rules represented in a natural or visual language

The *CIM level* comprises visual and verbal rendering and rule modeling, e.g. via graphical representation or a controlled natural language syntax for rules, mainly intended for human consumption. Graphical representations such as UML diagrams or template-driven/controlled languages can also be used as presentation languages.

In order to facilitate rule modeling, R2ML provides a UML-based Rule Modeling Language (URML) (Lukichev and Wagner, 2006) which allows visual rule modeling based on UML class models and OCL constraints. RuleML on the CIM level provides several tools that use a controlled natural rule language approach. Among them are TRANSLATOR (Hirtle, 2006), which is based on Attempto Controlled English (ACE) (Fuchs et al, 2006), the open source Reaction RuleML editor (http://ibis.in.tum.de/research/ReactionRuleML/index.htm#editor), which uses a template driven approach, and the commercial RuleManager (Ensig, 2007). The Protégé tool (http://protege.stanford.edu/) provides facilities for modeling SWRL rules, but only on the PIM level, i.e. rules are directly written in the concrete SWRL XML syntax. RIF, being a rather new standard under development, currently does not provide any such tool support.

The *PIM level* should enable platform-independent machine interpretation, processing, interchange and translation into multiple PSM execution syntaxes of concrete rule engines. Hence, the concrete XML (or RDF/XML-based) syntax of a Web rule language such as RuleML, SWRL or R2ML resides on this level, whereas the abstract syntax is on the borderline between the PIM and CIM levels. The abstract syntax can be defined, e.g., with the help of a suitably general grammar definition language such as the EBNF formalism, used, e.g., in the definition of the abstract syntax of OWL, RuleML, RIF, and SWRL, or with the help of a MOF/UML model, as, e.g., in PRR, R2ML, and RuleML. (Wagner et al, 2004) (Giurca and Wagner, 2005).

The *PSM level* is the result of translating/mapping PIM rule (interchange) languages into execution syntaxes which can be directly used in a specific execution environment such as a rule engine. A general distinction can be made between a compiled language approach, where the rules are statically translated into byte code (at compile time), as e.g. done in the rule engines Take (http://code.google.com/p/take/) and Drools (www.jboss.org/drools/) versus interpreted scripting languages, which are dynamically interpreted (at run-time), as e.g. in the rule engines Prova (Paschke, 2006b) and OO jDREW (Ball et al., 2005). While the compiled approach has obvious efficiency benefits, the interpreted approach is more dynamic and facilitates, e.g., updates at run-time. Often, Semantic Web Rule Languages are directly executable by their respective rule engines; hence reside on the PSM level. As an intermediate step between the concrete PSM level and the PIM level an abstract representation is often introduced, such as N3, which provides an abstract rule syntax based on the RDF syntax, or POSL and Prova, which both provide ANTLR grammars (http://www.jdrew.org/oojdrew/demo/translator, http://www.prova.ws/gram.html) which are transformed into ASTs as the basis for further translation into interchange markup languages such as RuleML or other, specific execution formats.

The correct execution of an interchanged PIM-level rule set serialized in a rule markup language depends on the semantics of both the rule program and the platform-specific rule inference engine (IE). To address this issue, the IE and the interchanged rule set must reveal their intended/implemented semantics. This may be solved via explicit annotations based on a common vocabulary, e.g. an (Semantic Web) ontology which classifies the semantics. Annotations describing the semantics of an interchanged rule set could even be used to find appropriate IEs on the Web to correctly and efficiently interpret and execute the rule program; for example, (1) by configuring the rule engine for a particular semantics in case it supports different ones, (2) by executing an applicable variant of several interchanged semantic alternatives of the rule program, or (3) by automatic transformation approaches which transform the interchanged rule program into a rule program with an applicable semantics; cf. XTAN (http://www.w3.org/2008/02/xtan/). Another approach is to specify additional meta test cases for testing typical properties of well-known semantics, where by the combination of succeeded and failed meta tests the unknown semantics of an IE can often be uniquely determined (Paschke, 2006).

We remark that, traditionally, rule-based systems have been supported by two types of inferencing algorithms: forward-chaining and backward-chaining. A general rule markup language, as a lingua franca, should support translation and interpretation of both reasoning directions, perhaps again using pragmatic annotations (where by default chaining should be bidirectional, as with the direction attribute in RuleML).

Independently from the semantics of an interchanged rule program, the pragmatic context in which the interchange takes place is important for the target environment, in order to know how the received information should be used and which actions should be taken with respect to the pragmatic aspects. A standard nomenclature of pragmatic performatives is defined by the Knowledge Query and Manipulation Language (KQML) (www.cs.umbc.edu/kqml/) and the FIPA Agent Communication Language (ACL) (FIPA, 2000), which define several speech-act-theory-based communicative acts.

3. Modular Specialized Schema Layers vs. Flat General Schema

There are two basic design principles for the concrete rule markup syntax. The language (or language family) may be implemented in one flat (monolithic) general XML schema or in a layered structure, where semantically related constructs are defined within separate modules that are added to the different language layers of the Web rule language (cf. http://www.ruleml.org/modularization/) . This leads to a hierarchical structure where higher language layers build on sublayers and add more expressiveness by extending them. The layers are not necessarily organized around expressiveness/efficiency to the language core.

R2ML follows the first approach and provides one quite large, flat XML schema for all different rule types and language constructs. In contrast, RuleML (also SWRL and RIF) follow the layered design principle and define new constructs within separate modules which are added to the respective layers in the RuleML language family. The layered and uniform design makes it easier to learn the language and to understand the relationship between the different features, and it provides a certain guidance to users who might be interested only in a particular subset of the features and who do not need support for the full expressiveness of the language. The modularization allows for easy extension of the language's representation capabilities, using the extensibility of XML Schema (e.g. a redefine of an XML Schema group definition), without breaking the core language standard. This development path provides a stable, useful, and implementable language design for rule developers to manage the rapid pace of change on the Semantic Web and modern rule systems. Apart from that, modules facilitate the practical and extensible development of a rule language family by bundling language constructs into layers which can be developed, compiled, tested and managed separately. The modularization also enforces the principle of information hiding and can provide a basis for data abstraction. However, a monolithic schema is easier to read by humans than an unevenly modularized one and, by now, some of the extant XML processing tools and editors do not fully support modular XML Schema definitions. This calls for flattening a layered schema on demand via automatic modular-to-monolithic translators, thus combining the advantages of modular development and maintenance with the advantages of monolithic delivery (for some validators and (object model) transformers, e.g. based on JAXB https://jaxb.dev.java.net/).

4. XML Elements vs. Attributes

A general question regarding the implementation of a concrete rule markup language is where to use XML elements and where attributes to define the rule constructs and the rule information content. A general discussion on this element-vs.-attribute issue can be found in the OASIS Cover pages (http://xml.coverpages.org/elementsAndAttrs.html):

- If the information in question could be itself marked up with elements, put it in an element.
- If the information is suitable for attribute form, but could end up as multiple attributes of the same name on the same element, use child elements instead.
- If the information is required to be in a standard DTD-like attribute type such as ID, IDREF, or ENTITY, use an attribute.
- If the information should not be normalized for white space, use elements. (XML processors normalize attributes in ways that can change the raw text of the attribute value.) (cf. http://www.ibm.com/developerworks/xml/library/x-eleatt.html)

Accordingly, RuleML's general markup conventions provide common principles for its language hierarchy. XML elements are used for representing language constructs as trees while XML attributes are used for distinguishing variations of a given element and, as in RDF, for webizing. Variation can thus be achieved by different attribute values rather than requiring different elements. Since the same attribute can occur in different elements, an orthogonal, two-dimensional classification ensues, which has the potential of quadratic tag reduction.

For example, recent work in RuleML led to orthogonal dimensions extending the RuleML 0.9 role tags for argument, <arg . . .>, and slots, <slot>. So far, the *unkeyed* <arg index="..."> was always *ordered*, as indicated by the index attribute, and the *keyed* <slot> was always *unordered*, as indicated by the lack of an index attribute. This was generalized by allowing an optional index attribute for both role tags, as shown by the independent distinctions in the following key-order matrix:

```
ordered unordered
keyed <slot index=”...”> <slot>
unkeyed <arg index=”...”> <arg>
```

Two extra orthogonal combinations are obtained from this system. First, *keyed*, *ordered* children permit positionalized slots, as in this cost fact (see Figure 2).

Here, slot names item, price, and taxes are provided, e.g. for readability, as well as index

Figure 2.

```
<Atom>
  <Rel>cost</Rel>
  <slot index="1"><Ind>item</Ind><Ind>jewel</Ind></slot>
  <slot index="2"><Ind>price</Ind><Data>6000</Data></slot>
  <slot index="3"><Ind>taxes</Ind><Data>2000</Data></slot>
</Atom>
```

positions 1-3, e.g. for efficiency.

Second, *unkeyed*, *unordered* children permit elements acting like those in a bag (finite multiset), as in this transport fact:

```
<Atom>
<Rel>transport</Rel>
<arg><Ind>chair</Ind></arg>
<arg><Ind>chair</Ind></arg>
<arg><Ind>table</Ind></arg>
</Atom>
```

Here, the arguments are specified to be commutative and 'non-idempotent' (duplicates are kept).

For a general discussion of positional vs. unordered representations see (Boley, 2006).

R2ML differs from RuleML, SWRL and RIF as it implements an attribute solution and defines user information content in attributes. For instance a typed object variable "?driver" is represented as follows:

```
<r2ml:ObjectVariable
r2ml:name="driver"
r2ml:classID="userv:Driver" />
```

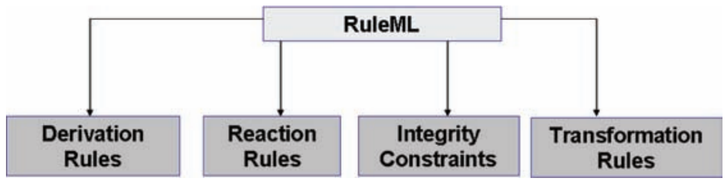
A distinction of positional vs. slotted (named-argument) predicates and functions, as in RuleML and RIF, does not exist in R2ML.

Expressive Layering

From the perspective of knowledge representation, the main adequacy criterion for a rule markup language is its *epistemological adequacy*, which addresses the ability of the language to represent all relevant knowledge under consideration. Among other representation issues, a general rule interchange format should allow to coherently represent derivation rules, reaction rules, integrity rules, and deontic rules in a homogeneous syntax (Wagner et al, 2005). We use the following general rule classification:

- Facts may comprise various kinds of information such as asserted atoms (formulas), individual-class memberships (of ontology classes), (object-oriented) instances, stored data (e.g., relational, XML), states and event occurrences which might be qualified, e.g., by priorities, temporally, etc.
- Derivation rules infer conclusions from conditions (as in Datalog and Horn logic), where facts (see above) are a special case with constantly true conditions
- Transformation rules specify term rewriting, which can be considered as derivation rules of logics with (oriented) equality
- Integrity rules (or integrity constraints) are assertions which express conditions (or queries) that must always be satisfied. Besides enforcing data integrity, they can constrain, e.g., the rule system structure, its information content, or its behavior:

Figure 3. RuleML rule language family



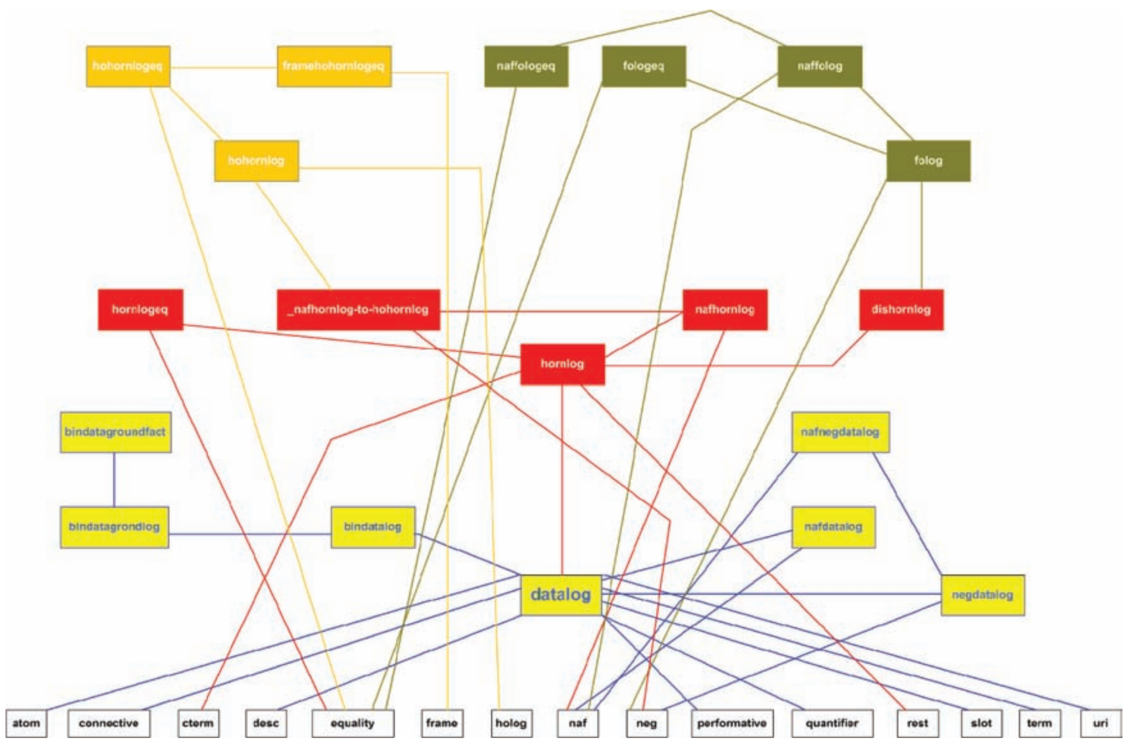
   - *Structural constraints* (deontic assignments)
   - *State constraints*
   - *Process constraints*
- Deontic rules describe rights and obligations, e.g., of institutions and agents in the context of evolving states (situations triggered by events/actions) and state transitions, where integrity rules (see above) are a special case ('introspectively') affecting the rule set itself
- Reaction rules are (behavioral / action) rules that react on occurred events (external events or changed conditions) by executing actions, where production rules are a special case with events restricted to changed conditions

Because of this broad variety of rules relevant to the Semantic Web, a general rule markup language such as RuleML should have a hierarchical structure that reflects the relevant rule dialects or sublanguages, covering the knowledge representation needs of various subcommunities (cf. Figures 3 and 4).

The main branches contain subbranches, e.g. derivation rules monotonically disallowing or non-monotonically allowing negation as failure (NAF). Branches also exhibit a layering structure with sublanguages of different expressive power, e.g. for monotonic derivation rules proceeding from function-free Horn logic (Datalog) through either full Horn logic or disjunctive Datalog up to first-order logic (FOL). This hierarchy does not form a strict tree (but a directed acyclic graph), i.e. some sublanguages are shared by several more expressive languages; e.g., a sublanguage of conditions is shared by derivation rules and reaction rules. A basic classification of rule languages is introducted by the RuleML family with Derivation Rule Markup Languages (that is, Derivation RuleML), Production Rule Markup Languages (that is, Production RuleML which is a subfamily of Reaction RuleML), Reaction RuleML Markup Languages (that is, Reaction RuleML, http://ibis.in.tum.de/research/ReactionRuleML/), and Transformation Markup Languages (that is, Functional RuleML, http://www.ruleml.org/fun/), as well as other specializations, e.g. dialects for deontic rules (e.g. covered by the RBSLA language (Paschke, 2005)), defeasible rules (Defeasible RuleML, http://defeasible.org/RuleML/) and uncertainty / fuzziness (Fuzzy RuleML, http://www.image.ntua.gr/FuzzyRuleML/).

The coverage of rule markup languages can be roughly divided into logic-based rule formalisms, usually variants of first-order predicate logic, and not logic-based rule formalisms such as (early) production rule systems. In this chapter, we mainly focus on logic-based derivation rule markup languages, but most of the general rule markup languages such as RuleML and R2ML also support the serialization of production rules and reaction rules (e.g., Reaction RuleML covers different types of reaction rule languages). In fact, RuleML and R2ML provide a roughly similar coverage, whereas, e.g., SWRL acts as a more specialized language for homogeneously combining Datalog rules with OWL, hence does not cover, e.g., reaction rules.

Figure 4. RuleML 0.91 derivation RuleML subfamily



Rules and Object Descriptions

With its URIs (http://www.w3.org/Addressing/), the Web provides a global addressing (URL) and naming (URN) mechanism for objects. A URI consists of a URI scheme name (http:, file:, etc.) followed by other information that is interpreted relative to the URI scheme. The method for assigning meanings to names varies from one URI scheme to the next, and within each scheme for different sets of names. Each scheme's specification describes how its URIs are intended to be used in certain contexts. As a result, any naming framework must provide mechanisms to enable the creation of new names while avoiding conflicts with existing ones. URIs are also central to the Semantic Web, where RDF metadata are used to describe those objects or resources with classes and properties, which are themselves defined by ontologies (in RDF Schema or OWL). Since SHOE (Heflin et al, 1999), Semantic Web rule languages have associated URIs with constant symbols, predicate names, and other language constructs for reference and disambiguation. For example, the constant symbol Georgia could be associated with the unique 'homepage' URI http://www.georgia.gov to refer to and disambiguate the state in the Southeastern U.S. in contrast to other entities having the same English name such as the country at the east coast of the Black Sea.

There have been attempts to differentiate the Web notion of URIs into two subnotions, as discussed in (Halpin, 2006): URLs (Uniform Resource Locators), for access, and URNs (Uniform Resource Names), for naming. This distinction is independent from the recent IRI (Internationalized) versions of URIs. In the context of Web knowledge representation, especially for Web rules as explored in POSL, RuleML, and RIF, three central URI uses are emerging (Boley, 2007), given here in the order of further needed research (orthogonal to research in URI normalization (Boley, 2003)).

First, a URI can be used, URL/access-style, for module import (transitive import for nested modules), where it is an error if dereferencing the URI does not yield a knowledge base valid with respect to the expected representation language.

Second, a URI can be used, URN/naming-style, as the identifier of an individual constant in the representation language, where URI dereferencing is not intended as part of the formal knowledge representation. If dereferencing is attempted as part of the metadata about the informal knowledge representation, it should retrieve a descriptive 'homepage' about the individual.

Third, a URI can be used, naming-style, as the identifier of a class, property, relation, or function, and at the same time, access-style, where dereferencing yields (a "#"-anchor into) a knowledge base formally defining that identifier (albeit perhaps partially only, as for an RDF Schema knowledge base just giving the superclasses of a class).

Here are examples for the three URI uses in connection with rules.

First, a module of U.S. states could be imported into the current rulebase using the URL/access-style URI http://modeg.org#us-state.

Second, the URI http://en.wikipedia.org/wiki/Pluto can be used URN/naming-style to refer to a celestial body originally considered a planet, as in this rule specifying its years of planethood (a URI is enclosed in a pair of angular brackets, <... >):

```
planet(<http://en.wikipedia.org/
wiki/Pluto>,AD[?year]):-
lessThanOrEqual(1930,?year),
lessThanOrEqual(?year,2006).
```

As part of the formal rule knowledge, the Pluto URI is employed only for naming. The rule can also be employed as metadata about informal knowledge through ('semantic search engine') queries like planet(?which,2005), because one of its solutions will bind ?which to the URI, whose dereferencing ('clicking') will then retrieve Pluto's Wikipedia entry.

Third, for certain formal purposes a URI like http://termeg.org#MiniVan is needed just to provide a name; for other formal purposes, also to provide a total or partial definition found by using that same URI access-style (say, the partial definition of being rdfs:subClassOf both http://termeg.org#Van and http://termeg.org#PassengerVehicle).

In most rule markup languages as well as the specific Semantic Web rule languages, the (user-defined) vocabulary names are globally unique standard identifiers in the form of URI references. Morover, they often define specific builtins for handling URIs such as the SWRL builtins *swrlb:resolveURI (from XQuery op:resolve-uri)*, which is satisfied iff the URI reference in the first argument is equal to the value of the URI reference in the second argument resolved relative to the base URI in the third argument, or *swrlb:anyURI*, which is satisfied iff the first argument is a URI reference consisting of the scheme in the second argument, host in the third argument, port in the fourth argument, path in the fifth argument, query in the sixth argument, and fragment in the seventh argument.

All classes in R2ML are URI references. A class is a type entity for R2ML objects and object variables. Similarily, a *reference property* as well as a *datatype predicate* in R2ML is a URI reference.

RIF uses *internationalized resource identifiers* or *IRIs* (symbol space rif:iri) as constants similar to RDF resources.

Rule-Ontology Combination

A rule markup language should be reasonably integrated with the Semantic Web and should be able to refer to external Semantic Web vocabularies by means of URIs or IRIs, e.g. to use their taxonomic vocabularies as type systems and their individuals as external constants/objects. Domain-independent rules can then be interpreted (relative to each vocabulary) in a domain-dependent manner (with a precise semantics). Accordingly, the original rule set can be much easier interchanged and managed/maintained in a distributed environment. Also, the core Web rule language stays compact and can be easily extended for different vocabulary languages (RDFS, OWL, OWL 2, etc.) on a "per-need-basis".

In recent years, quite an effort has been made to develop a dual expressiveness layering of assertional and terminological knowledge as well as their blends (Antoniou et al., 2005, Kifer et al, 2005). To retain decidability of querying, the *assertional bottom layer* usually consists of Datalog (function-free) assertions, perhaps restricted to unary/binary predicates. For the *terminological bottom layer*, an irreflexive version of RDF Schema's subClassOf can be employed, which could later be extended towards the rhoDF (Munoz et al., 2007) fragment of RDF. The two layers can be blended through a hybrid combination (rhoDF classes used as types for Datalog constants and variables, and subClassOf defined with order-sorted semantics) or a homogeneous integration (rhoDF classes used as unary predicates in the body of Datalog rules, and subClassOf defined as special rules with Herbrand-model semantics).

The higher layers can develop Datalog into Horn (as in Prova's or OO jDREW's hybrid implementation) and FOL (First-Order Logic) assertions, rhoDF into ALC and SHIQ terminologies with classes and properties, and appropriate blends (Rosati, 2006) (Rosati, 2006a), e.g. as advancements of our hybrid DatalogDL (Mei et al, 2007b) or homogeneous ALCuP (Mei et al, 2007). For certain purposes, especially in the early modeling phases, the assertional layers can move even beyond FOL, including towards higher-order and modal logics, as started as part of the RuleML family (Boley, 2006).

To permit the specification of terminologies independent of assertions, a hybrid approach is proposed here adopting the CARIN (Levy and Rousset, 1998) principle as a working hypothesis: A terminological predicate is not permitted in the head of a rule. Intuitively, terminological classes cannot be (re)defined by assertional clauses, because a terminology establishes more stable 'background' knowledge extended by assertions that constitute more volatile 'foreground' knowledge.

Such a hybrid lower layer can use sort restrictions as simple terminological queries in Datalog rule bodies, which in higher layers are extended to terminological queries involving properties, ALC expressions, etc. In the spirit of (Kifer et al, 2005), this should lead to a more realistic Semantic Web architecture with simplified foundations and better computational properties. Our fine-grained bottom-up approach also complements the recent differentiation of OWL-Lite into OWL 1.1 (later: OWL 2) Tractable Fragments (Grau et al., 2006).

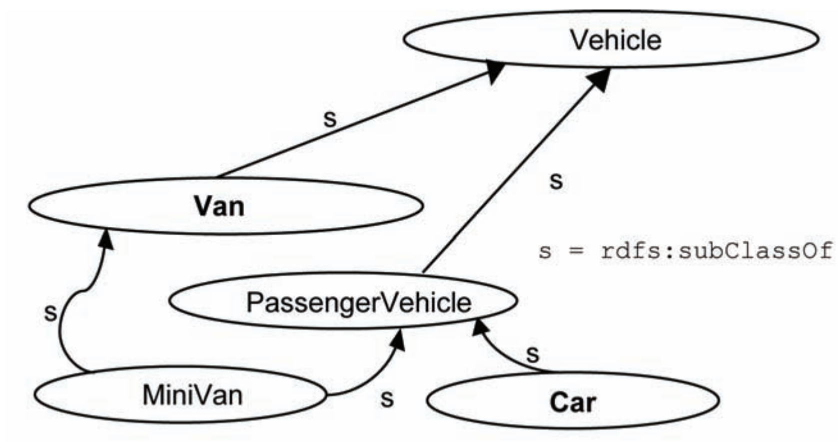
The following example uses classes of a subClassOf terminology as variable sorts (types) of slightly extended Datalog rules, namely of Horn logic rules employing (unary) functions only for measurement units.

The terminology forms a DAG (cf. Figure 5) that introduces Vehicle-rooted classes and exemplifies multiple inheritance of MiniVan from Van and PassengerVehicle (the ">" infix is used between a superclass and a subclass):

```
Vehicle > Van
Vehicle > PassengerVehicle
Van > MiniVan
PassengerVehicle > MiniVan
PassengerVehicle > Car
```

The following RuleML/POSL rules specify registration fees for vehicles. The first rule specifies a vehicle variable typed by the Van class, while the second refers to the Car class (the ":" infix is used between a variable and its type):

Figure 5. A vehicle terminology (cf. (W3C, 2000))




```
registration(?V:Van,CAD[?R:Deci
mal]):-
emission(?V,CO2[?E]),
weight(?V,kg[?W]),
emiweight(CAD[?R],CO2[?E],kg[?W
]).
registration(?V:Car,CAD[?R:Deci
mal]):-
emission(?V,CO2[?E]),
speed(?V,kmh[?S]),
emispeed(CAD[?R],CO2[?E],kmh[?S
]).
```

A registration query for a given vehicle class will thus unify only with correspondingly sorted rule conclusions, hence directly branch into the appropriate rule premises (the emiweight and emispeed premises compute the fees from the emissions as well as the weights and speeds for Vans and Cars, respectively). The previous section has shown URI-'webized' versions of these terminological classes.

R2ML rules may refer to a vocabulary which can be R2ML's own vocabulary or an imported one (such as RDF(S) and OWL). R2ML's internal default vocabulary is a serialization of a UML fragment of class diagrams. R2ML uses XML Schema datatypes as its standard datatype set. R2ML distinguishes between plain and typed literals and typed and untyped variables. A *DataTerm* is a *DataVariable*, a *DataLiteral*, or a *DataFunctionTerm*. An *ObjectVariable* is a variable that stand for objects of a particular class type, a *DataVariable* is a variable that stand for data literals, and a *GenericVariable* is a variable that does not have a type. For instance, an *ObjectVariable* contains an optional reference to a class which is used as its type:

```
<r2ml:ObjectVariable
r2ml:name="driver"
r2ml:classID="userv:Driver" />
```

R2ML allows both typed and untyped individuals. For instance, a *TypedLiteral* consist in a lexical value and a type that is an RDF datatype or a user defined datatype (subclass of rdfs:Literal):

```
<r2ml:TypedLiteral r2ml:type
Literal="xs:positiveInteger"
r2ml:lexicalValue="90"/>
```

RuleML supports order-sorted terms permitting typed individuals, variables (exemplified above) and data literals (Boley, 2003). Therefore, RuleML/XML defines an optional type attribute for specifying a term's (user-defined) type. Besides referring to the default XML Schema datatypes, typed terms may also link into external object class hierarchies via their fully qualified class names (e.g. Java classes) or taxonomies such as RDF Schema class hierarchies, thus reusing the OO class models and Semantic Web's light-weight ontologies as pluggable external order-sorted type systems. For example:

```
<Var type="rdf://
owl:Vehicle">V</Var>
<Ind type="xml://
xsd:string">abc</Ind>
<Var type="java://java.lang.
Number">X</Var>
```

A *Data* term in RuleML contains a fixed value, like an RDF literal. It may be optionally associated with an XML Schema built-in datatypeusing the xsi:type attribute. For example:

```
<Data
xsi:type="xs:dateTime">2002-10-
10T17:00:00Z</Data>
```

This open order-sorted typing approach of RuleML provides higher levels of abstractions and allows ad-hoc polymorphism with respect to coercion, i.e., automatic type conversion between subtypes, and overloading, i.e., defining multiple cases (rules with the same head except for types) taking different argument types. The ability to integrate external Semantic Web vocabularies, data types and object oriented class hierarchies as type systems provides syntactic expressiveness for easy extension of the language with domain-specific terminologies, and it facilities rule interchange across domain boundaries due to the explicit semantic definition of the used vocabulary, e.g., a Semantic Web ontology.

As in RuleML, RIF provides an optional type attribute for typed constants / individuals and a set of default XML Schema primitive data types such as xsd:long, xsd:integer, xsd:decimal, xsd:string, rdf:XMLLiteral and rif:text. For example:

```
<Const
type="rif:iri">dc:creator</
Const>
<Const type="xsd:string">abc</
Const>
```

However, variables are not typed directly in a prescriptive form using, e.g., the type attribute in the variable construct to denote that a variable *X* is of type *T*, i.e. *X:T*. Instead, RIF defines classification terms for class memberships, and also for subclass relationships (cf. F-logic's ":" and "::"):

- t#s is a *membership term* if t and s are terms.
- t##s is a *subclass term* if t and s are terms.

These classification terms are used to describe subclass hierarchies and membership constraints, e.g. expressing that a variable is of a certain class (type).

In SWRL, a homogeneous combination of OWL (OWL DL and OWL Lite) and RuleML (Unary/Binary Datalog), atoms can be of the form *C(x)*, *P(x,y)*, *sameAs(x,y) differentFrom(x,y)*, or *builtIn(r,x,...)* where *C* is an OWL description or data range, *P* is an OWL property, *r* is a built-in relation, *x* and *y* are either variables, OWL individuals or OWL data values, as appropriate. In the context of OWL Lite, descriptions in atoms of the form *C(x)* may be restricted to class names. That is, SWRL defines a rule language on top of OWL ontologies and hence directly supports the definition of class ontologies and their properties which can be used to type variables. For an example, see Figure 6.

Figure 6.

```
<!-- Each person that is a qualified driver can be added to a car rental as
additional driver-->
<ruleml:Implies
  xmlns:ruleml="http://www.ruleml.org/0.91/xsd"
  xmlns:owlx="http://www.w3.org/2003/05/owl-xml"
  xmlns:swrlx="http://www.w3.org/2003/11/swrlx"
  xmlns:srv="http://www.eurobizrules.org/ebrc2005/eurentcs">

  <ruleml:body>
    <swrlx:classAtom>
      <owlx:Class owlx:name="srv:Rental"/>
      <ruleml:Var>rental</ruleml:Var>
    </swrlx:classAtom>
    <swrlx:classAtom>
      <owlx:Class owlx:name="srv:Person"/>
      <ruleml:Var>person</ruleml:Var>
    </swrlx:classAtom>
    <swrlx:classAtom>
      <owlx:Class owlx:name="srv:QualifiedDriver"/>
      <ruleml:Var>person</ruleml:Var>
    </swrlx:classAtom>
  </ruleml:body>
  <ruleml:head>
    <swrlx:individualPropertyAtom swrlx:property="srv:additionalDriver">
      <ruleml:Var>rental</ruleml:Var>
      <ruleml:Var>person</ruleml:Var>
    </swrlx:individualPropertyAtom>
  </ruleml:head>
</ruleml:Implies>
```

This homogeneous integration approach is adopted by R2ML from SWRL, while in RuleML without OWL one would refer to the external ontology which defines the vocabulary classes and their properties.

EXTERNAL DATA INTEGRATION AND DATA PROCESSING

Often Web rules refer to or describe functions and queries over data stored in an external database which can be anything from log files to Web sources or relational databases, data warehouses, or XML or RDF databases such as native XML databases or RDF triple stores. The rule language must allow for the direct dynamic integration of these secondary data storages as facts or object values into the rules in order to reduce redundancy and high memory consumption. It should also support outsourcing of expensive (pre-)processing of data to external systems, e.g., of mathematical functions to procedural implementations such as Java, or of SQL/SPARQL aggregation queries (constructive views) to database management systems or RDF triple stores. A tight combination of declarative and object-oriented programming with rich procedural attachments and language built-ins, e.g. for querying, will facilitate the integration of existing functionalities, tools, and external data sources into rule executions at run time.

Procedural attachments are procedure calls to external user-defined computational models of a standard programming language, e.g., directly to

Figure 7.

```
<swrlx:builtinAtom swrlx:builtin="&swrlb;#multiply">
    <ruleml:var>inches</ruleml:var>
    <ruleml:var>feet</ruleml:var>
    <owlx:DataValue owlx:datatype="&xsd;#int">12</owlx:DataValue>
</swrlx:builtinAtom>
```

Java or C# methods. Therefore, procedural attachments are a crucial extension of a modern Web rule language. They permit the combination of the benefits of declarative (rule-based) as well as procedural and object-oriented languages, e.g., to delegate computation-intensive tasks to optimized object code or to invoke procedure calls on object methods which cannot be easily expressed in a declarative rule-based way. Procedural attachments should be supplemented with a typed logic approach with external type systems such as Java or Semantic Web ontologies, e.g. to assign external objects to typed variables, and with mode declarations in order to safeguard the usage of built-ins and calls to external functionalities.

(Procedural Attachments). A procedural attachment is a function or predicate whose implementation is given by an external procedure. Two types of procedural attachments are distinguished:

- *Boolean-valued attachments (or predicate attachments)* which call methods that return a Boolean value, i.e., that are of Boolean sort (type).
- *Object-valued attachments (or functional attachments)* which are treated as functions that take arguments and return one or more objects, i.e., that are of a function sort. This also includes access to public object fields.

(Built-Ins). Built-in predicates or functions are special restricted predicate or function symbols in the rule language for concrete domains, e.g., built-ins for strings, numerics, Boolean values, date, time, intervals, lists, etc.

All Web rule languages discussed in this chapter provide support for built-ins and some of them also for general procedural attachments.

SWRL provides an extensible library of built-in functions (http://www.daml.org/2004/04/swrl/builtins.html) co-developed with RuleML. SWRL's built-ins approach is based on the reuse of existing built-ins in XQuery and XPath, which are themselves based on XML Schema Part 2: Datatypes. SWRL built-ins are called via a built-in atom, swrlx:builtinAtom, which identifies a built-in using the swrlx:builtin attribute and lists its arguments as subelements. SWRL built-ins are identified using the http://www.w3.org/2003/11/swrlb namespace, currently also used by RuleML. This is an example of calling the multiply built-in (see Figure 7).

SWRL does not provide direct support for procedural attachments, but it could easily adopt this feature from Reaction RuleML.

R2ML by default supports SWRL and XPath2 built-ins as predicate names of atoms *r2ml:DatatypePredicateAtom* and symbols of functions *r2ml:DatatypeFunctionTerm*. Functions and operators like addition, subtraction, etc. are translated into corresponding R2ML function terms. The operands of the functions implied by the built-ins are enclosed by *r2ml:dataArguments* and might be class attributes, class operations, data variables, typed variables, or further nested built-in functions. However, mode declarations are missing (see Figure 8).

R2ML supports procedural attachments in order to access public data fields of objects which might be bound to object variables (see

Figure 8.

```
<r2ml:DatatypePredicateAtom
 r2ml:datatypePredicate="swrlb:lessThan">
   <r2ml:dataArguments>
      <r2ml:DataVariable r2ml:name="y"/>
      <r2ml:TypedLiteral r2ml:lexicalValue="4"
                         r2ml:datatype="xs:positiveInteger"/>
   </r2ml:dataArguments>
</r2ml:DatatypePredicateAtom>
```

Figure 9.

```
<r2ml:AttributeFunctionTerm r2ml:attributeID="userv:Car.price">
   <r2ml:contextArgument>
      <r2ml:ObjectVariable r2ml:name="car"
                           r2ml:classID="userv:Car"/>
   </r2ml:contextArgument>
</r2ml:AttributeFunctionTerm>
```

Figure 9).

For reactive rules such as production rules, R2ML supports assignments of action expressions in order to call object methods as actions in the action part (see Figure 10).

The RIF built-ins (http://www.w3.org/2005/rules/wiki/DTB) overlap with the functions and predicates defined in XQuery 1.0 and XPath 2.0 Functions and Operators.

Syntactically, built-in predicates and functions in RIF are enclosed by external terms of the form:

```
'External' '(' Expr ')'
```

where Expr is a UNITERM, i.e. either a Boolean-valued function expression / predicate or an object-valued functional expression. Since RIF does not support a general typed rule language, it requires special guard predicates for all of its supported datatypes to ensure the correct usage of the arguments of built-ins:

Figure 10.

```
<r2ml:AssignActionExpression r2ml:propertyID="status">
   <r2ml:contextArgument>
      <r2ml:ObjectVariable r2ml:name="ticket"/>
   </r2ml:contextArgument>
   <r2ml:TypedLiteral r2ml:lexicalValue="Escalate"
                      r2ml:datatypeID="xs:string"/>
</r2ml:AssignActionExpression>
```

Figure 11.

```
<!— a call to a builtin function -->
<Expr>
   <Fun per="builtin" uri="swrlb:stringConcat"/>
   <Var type="java://java.lang.String" mode="+">String1</Var>
   <Var type="java://java.lang.String" mode="+">String2</Var>
</Expr>

<!-- a call to a builtin predicate -->
<Atom>
   <Rel per="builtin" uri="rif:dateTime-equal"/>
   <Var type="xml://xs:dateTime" mode="+">Time1</Var>
   <Var type="xml://xs:dateTime" mode="+">Time2</Var>
</Atom>
```

```
External("op:numeric-greater-
than"^^rif:iri(
?diffdays
"10"^^xsd:integer))
External(
"www.w3.org/2007/rif-builtin-
predicates#isInteger"^^rif:iri(
?diffdays))
```

Note that the above example shows the RIF presentation syntax, not the concrete RIF XML syntax.

RIF-FLD foresees procedurally attached user-defined function terms or predicates to be wrapped as external terms but does not define a concrete approach for calling procedural actions yet. However, it supports frame terms *t[p1->v1 ... pn->vn]* which can be used to describe properties of objects.

The RuleML family (through its Reaction RuleML branch) provides an open flexible approach for pluggable external built-in libraries safeguarded by type and mode declarations. It explicitly denotes the usage by the attribute per= "*plain|value|effect|modal|builtin*" on functional expressions and atomic relations. A <*Rel*> or <*Fun*> using "plain" is left uninterpreted, using "*value*" is interpreted purely for its value, using "*effect*" is interpreted impurely both for its value and its (side-)effect action, e.g. by a procedural attachment, using "*modal*" is interpreted as pure modality, and using "*builtin*" as a built-in (Figure 11).

The mode ("+": input; "-": output; "?": input or output) and type declarations ensure the correct usage of arguments in built-ins, i.e. that built-ins are called with ground values (not free variables) of the expected types.

RuleML provides a concise integration of procedural attachments. Methods of external object classes can be called, including calls to object constructors and calls to object instance and static object methods as well as access to public object fields. Constructed objects and returned result objects can be assigned to variables. Nested selection patterns can be defined over the result object collections such as "forall ?X where ?X=Person(age > 30 and age < 40)" (see Figure 12).

Most of the specific Semantic Web rule languages such as Jena and Prova support all "standard" built-ins of Web Rule languages as well as many additional built-ins e.g. for meta interpretations of literals, exception handling, console printouts, collections, iterations/enumerations, object property constraints, or access to system environment properties. For instance, Prova supports several query built-ins to access

Figure 12.

```
<!-- Assign the constructed Java object to the variable
     Date = java.io.Calendar.getInstance() -->
<Equal>
  <Var>Date</Var>
  <Expr>
     <!-- class -->
     <oid><Ind uri="java://java.util.Calendar"/></oid>
     <!-- constructor -->
     <Fun per="effect">getInstance</Fun>
  </Expr>
</Equal>

<!-- Use the bound object of the variable and call a function
     "isSet" of the object -->
<Atom>
   <!-- object previously assigned to Date -->
   <oid><Var>Date</Var></oid>
   <Rel per="effect">isSet</Rel>
   <Data>1</Data>
</Atom>

<!-- Call a static C# method -->
<Atom>
  <oid><Ind uri="c-sharp://System.Console"/></oid>
  <Rel uri="WriteLine"/>
  <Data>Hello World</Data>
</Atom>
```

files, XML data sources via DOM, XPath, and XQuery, RDF data sources via RDF triples and SPARQL and RDFS/OWL ontologies, as well as various homogeneous or heterogeneous inference queries using external DL reasoners such as, e.g., Pellet. Prova also provides a tight and natural Java integration. Methods of classes in arbitrary Java packages can be dynamically invoked from Prova rules. The method invocations include calls to Java constructors creating Java variables and calls to instance and static methods for Java classes as well as public object data fields.

FUTURE TRENDS

A general rule markup language such as RuleML or RIF covers many different rule types and rule families. Some of the language families such as classical production rules historically only define an operational semantics, while other rule families such as logical rules (see RuleML family in the section about expressive layering) are based on a model-theoretic and/or proof-theoretic semantics. A general research question is whether there exists a unifying semantic framework for all different rule types. Work in this direction is pursued, e.g. in RIF (http://www.w3.org/2005/rules/wiki/FLD) and Reaction RuleML (transactional transition semantics for reaction rules subsuming all other RuleML rules). However, since there is no general consensus on one particular semantics for all expressive rule languages, an exclusive commitment to one particular semantics for a Web rule language should be avoided (even in well-researched fields such as logic programming several semantics such as well-founded semantics and answer set semantics are competing). Nevertheless, for certain subfamilies a preferred semantics can still be given and semantic mappings between rule families be defined.

General rule markup languages need to include practical language constructs which might not (yet) have a standard formal semantics based on classical model-theoretic logic. For instance, procedural calls to external (object) functions, operational systems, data sources and terminological descriptions, are often vital to deal with practical real-world settings of distributed Web applications. Recent research, e.g. in RuleML and RIF-PRD, is done on adopting such practical language constructs without a standard formal semantics but with a non-standard one. While there is a risk that these concessions to non-standard semantics might endanger the benefits of formal semantics for the overall rule language, they turn out to be a crucial means to avoid limitations of standard rule representations in the exploration of rule markup languages. The rule component will rarely run in isolation, but interact with various external components, hence call for functionalities such as efficient object-oriented or relational/SQL-style retrieval and aggregation methods that are common in modern information systems.

Further examples of useful practical constructs are the annotation of rules and rule sets with additional metadata such as rule qualifications, rule names, module names, Dublin Core annotations, etc., which eases, e.g., the modularization of rules into rule sets (bundling of rules), the creation of constructive views over internal and external knowledge (scoped reasoning), as well as the publication and interchange of rules / rule sets on the Web. Advanced rule qualifications such as validity periods or rule priorities might for example safeguard dynamic updates (e.g. the incorporation of interchanged rules into the existing rule base), where conflicts are resolved by rule prioritizations.

Another domain of research is the engineering and maintenance of large rule-based applications, where the rules are serialized and managed in a distributed manner, and are interchanged across domain boundaries. This calls for support of verification, validation and integrity testing (V&V&I), e.g. by test cases that are written in the same rule markup language and are stored and interchanged together with the rule program. A proposal for self-validating rule bases adapting a test-driven development approach from extreme programming in Software Engineering has been made for RuleML (Paschke et al, 2006) and for RIF (Paschke et al, 2005).

CONCLUSION

In this chapter several important requirements and design choices for a rule markup language have been described. It was shown how the current Web rule language proposals address these issues and what characteristics derive from those solutions. Commonalities as well as differences between the languages were presented and illustrated with concrete examples. Discussions of the advantages and disadvantages of the language design approaches reveal that all approaches legitimately coexist at this stage, as all have their strengths and weaknesses.

REFERENCES

W3C (2000). W3C Resource Description Framework. Candidate Recommendation. Retrieved from, http://www.w3.org/TR/2000/CR-rdf-schema-20000327/

W3C (2007). Semantic Web Layer Cake. Retrieved from, http://www.w3.org/2007/03/layerCake.png

Antoniou, G., Damasio, C. V., Grosof, B., Horrocks, I., Kifer, M., Maluszynski, J., et al. (2005). *Combining rules and ontologies — A survey.* Deliverables I3-D3, REWERSE. Retrieved from, http://rewerse.net/deliverables/m12/i3-d3.pdf

Ball, M. Boley, Hirtle, H.D., Mei, J., & Spencer, B. (2005). The OO jDREW reference implementation of RuleML. In *Proceedings of Rules and Rule Markup Languages for the Semantic Web* (RuleML-2005) (pp. 218-223). (LNCS 3791).

Boley, H. (2003). *Object-oriented RuleML: User-level roles, URI-grounded clauses and order-sorted terms* (. *LNCS*, *2876*, 1–16.

Boley, H. (2006). The RuleML family of Web rule languages. In *Proceedings of the Fourth Workshop on Principles and Practice of Semantic Web Reasoning*, Budva, Montenegro (LNCS 4187, pp. 1-15).

Boley, H. (2007). *Are your rules online? Four Web rule essentials*. RuleML (pp. 7-24).

Bry, F., & Marchiori, M. (2005). *Ten theses on logic languages for the Semantic Web*. PPSWR (pp. 42-49).

Codd, E. (1971). Alpha: A data base sublanguage founded on the relational calculus of the database relational model. In *Proceedings of ACM SIGFIDET Workshop on Data Description, Access and Control*, San Diego, CA.

Ensing, M. (2007). *The rule manager: A graphical business rules environment*. Demo at RuleML-2007. Retrieved from, http://2007.ruleml.org/docs/Acumen%20Business%20-%20RuleML.pdf

FIPA. (2000). *FIPA agent communication language*. Retrieved December 2001, from http://www.fipa.org/

Fuchs, N.E., Kaljurand, K., & Schneider, G. (2006). *Attempto controlled English meets the challenges of knowledge representation, reasoning, interoperability and user interfaces*. FLAIRS 2006.

Giurca, A., & Wagner, G. (2005). Towards an abstract syntax and direct-model theoretic semantics for RuleML. *RuleML, 2005*, 45–55.

Grau, B. C., Calvanese, D., De Giacomo, G., Horrocks, I., Lutz, C., Motik, B., et al. (2006). *OWL 1.1 Web ontology language tractable fragments.* W3C Member Submission. Retrieved from, http://www.w3.org/Submission/owl11-tractable/

Halpin, H. (2006, May). *Identity, reference, and meaning on the Web*. WWW2006 Workshop on Identity, Reference, and the Web. Retrieved from, http://www.ibiblio.org/hhalpin/irw2006

Heflin, J., Hendler, J., & Luke, S. (1999). *SHOE: A knowledge representation language for Internet applications* (Technical Report CS-TR-4078) (UMIACS TR-99-71).

Hirtle, D. (2006). *TRANSLATOR: A TRANSlator from LAnguage TO rules*. Canadian Symposium on Text Analysis (CaSTA), Fredericton, October 2006.

Kifer, M., de Bruijn, J., Boley, H., & Fensel, D. (2005). A realistic architecture for the Semantic Web. In A. Adi, S. Stoutenburg, & S. Tabet (Eds.), *RuleML* (LNCS 3791, pp. 17-29).

Levy, A. A., & Rousset, M.-C. (1998). CARIN: A representation language combining horn rules and description logics. *Artificial Intelligence*, *104*(1-2), 165–209. doi:10.1016/S0004-3702(98)00048-4

Lukichev, S., & Wagner, G. (2006). Visual rules modeling. In I. Virbitskaite & A. Voronkov (Eds.), *Proceedings of the 6th International Conference Perspectives of Systems Informatics* (LNCS 4378, pp. 467-673).

Mei, J., Lin, Z., & Boley, H. (2007). *ALCuP: An integration of description logic and general rules, Web reasoning and rule systems*. In First International Conference, RR 2007 (LNCS 4524, pp. 163-177).

Mei, J., Lin, Z., Boley, H., Li, J., & Bhavsar, V. C. (2007b). The DatalogDL combination of deduction rules and description logics. *Computational Intelligence*, *23*(3), 356–372. doi:10.1111/j.1467-8640.2007.00311.x

Munoz, S., Perez, J., & Gutierrez, C. (2007). Minimal deductive systems for RDF. In E. Franconi, M. Kifer, & W. May (Eds.), *ESWC* (LNCS 4519, pp. 53-67).

Paschke, A. (2005). *RBSLA: A declarative rule-based service level agreement language based on RuleML*. CIMCA/IAWTIC 2005 (pp. 308-314).

Paschke, A. (2006). *Verification, validation and integrity of distributed and interchanged rule based policies and contracts in the Semantic Web*. Int. Semantic Web and Policy Workshop (SWPW'06), Athens, Georgia.

Paschke, A. (2006b). *A typed hybrid description logic programming language with polymorphic order-sorted DL-typed unification for Semantic Web type systems*. OWL-2006 (OWLED'06), Athens, Georgia.

Paschke, A. (2007). *Rule-based service level agreements - Knowledge representation for automated e-contract, SLA and policy management*. IDEA Verlag GmbH, München.

Paschke, A., Boley, H., & Dietrich, J. (2005). *RIF use case. Rule interchange through test-driven verification and validation*. Retrieved from, http://www.w3.org/2005/rules/wg/wiki/Rule_Interchange_Through_Test-Driven_Verification_and_Validation, 2005.

Paschke, A., Boley, H., Kozlenkov, A., & Craig, B. L. (2007). Rule responder: RuleML-based agents for distributed collaboration on the pragmatic Web. *ICPW*, *2007*, 17–28. doi:10.1145/1324237.1324240

Paschke, A., Dietrich, J., Giurca, A., Wagner, G., & Lukichev, S. (2006). *On self-validating rule bases.* Int. Semantic Web Enabled Software Engineering Workshop (SWESE'06), Athens, Georgia.

Rosati, R. (2006). The limits and possibilities of combining description logics and datalog. In T. Eiter, E. Franconi, R. Hodgson, and S. Stephens (Eds.), *RuleML* (pp. 3-4). IEEE Computer Society.

Rosati, R. (2006a). *Integration ontologies and rules: Semantic and computational issues*. Reasoning Web.

Wagner, G., Antoniou, G., Tabet, S., & Boley, H. (2004). The abstract syntax of RuleML - towards a general Web rule language framework. *Web Intelligence*, *2004*, 628–631.

Wagner, G., Damásio, C. V., & Antoniou, G. (2005). Towards a general Web rule language. *Int. J. Web Eng. Technol*, *2*(2/3), 181–206. doi:10.1504/IJWET.2005.008483

KEY TERMS AND DEFINITIONS

Built-in Predicates or Functions: Special restricted predicate or function symbols in the rule language for concrete domains, e.g., built-ins for strings, numerics, Boolean values, date, time, intervals, lists, etc.

Deontic Rules: Describe rights and obligations, e.g., of institutions and agents in the context of evolving states (situations triggered by events/actions) and state transitions, where integrity rules (see above) are a special case ('introspectively') affecting the rule set itself.

Derivation Rules: Infer conclusions from conditions (as in Datalog and Horn logic), where facts are a special case with constantly true conditions.

Facts: Various kinds of information such as asserted atoms (formulas), individual-class memberships (of ontology classes), (object-oriented) instances, stored data (e.g., relational, XML), states and event occurrences which might be qualified, e.g., by priorities, temporally, etc.

Integrity Rules (or integrity constraints): Assertions which express conditions (or queries) that must always be satisfied.

Procedural Attachment: A function or predicate whose implementation is given by an external procedure.

Reaction Rules are (Behavioral / Action) Rules: (Re)act on occurred events (external events or changed conditions) by executing actions, where production rules are a special case with events restricted to changed conditions.

Rule Interchange Format: A common interchange format, such as e.g. W3C RIF or RuleML, for different rule types and rule families.

Rule Markup Language: A concrete markup-based rule syntax using e.g. XML for the Web.

Semantic Web Rule Language: A rule language specifically tailored for the semantic web with a human-friendly syntax, e.g., a scripting syntax (as opposed to a Rule Markup Language for the Semantic Web).

Transformation Rules: Specify term rewriting, which can be considered as derivation rules of logics with (oriented) equality.

This work was previously published in Handbook of Research on Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches, edited by A. Giurca, D. Gasevic, and K. Taveter, pp. 1-24, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 2.20
Semantic Web Rule Languages for Geospatial Ontologies

Philip D. Smart
Cardiff University, UK & University of Glamorgan, UK

Alia I. Abdelmoty
Cardiff University, UK & University of Glamorgan, UK

Baher A. El-Geresy
Cardiff University, UK & University of Glamorgan, UK

Christopher B. Jones
Cardiff University, UK & University of Glamorgan, UK

ABSTRACT

Geospatial ontologies have a key role to play in the development of the geospatial-Semantic Web, with regard to facilitating the search for geographical information and resources. They normally hold large volumes of geographic information and undergo a continuous process of revision and update. Limitations of the OWL ontology representation language for supporting geospatial domains are discussed and an integrated rule and ontology language is recognized as needed to support the representation and reasoning requirements in this domain. A survey of the current approaches to integrating ontologies and rules is presented and a new framework is proposed that is based on and extends Description Logic Programs. A hybrid representational approach is adopted where the logical component of the framework is used to represent geographical concepts and spatial rules and an external computational geometry processor is used for storing and manipulating the associated geometric data. A sample application is used to demonstrate the proposed language and engine and how they address the identified challenges.

DOI: 10.4018/978-1-60566-402-6.ch007

INTRODUCTION AND BACKGROUND

The Internet is the single largest information resource in the world that is however still not being used to its full potential. To fully unlock the potential of such a large knowledge resource and to enable its effective utilisation by both human and machine agents, information on the Web needs to be machine-understandable using semantic as opposed to syntactic (e.g. HTML) markup languages and tools. At the heart of this vision are ontologies which, in the context of the web, are logical theories that act to constrain and derive information (Guarino,1995). They provide the necessary semantics and machine understanding to the sheer volumes of information contained on the Web.

A significant proportion of information resources on the web are geographically referenced. Nearly 17% of all web queries contain place names (Sanderson & Kohler, 2004) and the web, powered by the simplicity of recent applications such as Google Maps, is increasingly being seen as a medium for the storage and exchange of geographic data in the form of maps. A geographic or geospatial ontology is a model of terminology and structure of geographic space as well as records of entities in this space (Egenhofer, 2002). This chapter considers the development and management of geospatial ontologies on the Semantic web. By analyzing the nature and complexity of the geographical concepts and data to be handled by these ontologies, we evaluate the suitability of the current semantic web tools and suggest an appropriate platform to represent and develop these ontologies.

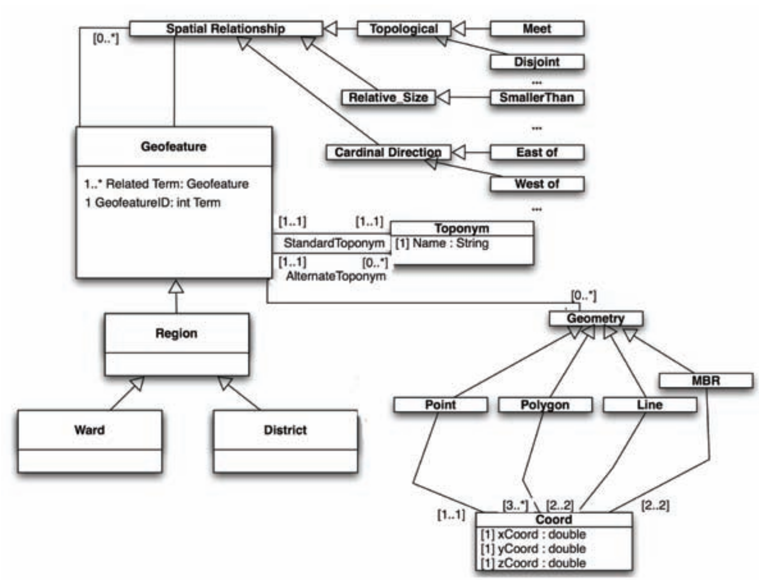
In particular, geographical concepts are complex, normally associated with geometric representations of their boundaries and location and exhibit implicit spatial relationships that need to be computed and derived. Qualitative spatial reasoning as well as computational geometry procedures are both established complementary techniques for the representation and manipulation in this domain. In addition, maintaining the spatial integrity of large geospatial ontology bases is crucial for their realization. Ontology representation languages such as OWL are limited in their ability to handle the challenges in this domain. In this chapter, a survey of current approaches to integrating rules and ontologies is presented. Two approaches are identified, namely a hybrid approach where both systems of ontologies and rules are kept distinct and communicate only through an interface, and a homogeneous approach where one system is mapped to and becomes accessible from the other.

In the second section, we first discuss the representational and manipulation challenges facing ontology management systems that aim to support geospatial domains. OWL as an ontology representation language is evaluated against those challenges and the need for a integrated rule layer is highlighted. In the third section, current approaches to integrating rules and ontologies (logic programming and Description Logic) are identified and classified. Based on a comparative evaluation of both approaches, a homogenous approach to integration, namely, Description Logic Programs is chosen as a suitable platform for the development of geospatial ontology management systems. In the fourth section, the potential and further extensions of this new approach are described. In the fifth section, the implementation of the approach is briefly sketched and demonstrated using a sample geospatial ontology described in the chapter, followed by conclusions and future outlook in the final sections.

MANAGING GEOSPATIAL ONTOLOGIES

In this section we consider a typical geospatial ontology model, as shown in Figure 1. The model is based on OGC guidelines for simple geographic features, see (OGC Technical Committee, 1999; Vretanos, 2005), and other models commonly used in existing geospatial ontology development (Fu et al., 2005, Smith & Frew, 1995). The terminology of the geo-ontology is relatively plain with regards to the number and type of constructs used. This reflects typical geographic ontology developments which, beyond the complex representation of geometry, are relatively sparse (parsimonious (Jones

Figure 1. An example geospatial ontology



et al., 2001)) and fit to purpose. In this section, issues related to the representation and management of such geospatial ontologies are discussed and OWL's ability to handle them is evaluated.

In Figure 1, a geofeature is a representation of any geographic phenomena that exist in space, e.g. a forest, a building or a road. As such, its location and boundary can be specified using a geometric entity of point, line or polygon. Also, as it is located in space, the relationships it exhibits with other geographical features are of interest, e.g. it may be inside (topological), north of (directional) or near to (proximity) another feature. Some of these relationships may be stored explicitly or need to be computed from the features geometric entities.

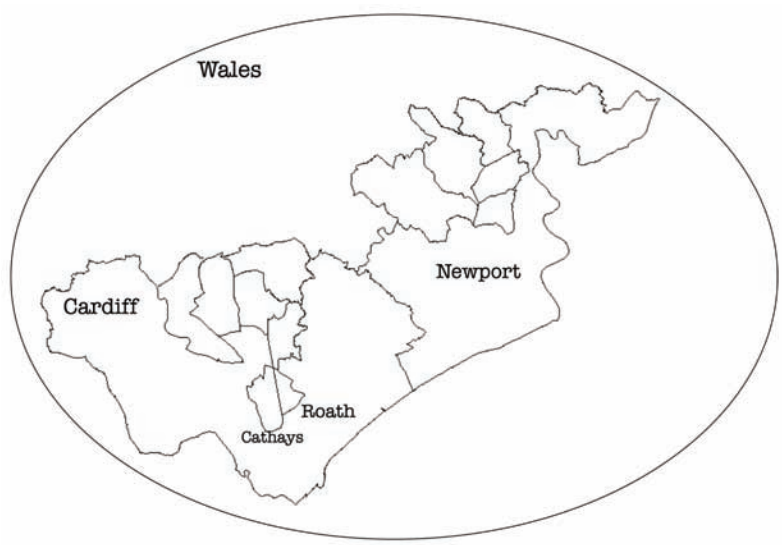
Representational Requirements

Consider the map scene in Figure 2. The map shows sample administration regions in Wales. Two Unitary Authorities are shown; Cardiff and Newport along with some of their contained Wards.

The ontology in Figure 1 can be used to represent this map of Places. In particular, it can be represented using OWL as shown below. Examples from the axioms in the TBox (terminology or model) and the ABox (asserted knowledge) are shown. The logic-based syntax used here is that of OWL-DL.

TBox
$Geofeature \subseteq\ =1.GeofeatureID \cap \geq 1.RelatedTerm \cap \forall RelatedTerm.Geofeature$
…
Region $\subseteq$ *Geofeature*
Country $\subseteq$ *Region*
Ward $\subseteq$ *Region*
UnitaryAuthority $\subseteq$ *Region*
ABox
Cardiff: UnitaryAuthority
Newport: UnitaryAuthority
Roath: Ward
Wales: Country
Cathays: Ward
Cardiff Meets Newport
Roath Inside Cardiff
Cardiff Inside Wales
Roath Meets Cathays

Figure 2. A sample geographic scene showing administrative regions in Wales



The geometry of geographic regions is normally complex and can consist of a large number of points. For example, the Unitary Authority of Cardiff in Figure 2 is represented by over 1,000 points, presenting a large storage and management overhead on the ontology language. An experiment with a small geospatial ontology of geographic regions was constructed in OWL, using the Jena2 library. The base OWL ontology, without the associated geometry, contained 1,000 individual places with two data attributes (ID and name). The ontology occupied approximately 2.2mb of persistent storage space and 16mb of system memory to reason with. Adding the associated geometry for each place, represented as general classes and data properties, the ontology grew to around 100mb on persistent storage and the memory overhead increased to over 1gb. Even then, spatial queries e.g. all Wards in Wales, are not supported within OWL as currently its associated reasoning engines e.g. Racer (Haarslev et al., 2001) or Fact (Horrocks, 1998), do not contain the necessary algorithms to perform polygon in polygon geometry operations etc.

To support the geospatial domain, an ontology management system needs to consider the following requirements.

- Support the representation of basic spatial data models, spatial data types and relationships.
- Provide a scalable capacity for handling and searching over large geometric data stores.

Manipulation Requirements

Manipulating geospatial ontologies involves the search, computation and retrieval of spatial properties and relationships. Two paradigms are possible: quantitative, using computational geometric procedures for structuring and search along with qualitative, using qualitative spatial reasoning techniques. Indeed, both paradigms are complementary and can be used together.

In this section, both paradigms are explored to understand the issues they pose to geospatial ontology management systems.

Table 1. Part of a spatial composition table for topological relations between simple regions. 1 denotes the universal relation.

	DISJOINT	**MEET**	**OVERLAP**	**INSIDE**	**CONTAINS**
DISJOINT	1	DISJOINT, MEET, OVERLAP, COVEREDBY, INSIDE	DISJOINT, MEET, OVERLAP, INSIDE, COVEREDBY	DISJOINT, MEET, OVERLAP COVEREDBY, INSIDE	DISJOINT
MEET	DISJOINT, MEET, OVERLAP COVERS, CONTAINS	DISJOINT, MEET, OVERLAP COVEREDBY,COVERS, EQUAL	DISJOINT, MEET, OVERLAP, INSIDE, COVEREDBY	OVERLAP, COVEREDBY, INSIDE	DISJOINT
OVERLAP	DISJOINT, MEET, OVERLAP CONTAINS COVERS	DISJOINT, MEET, OVERLAP, COVERS, CONTAINS	1	OVERLAP, COVEREDBY, INSIDE	DISJOINT, MEET, OVERLAP COVERS, CONTAINS
INSIDE	DISJOINT	DISJOINT	DISJOINT, MEET, OVERLAP INSIDE, COVEREDBY	INSIDE	1
CONTAINS	DISJOINT,MEET, OVERLAP CONTAINS COVERS	OVERLAP,COVERS CONTAINS	OVERLAP, COVERS CONTAINS	OVERLAP,COVERS COVEREDBY, INSIDE, CONTAINS EQUAL	CONTAINS

Geometric Manipulation

As noted in the previous section, geospatial ontologies will normally be associated with large geometric ontology bases representing the ground location associated with geographic phenomena. Simple manipulation of these phenomena will involve the computation of their spatial properties, such as length or area and relationships such as near or inside. Traditional computational geometry algorithms need to be implemented to compute these properties. Also, spatial databases and information systems normally employ different forms of spatial data structures and indexing techniques to facilitate searching over large geometric stores.

Hence, a geospatial ontology management system needs to consider the following requirement.

- Support basic geometric computational and spatial search functions to manipulate the geometric data stores.

Qualitative Spatial Representation and Reasoning

Over the past two decades much work has been conducted on the development of qualitative spatial approaches to represent and reason over space and spatial relations (Frank, 1992; Freska, 1992; Gahegan, 1995; Egenhofer et al., 1999; El-Geresy, 2004). Qualitative spatial manipulation is import and complements quantitative geometric processing in space, especially so when precise geometric information is missing or simply not needed for the context of operation. Qualitative approaches are based on the exploitation of the nature of the structure of space and the qualities of the spatial relationships themselves for deriving implicit information. Results of these approaches are documented in what is known as composition table of spatial relations such as the one shown in Table 1.

Entries in the table are possible relationships between two regions (A and C) resulting from the composition of relationships with another region B, i.e. $R_1(A,B) \otimes R_2(B,C) \rightarrow R_3(A,C)$.

As an example, consider the spatial relations defined in the ontology in the previous section. A new relationship can be derived as follows, where $\otimes$ denotes the composition of two relations:

Inside $\otimes$ ***Inside*** $\rightarrow$ ***Inside***

*(Roath**Inside**Cardiff)* $\otimes$ *(Cardiff**Inside**Wales)* $\rightarrow$ *(Roath**Inside**Wales)*

Entries in spatial composition tables can be seen as a set of first order compositional inferences, and can be represented as a set of deduction or inference rules of the form:

$\forall\ x,y,z$: $R_1(x,y) \wedge R_2(y,z) \rightarrow R_3(x,z)$

where R_1, R_2 and R_3 are spatial relations, for example,

$\forall\ x,y,z$ Inside(x,y) $\wedge$ Meet(y,z) $\rightarrow$ Disjoint(x,z)

where x,y and z are region variables, substituted for geofeature instances in the geospatial ontology. The rule entails that the region bound to the variable x is disjoint from the region bound to the variable z, if x is inside another region y, and y meets (touches) z.

Hence, another requirement for the geospatial ontology management system is as follows.

- Support the representation of spatial composition rules for qualitative spatial reasoning.

Maintaining the Consistency of Geospatial Ontologies

'Consistency describes the absence of any logical contradictions within a model of reality' (Nectaria & Egenhofer, 1996). Errors in the description of the location and shape of geographical entities are common, especially on the web, when data provided and manipulated by users may not be complete or accurate. Such errors can propagate to inconsistencies in the spatial relationships and consequently to wrong information being stored in the ontology bases. Erroneous updates to the data may go undetected unless appropriate spatial integrity rules are declared and applied.

Cockcroft (1997) catagorised spatial integrity rules as 1) topological, maintaining the accuracy of topological information (which applies to all spatial relations) 2) Semantic, concerning the meaning of geographical features and how they should legally be allowed to interact 3) user defined, analogues to user defined business rules. Topological constraints can be further subdivided to structural errors, geometric errors, and topo-semantic constraints (Servigne et al., 2000).

Consider again our example ontology and a new fact to be inserted as follows.

Roath Disjoint Wales

Although this fact is valid from the point of view of the ontological model, i.e. it asserts the existence of a spatial relationship between two regions, its spatial consequence implies an inconsistency as the explicit relationships already stored implies that Roath must in fact be inside Wales.

*(Roath**Inside**Cardiff)* $\otimes$ *(Cardiff**Inside**Wales)* $\rightarrow$ *(Roath**Inside**Wales)*

To detect this inconsistency, we need not only to encode the appropriate qualitative spatial rule but also the following integrity constraint

$\forall x,y,z$ Inside(x,y) $\wedge$ Inside(y,z) $\wedge$ Disjoint(x,z) $\rightarrow$ error

Different types of spatial integrity constraints can be encoded to maintain the consistent spatial structure and properties of the geospatial ontology bases. For example the following constraint indicates that if an object is inside another it must also be smaller in size.

$\forall x,y$ Inside(x,y) $\wedge$ SmallerThan(y,x) $\rightarrow$ error

A final consideration for the geospatial ontology management system can therefore be as follows.

- Support the expression and implementation of spatial integrity constraints over geospatial ontology bases.

Summary

Description logics (DL) are a powerful representational tool for describing real world concepts, their attributes and relationships. Key reasoning mechanisms of any DL are checking concept satisfiability and inferring subsumption hierarchies. On the terminological level a DL reasoner will infer concept hierarchies based on concept subsumption. On the level of asserted knowledge (instance level) each individual's type is inferred if not already explicit.

The description logic underpinning OWL-DL (*SHOIN*(D)) has purposefully been restricted to preserve decidability and in large tractability of the language. Such restrictions lead to representation and reasoning limitations. In addition to these, several limitations can be recognized with OWL as a platform to support the management of geospatial ontologies (Abdelmoty et al., 2005). These limitations, some as highlighted in previous sections, are now summarized below.

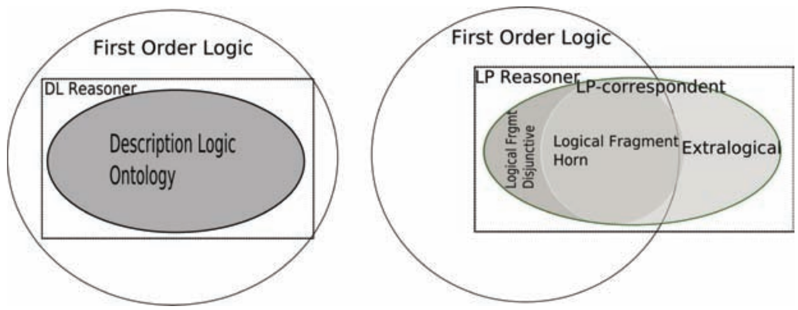
- OWL's first order, open world semantics in combination with the non unique name assumption is not suitable for constraint checking (Bruijn et al., 2005). Extensions to OWL have been proposed to overcome this limitation, for example by translating subsets of OWL to a logic program that assumes both unique name and closed world assumptions (Bruijn et al., 2005; Grosof et al., 2003), as well as by locally closing certain domain concepts using autoepistemic constructs.
- "Triangular knowledge" can not be represented directly in OWL (Horrocks, 2005). In particular, inference patterns of the form, $\forall$ x,y,c: $rel_1(x,y) \wedge rel_2(y,c) \rightarrow rel_3(x,c)$ can not be represented. This is the typical form of a spatial compositional inference rule.
- The description logic underpinning OWL-DL does not support spatial data types. Representation of geometric objects using generic class and property constructs is not ideal and will have potentially high implications on storage overheads (Haarslev et al. 1998).
- OWL does not support geometric computation, analysis or spatial indexing. Simple computational geometry calculations such as area and distance calculations are not possible to represent and consequently more complex spatial search queries are also not possible.
- Tableaux based reasoners (as used in most DL reasoners) are poor for query answering over individuals (Bruijn et al., 2005). Instance bases of geospatial ontologies are likely to be very large. Logic programming reasoning engines are more suitable platforms for reasoning in this case.

Realising the limitation of OWL, a rule layer has recently been proposed in the semantic web stack. Adding a rule layer to the existing ontology layer (typically a DL variant) will help to overcome some of the representational and reasoning limitations of OWL. In the next section a survey is given of existing approaches to the integration of logic programs and description logics, or less formally rules and ontologies.

INTEGRATION OF RULES AND ONTOLOGIES

Work on adding a rule layer to the semantic web technology stack was initiated by the W3C at the turn of the century. A fundamental challenge in facilitating the addition of this rule layer is finding means of integrating rules (Logic Programs - LPs) and ontologies (a descriptions logic- DL),

Figure 3. Different paradigms; a) ontologies in DLs and b) Logic programming systems



and in doing so handling the following semantic differences between logic programs and the DL subset of classical first order logic.

- First order languages have open world semantics, while LPs adhere to closed world semantics.
- DLs do not assume the unique name assumption, whereas all individuals in a logic program are assumed unique.

In addition, an integrated framework for rules and ontologies will face the following issues.

- Computational complexity of the resulting system. The decidability and tractability of the resulting reasoning system. The realisation and practical utilisation of the combined frameworks need to provide pragmatic reasoning procedures (with at most a polynomial time complexity).
- The modularity of the reasoning process. The choice is between combining both systems into a single logical language with a uniform reasoner or retaining different reasoners.

Figure 3 illustrates the differences between ontologies, as DL fragments of first order logic and rule systems as logic programming fragments of the same logic. The integration of the DL structural component with the LP relational component is the subject of much work and debate. Current approaches can broadly be classified into two categories based on the degree of overlap between the two models in the resulting system; the *hybrid approach* and the *homogeneous approach* as described below.

Rule + Ontology: A Hybrid Approach

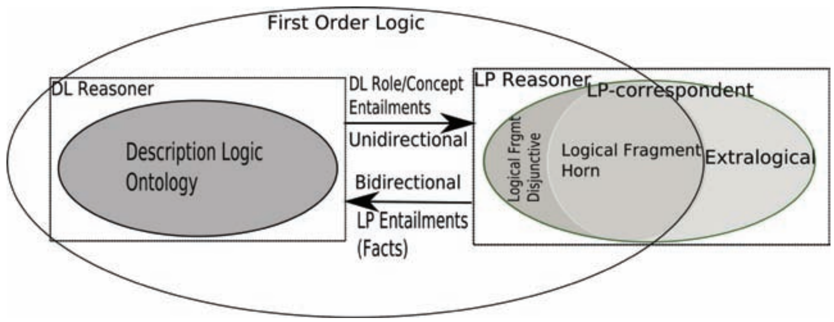
A hybrid approach is essentially a modular approach to the integration of rules and ontologies. It is sometimes referred to as loose integration or the integration of rules and ontologies through strict semantic separation (Eiter et al., 2006). Reasoning distinction between the ontology (DL) component and the relational (rule) component is maintained (Rosati, 2005). The ontology component is a description logic variant i.e. ALC and upwards, and the rule component is typically some identified flavor of Datalog (for example Datalog or Datalog$^{\vee}$).

A complete hybrid knowledge base K is represented by a pair K = <Σ,Π>, where Σ represents the ontological (DL) component and Π represents the relational (rule) component. Π contains rule and ontology predicates maintaining a strict separation between both. Typically a rule r in the relational component Π has the form:

$H \leftarrow B_1 \wedge \ldots \wedge B_n\colon O_1 \wedge O_m \; m \geq 1, \; n \geq 1$

where H and B are both rule predicates (head and body predicates respectively) and O

Figure 4. Hyrbid Rule + Ontology Integration



represents an ontology predicate. The ontology or structural predicates act as constraints on the interpretation of the relational component. Interaction between the rule and ontology reasoners take place through a safe interface (Eiter et al., 2006). The flow of information through the interface is either unidirectional or bidirectional.

In the unidirectional approach, reasoning is performed over the ontology using an ontological reasoning engine (DL reasoner). Entailments from the DL reasoner ($\Sigma \models \omega$) are fed, as a starting point, into the rule reasoner. Rules are interpreted such that they must satisfy the ontology predicates p in ω. Early unidirectional approaches combined unexpressive structural variants with unexpressive relational variants. For example AL-Log (Donini et al., 1998) uses the foundational DL namely *ALC* with Horn Datalog while introducing the, now common, rule safety condition to maintain decidability. That is, rule safety constraints the use of each variable that appears in the head of the rule to also appear in the body of the rule. Moreover, only concepts are allowed as constraints in the relational component. CARIN (Levy & Rousset, 1995) overcame this limitation with a more expressive DL (*ALCNR*), and allowed both concept and role constraints whilst maintaining decidability using a form of role safety.

To further enhance reasoning, a complete synergy of structural and relational reasoning can be catered for by introducing a bidirectional flow of information. Iterative reasoning is then performed on both components until no more inferences can be drawn. This approach was adopted by Eiter at al. (2004), namely Description Logic Programs, and DL+log (Rosati, 2006).

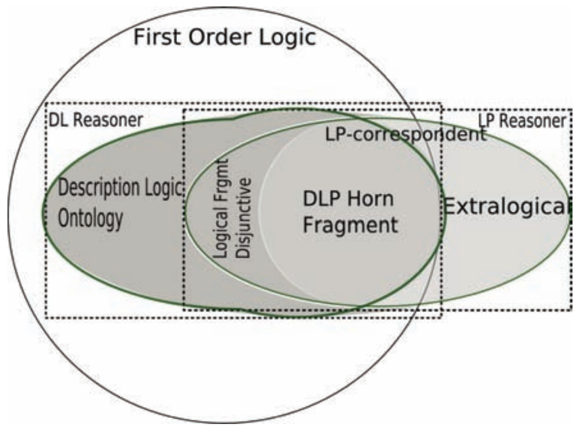
Homogenous Rule + Ontology Integration

Homogenous approaches present a complete or tight integration of both the ontology Σ and rule Π components resulting in a single logical language *L*. No syntactic or semantic distinctions are made in *L* between the ontology and rule predicates and both can be interpreted under the same reasoning umbrella.

All works based in this approach employ a mapping (typically a recursive mapping) from one language to the other. Approaches that involve mapping the rule language to the ontology language exist. However, more common are approaches mapping from the ontology language to the rule language, thus opening the possibility of using existing rule engines for reasoning tasks (query answering etc). The mapping process involves either completely combining both languages (expressive union), or embedding one language into the other (intersection).

Expressive union of the structural and relational component is depicted in Figure 5 as the union of the entire LP and DL fragments within FOL. The union brings about substantial computational

Figure 5. Homogenous Rule + Ontology Integration



complexities that often leads to undecidability, even if the languages of each are simplistic, see (Schmidt-Schauß, 1989) for the proof. W3C's candidate rule language SWRL (Patel-Schneider et al., 2004) is an example of unifying OWL-DL with a Horn based rule language which does lead to undecidability.

Intersecting the structural and relational components into their common fragment can help to retain decidability and tractability if for example, their common fragment corresponds to the horn subset of FOL. Description Logic Programs (Grosof et al., 2003) represent a complete ontology paradigm that is formed by the intersection of the description logic underpinning OWL-DL with Horn Datalog. The result is a sound, highly tractable and practical paradigm from which further extensions can be be layered (Hitzler, 2005). Over the past few years a number of extensions have been considered to DLP. For example, by considering disjunctive logic programs a larger fragment of DL can be mapped into the combined language *L* (Motik & Volz, 2003), and in (Krötzsch et al., 2008) which extended the reasoning potential of DLPs, while not adversely effecting the decidability and tractability of the language.

DLPs are also at the heart of the W3C's alternative (to SWRL) candidate rule language WRL (Angele et al., 2005). WRL-core is the base representation paradigm of WRL and represents an ontology language in its own right.

Comparison of the Two Approaches

Augmenting ontologies with rules is of benefit in general and offer potential solutions to the representational limitations of ontology languages such as OWL, or indeed to most description logics.

Of the hybrid approaches, early techniques either used unexpressive DL variants (e.g. AL-Log *ALC)*, or only permitted class constraints in the relational component. Unidirectional approaches produce only a subset of all inferences possible from the combination of both components. On the other hand, bidirectional approaches suffer from higher, often intractable, and therefore unacceptable worst case complexities (EXPTIME or NEXPTIME see (Eiter et al., 2004)). The use of two separate reasoning engines to obtain a unified output could be an inconvenient obstacle to reasoning and may add an additional cost to run-time performance.

Of the homogenous approaches, SWRL is undecidable and as such is not mature enough for full scale implementation. WRL-Flight is promising and supports integrity constraints thanks to both a closed word and unique name assumption. However, its perfect model semantics is not as compatible with most existing ordinary logic programming engines. Disjunctive logic programs exhibit high computational complexities (typically higher than polynomial time i.e. NEXPTIME or $NEXPTIME^{NP}$ (Eiter et al., 1997)), due to the possibility of multiple minimal models (each minimal model increases the size of the search space exponentially).

The most tractable of all approaches is, unsurprisingly, the least expressive logic, namely, Description Logic Programs (DLPs). A DLP can be trivially mapped to existing logic programming or production system engines e.g. XSB (Sagona et al., 1993) or Rete (Forgy, 1982) respectively. Logic programs or productions systems do not assume classical first order semantics. Making intuitive closed world and unique name assumptions make them more suitable models for integrity checking tasks. Finally, logic programs will scale well to reasoning over individuals (Krotzsch et al., 2006); a very desirable property, in particular for the geospatial domains.

Approaches to Integrating Spatial Logics and Ontologies

Current approaches to utilizing spatial logic and reasoning in ontologies are based on the homogeneous approach to integration and propose extending existing description logics with spatial concrete domains and qualitative spatial reasoning algorithms. The method provides a means to employ spatial reasoning for deduction and satisfiability checking of spatial information asserted into the description logic.

Haarslev et al. proposed an extension to ALCRRP(D) DLs to include a concrete spatial domain and hence support spatio-terminological reasoning (at the concept level) (Haarslev et al., 1998). However, the proposal was limited to support only one specific spatial data type; a polygon.

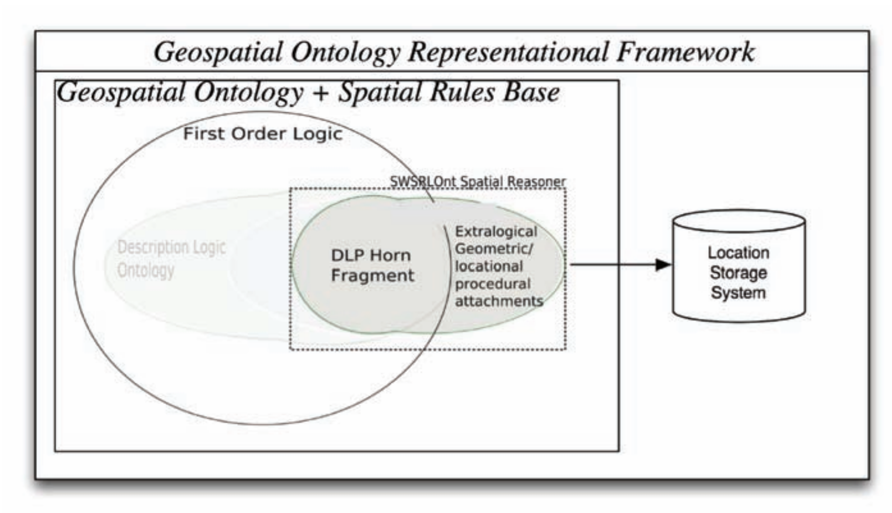
Wessels (2001) proposed the $ALCI_{rcc}$ DL that includes role axioms derived from the spatial composition tables, based on the Region Connection Calculus (RCC) (Cohn et al., 1997), a theoretical variant of the 9-intersection model shown in the previous section, again for topological reasoning. The RCC spatial role axioms can be used to check the topological satisfiability of spatial individuals. However, as noted in their work, the proposal is only decidable if RCC-5 or a more general family of the spatial logic is used – for example by replacing the distinction between within and coveredby with the more general part-of relation.

To be of practical use, Wessels (2003) also argues for layerd hybrid deductive geographic information system that employs the DL $ALCI_{rcc}$, and recognizes the limitations of adding geometry to a description logic, previously mentioned.

The full potential of using an integrated logical and geometric framework to support a general spatial ontology is yet to be realized. The potential of such a framework to serve geospatial domains has been recognized in some recent proposals (Chen et al., 2005; O'Dea et al., 2005; Smart et al.). Exiting approaches to combine spatial logics with description logics to form new geospatial ontology language, highlight some important issues. Both proposals above are limited in their application both from a representational and reasoning perspective. In particular, the following issues still need to be addressed.

- Existing DLs need to include complex role composition constructs in order to effectively store spatial compositional inference rules such as those described in table 1, as well as role compositions that allow head disjunctions. A separate rule representation layer is needed.
- A spatial ontology will need to support a general set of spatial data types, such as regions, points, and lines. Restricting geospatial ontologies to spatial logics that work over simple polygonal shapes limits their general applicability.

Figure 6. Proposed DLP approach to representing geospatial ontologies and rules



- Geometric processing of spatial data is better suited to dedicated external geometric reasoning engines. Hence, 'outsourcing' the geometric component of the ontologies to external components outside the DLs and logic programming seems a reasonable choice.

A SEMANTIC WEB RULE AND ONTOLOGY LANGUAGE FOR GEOSPATIAL DOMAINS

From the above survey of approaches to integrating rules and ontologies and the analysis of the requirements of the geospatial domains, we propose the use of Description Logic Programs DLP (a homogeneous approach to integration of rules and ontologies) as a basis for representing and reasoning in geospatial domains. To further support the particular requirements of these domains, an integrated framework combining both the logical and the computational approaches to geospatial data processing is also proposed. Figure 6 depicts the new proposed language framework for representing geospatial onotlogies. In this section, we examine how this framework can support the requirements identified in the second section and also outline any necessary further extensions.

DLP for Representing Geospatial Ontologies

The geospatial ontology in section 2 can be captured, without loss in a DLP. It is of interest to note that a large percentage of currently developed ontologies are also within the representational abilities of the DLP fragment (Volz, 2004). A transformation function, named DLP-Fusion (Grosof et al., 2003), is needed to translate OWL-DL to DLP. The representation of the geometric component of geo-features is decoupled from the ontology component and delegated to an external specialized system as described below.

DLP-Fusion, OWL-DL to DLP

DLP-Fusion is a syntactical, semantic preserving, bidirectional mapping between OWL-DL and the Horn fragment of FOL, resulting in a new ontology paradigm namely a Description Horn Logic

Table 2. Example DLP-Fusion mapping

DL Axiom	First Order Rule Syntax
Region $\subseteq$ Geofeature	Region(x) $\rightarrow$ Geofeature(x)
Ward $\subseteq$ Region	Ward(x) $\rightarrow$ Region(x)
Topological $\subseteq$ Spatial_Relationship	Topological(x,y) $\rightarrow$ Spatial_Relationship(x,y)
Meet $\subseteq$ Topological	Meet(x,y) $\rightarrow$ Topological(x,y)
Roath: Ward	Ward(Roath)
Wales: Country	Country(Wales)
Roath Inside Cardiff	Inside(Roath,Cardiff)

(DHL) ontology.

The mapping function, denoted *T*, takes DL axioms of the form ($C \subseteq D$), ($S \equiv B$), ($T \subseteq \forall P.D$), ($T \subseteq \forall P^{-}.D$), (a: D), (<a,b>: P), ($P \subseteq Q$), ($P \equiv Q$), ($P \equiv Q^{-}$)), ($P^{+} \subseteq P$) and converts them into a rule of the form $A \rightarrow B$. The geospatial ontology in figure 3 can be directly mapped using this transformation. Examples are shown in Table 2.

Where x and y are spatial variables from the domain of individuals in the geospatial ontology.

DLP for Qualitative Spatial Reasoning

Unlike OWL, the Horn fragment of first order logic is sufficient to represent triangular knowledge or role composition of the form (where R_1, R_2 and R_3 are spatial relations, and x,y and z are spatial variables):

$$\forall x,y,z\ R_1(x,y) \wedge R_2(y,z) \rightarrow R_3(x,z)$$

For example from the composition Table 1, we can now represent the compositional inference:

$$\forall x,y,z\ \text{Disjoint}(x,y) \wedge \text{Contains}(y,z) \rightarrow \text{Disjoint}(x,y)$$

Of note, OWL-DL can represent transitive roles, for example: $R_1(x,y) \wedge R_1(y,z) \rightarrow R_1(x,z)$, but not those involving three different relations as shown in the example above.

As DLPs are the logic programming equivalent of a Horn rule, they too can capture such spatial compositional inferences. However, Horn rules are limited to compositional inferences with a definite head (one head predicate). To employ more general qualitative spatial reasoning calculi disjunctive rules are required; rules that allow head disjunctive of the form:

$$R_1(x,y) \wedge R_2(y,z) \rightarrow R_3(x,z) \vee \ldots \vee R_n(x,z)$$

for example:

$$\text{Contains}(x,y) \wedge \text{Meet}(y,z) \rightarrow \text{Overlap}(x,z) \vee \text{Contains}(x,z)$$

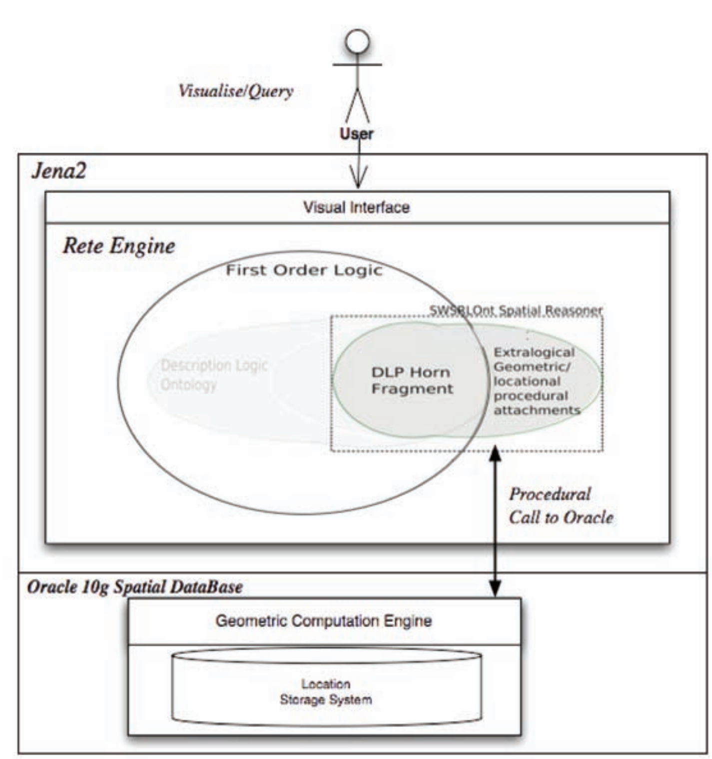
To work around this problem, spatial calculi have been proposed that can be directly mapped to Horn rules (Schockaert& Cock, 2007). Such spatial calculi are currently being tested with the proposed DLP approach, but the results are not included in this chapter.

In addition to representing qualitative rules, a DLP can be extended to include procedural attachments that implements, through calls to external geometric processing systems, spatial operators for the computation of spatial properties and relationships. A reasoning synergy is then possible to compute required properties which could not be derived automatically using the spatial logic.

DLP for Integrity Maintenance

Classical Horn Logic assumes a first order semantics, including the open world and non unique name assumptions. However, the logic programming equivalent of Horn logic, that used by DLPs, assumes a more intuitive closed world and unique name assumption and is consequently

Figure 7. DLP based geo-ontology software framework



a suitable language for expressing and implementing integrity constraints.

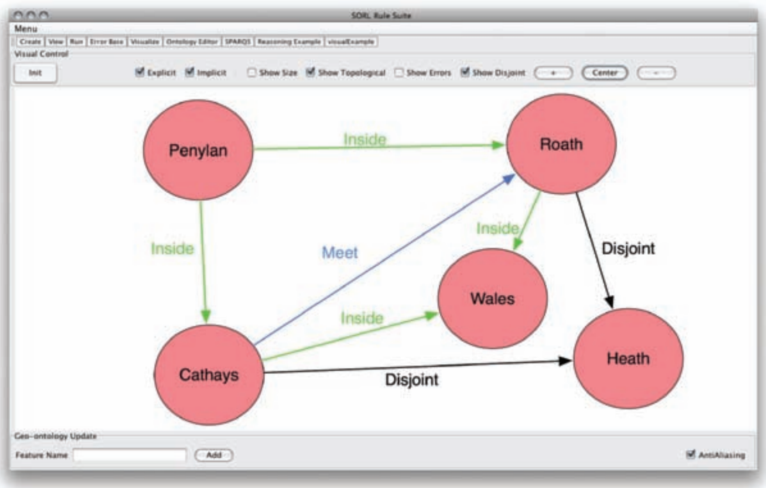
Integrity rules are headless rules, where satisfaction of all body predicates leads to a conflicting knowledge base. In addition to this general form of integrity rules, representation of 'default' rules and rule exceptions is also desirable. Consider for example, a rule that states that rivers and roads do not intersect. It may be desirable to express exceptions to this rule in the case where the entities in fact do intersect in a "ford". Enumerating rule exceptions is therefore a desirable quality. Extension to DLP to support a form of default logic, e.g. Courteous Logic (Grosof et al., 1997) is needed.

A backward chaining mode of inference is used in the implementation of logic programs, where querying the system results in exhaustive rule evaluation. On the other hand, integrity rules must continually monitor the ontology to find inconsistencies as and when they arise and as such need to work in a forward chaining mode. However, forward and backward reasoning modes can be made to work together, an interleaved mode (Eisenstadt & Brayshaw, 1990). Then, predicates in the forward system can be determined on the fly from the backward system, and no longer have to be contained explicitly in the geo-ontology as facts. This is advantageous in reducing storage and memory overheads, but has an adverse effect on reasoning speed. That is, not all facts need to be stored in the geo-ontology, but querying the backward system requires evaluation of, in the worst case, the entire rule set for each query.

PROTOTYPE APPLICATION

A system has been developed that implements the DLP proposed framework above within the Jena2

Figure 8. Geo-ontology visualisation tool



Semantic Web toolkit (Carroll et al., 2003). The rule engine is based on the Rete pattern matching production system (Forgy, 1982). A complete spatial rule base of topological spatial compositions has been developed. The geometric processing component is developed using the Oracle 10g Spatial database management system. Figure 7 shows a concrete instantiation of the framework in Figure 6.

Evaluation of the system is currently ongoing on synthetic as well as real data sets. A small example based on the ontology in section 2 is demonstrated below. The following are examples of both an integrity and deduction rule in the system.

Topological deduction rule:

[<label>ContainsEqualContains</label><ruleGroup>Topological</ruleGroup>: Equal(?x ?c) AND Contains(?c ?y) --> Contains(?x ?y)]

Topological integrity rule:

[[<label>InsideDisjointIntegrity</label><ruleGroup>Topological</ruleGroup>: Region(?x) AND Region(?y) AND Region(?z) AND inside(?x ?y) AND Disjoint(?y ?z) AND Inside(?x ?z) --> error(?x ?z)

Figure 8 shows a screenshot of the ontology depicted as a graph of regions (as red circles) and relations (different colored edges).

A new region (Penylan) as well as two new relationships are introduced to the ontology as found using the web mining technique in (Schockaert et al., 2008), these are:

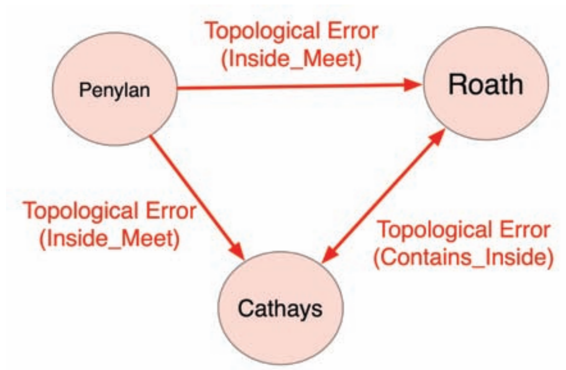
Penylan: Ward

Penylan inside Roath

Penylan inside Cathays

The reasoning engine detects the inconsistencies as a result of this update as shown in Figure 9, where edges resulting in the error are highlighted (in red).

Figure 9. Topological inconsistencies



The three relationships between Cathays, Roath and Penylan are all highlighted indicating an inconsistent spatial scene. In reality, *Penylan* and *Roath* are neighbours, as shown in the Google maps view in Figure 10. To find this inconsistency, the following integrity rules were triggered.

[[<label>Inside_Meet</label><ruleGroup>Topological</ruleGroup>: Region(?x) AND Region(?y) AND Region(?z) AND Inside(?x ?y) AND Meet(?y ?z) AND inside(?x ?z) --> error(?x ?z)]

Where, Penylan is inside Cathays and Roath meets Cathays means that Penylan can not be inside Cathays, but it is hence the rule implies an error.

[[<label>Contains_Inside</label><ruleGroup>Topological</ruleGroup>: Region(?x) AND Region(?y) AND Region(?z) AND Contains(?x ?y) AND Inside(?y ?z) AND Meet(?x ?z) --> error(?x ?z)]

Where, Roath contains Penylan and Roath meets Cathays means that Cathays can not contain Cathays, but it is hence the rule implies an error.

Figure 10. Google Maps View ©2008 Google Map Data ©2008 Tele Atlas

FUTURE TRENDS

Further to the current development of the proposed framework, the following are ongoing research tasks:

- Evaluating the scalability using large ontology bases. In particular, large place ontologies are developed using some real data sets obtained from national mapping agencies and multiple resources on the web.
- Evaluating the value of encoding hybrid spatial reasoning rules using multiple types of spatial relations.
- Explore more expressive DLP fragments for an enriched geo-ontology representation paradigm with increased reasoning and integrity potential.
- Deployment or testing of the proposed framework for realistic application scenarios, for example, representing and maintaining Wikipedia articles.

CONCLUSION

This chapter proposes a new framework for the development and management of geospatial ontologies on the Semantic web. Challenges in representing and manipulating geospatial knowledge in large onotology bases are identified and used to specify requirements for the new framework. OWL as the Semantic Web ontology language was evaluated and its limitations identified to handle those requirements.

To overcome the limitations of OWL, the current approaches to the integration of ontologies and rules are reviewed and classified. A rule layer will allow for the complex semantics of geospatial ontologies to be represented, including the expression of qualitative spatial logic rules for the specification and derivation of spatial properties and relationships.

Description Logic Programs is a homogeneous approach to integrating both OWL-DL ontologies with classical Logic Programs. DLPs was proposed as the most suitable approach as it provides a tractable, and hence scalable, base for the expression of spatial rules and integrity constraints.

The DLP approach was then examined to analyse its benefits and a new framework was proposed in which it can be implemented to serve this domain. The implementation of the framework is briefly sketched and demonstrated using a sample ontology used for demonstration purposes throughout the chapter. Initial testing of the approach demonstrates its efficiency and effectiveness. Evaluation experiments are currently being done to test the scalability of the approach to realistic geospatial ontology development on the Web.

REFERENCES

Abdelmoty, A. I., Smart, P. D., Jones, C. B., Fu, G., & Finch, D. A. (2005). Critical evaluation of ontology languages for geographic information retrieval on the Internet. *In Journal of Visual Languages & Computing* (pp., 331–-358).

Angele, J., Boley, H., Bruijn, J., & Fensel, D. (2005). *Web Rule rule Language language (WRL)*. Retrieved 2008, from http://www.w3.org/Submission/WRL

Bruijn, J., Lara, R., Polleres, A., & Fensel, D. (2005). OWL DL vs. OWL flight: Conceptual modeling and reasoning for the semantic Semantic Web. In *Proceedings of the 14th international conference on World Wide Web* (pp. 623-632). New York: ACM Press. New York, NY, USA.

Carroll, J. J., Dickinson, I., Dollins, C., & Reynolds, D. (2003). *Jena: Implementing the Semantic Web recommendations*. (Internet Report). Retrieved 2007 from http://dsonline.computer.org/0211/f/wp6jena.htm.

Chen, H., & Fellah, S. & Bishr., Y. A. (2005). Rules for geospatial Semantic Web applications. In *W3C Workshop on Rule Languages for Interoperability*.

Cockcroft, S. A. (1997). Taxonomy of spatial data integrity constraints. *GeoInformatica*, *4*(1), 327–343. doi:10.1023/A:1009754327059

Cohn, A. G., Bennett, B., Gooday, J., & Gotts, N. M. (1997). Qualitative spatial representation and reasoning with the region connection calculus. *GeoInformatica*, *1*, 275–316. doi:10.1023/A:1009712514511

Donini, F. M., Lenzerini, M., Nardi, D., & Schaerf, A. (1998). AL-log: Integrating datalog and description logics. *Journal of Intelligent Information Systems*, *10*(3), 227–252. doi:10.1023/A:1008687430626

Egenhofer, M. J. (2002). Toward the semantic geospatial web. In *Proceedings of the tenth ACM international symposium on Advances in geographic information systems* (pp. 1-4). ACM Press.

Egenhofer, M. J., Mark, D. M., & Herring, J. (1999). *The 9-Intersection: Formalism and its use for natural-language spatial predicates*. Report, National Center for Geographic Information and Analysis.

Eisenstadt, M., & Brayshaw, M. A. (1990). Knowledge engineering toolkit: The first of a two-part series presenting a knowledge-engineering toolkit for building expert systems. *Byte Magazine*, *15*(10), 268–282.

Eiter, T., Gottlob, G., & Mannila, H. (1997). Disjunctive Datalog. *ACM Transactions on Database Systems, 22* (pp., 364-418) Vol 22.

Eiter, T., Ianni, G., Polleres, A., Schindlauer, R., & Tompits, H. (2006). Reasoning with rules and ontologies. In *Reasoning Web, Second International Summer School 2006, Lisbon, Portugal* (LNCS 4126, pp. 93-127).

Eiter, T., Lukasiewicz, T., Schindlauer, R., & Tompits, H. (2004). Combining answer set programming with description logics for the semantic Semantic Web. In *KR2004: Principles of Knowledge Representation and Reasoning* (pp. 141–151). Menlo Park, CA: AAAI Press, Menlo Park, California.

El-Geresy, B. A., & Abdelmoty, A. I. (2004). SPARQS: A qualitative spatial reasoning engine. *In Journal of Knowledge-Based Systems, 17,* (pp. 89-102): Vol 17.

Forgy, C. (1982). Rete: A fast algorithm for the many patterns/many objects match problem. *Journal of Artificial Intelligence*, *19*, 17037. doi:10.1016/0004-3702(82)90020-0

Frank, A. (1992). Qualitative spatial reasoning about distances and directions in geographic space. *Journal of Visual Languages and Computing* (LNCS 639, pp. 343–371).

Freksa, C. (1992). Using orientation information for qualitative spatial reasoning. In *Theories and methods of spatio-temporal reasoning in geographic space* (pp. 162–178).

Fu, F., Jones, C. B., & Abdelmoty, A. (2005). Building a geographical ontology for intelligent spatial search on the Web. In *IASTED - International Conference on Databases and Applications part of the 23rd Multi-Conference on Applied Informatics, Innsbruck, Austria* (pp. 167-172).

Gahegan, M. (1995). Proximity operators for qualitative spatial reasoning. In *Spatial information theory: A theoretical basis for GIS* (LNCS 988, pp. 31-44).

Gerevini, A., & Renz, J. (2002). Combining topological and size information for spatial reasoning. *Artificial Intelligence*, *137*, 1–42. doi:10.1016/S0004-3702(02)00193-5

Grosof, B. N. (1997). Prioritized conflict handling for logic programs, In *Proceedings of the International Symposium on Logic Programming* (pp. 197-211). Cambridge: MIT Press.

Grosof, B. N., Horrocks, I., Volz, R., & Decker, S. (2003). Description logic programs: combining logic programs with description logic. In *Proceedings of the 14th international conference on World Wide Web 03* (pp. 48-57).

Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies*, *43*, 625–640. doi:10.1006/ijhc.1995.1066

Haarslev, V., Lutz, C., & Moller, R. (1998). Foundations of spatioterminological reasoning with description logics. In *Principles of knowledge representation and reasoning* (pp. 112–123).

Haarslev, V., & Muller, R. (2001). RACER system description. In *Proceedings of the First International Joint Conference on Automated Reasoning* (pp. 701-706).

Hitzler, P., Haase, P., Krtzsch, M., Sure, Y., & Studer, R. (2005). DLP isn't so bad after all. In *Proceedings of the 1st Workshop on OWL: Experiences and Directions* (OWLED-*05). CEUR Workshop Proceedings* (Vol. 188).

Horrocks, I. (2005). OWL rules, ok? In *W3C Workshop on Rule Languages for Interoperability*.

Horrocks, I. R. (1998). Using an expressive description logic: FaCT or fiction? In *KR'98: Principles of Knowledge Representation and Reasoning* (pp. 636-645).

Jones, C. B., Alani, H., & Tudhope, D. (2001). Geographical information retrieval with ontologies of place. In *Spatial Information Theory: Foundations of Geographic Information Science, International Conference: COSIT* (LNCS 2205, pp. 322-335).

Krotzsch, M., Hitzler, P., Vrandecic, D., & Sintek, M. (2006). How to reason with OWL in a logic programming system. In *Rules and Rule Markup Languages for the Semantic Web, Second International Conference* (pp. 17–28). IEEE Computer Society.

Krötzsch, M., Rudolph, S., & Hitzler, P. (2008). Description logic rules. In *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI-08)* (pp. 80–84). IOS Press.

Levy, A. Y., & Rousset, M. C. (1995). Combining rules and description logics: An overview of CARIN. In *Proceedings of the International Symposium on Logic Programming* (pp. 635). Cambridge: MIT Press.

Motik, B., & Volz, R. (2003). Optimizing query answering in description logics using disjunctive deductive databases. In *Proceedings of the 10th International Workshop on Knowledge Representation meets Databases, Hamburg, Germany* (Vol. 79).

Nectaria, T., & Egenhofer, M. J. (1996). Multi-resolution spatial databases: Consistency among networks. In *Integrity in Databases-Sixth International Workshop on Foundations of Models and Langauges for Data and Objects* (pp. 119–132).

O'Dea, D., Geoghegan, S., & Ekins, C. (2005). Dealing with geospatial information in the Semantic Web. In *Proceedings of the 2005 Australasian Ontology Workshop: AOW '05* (pp. 69-73). Australian Computer Society, Inc.

OGC Technical Committee. (1999). *OGC, abstract feature specification* (Tech. Rep.) Retrieved 2008, from: http://www.opengis.org/techno/specs.htm.

Patel-Schneider, P. F., Horrocks, I., Tabet, H. B. S., Grosof, B., & Dean, M. (2004). *SWRL: A Semantic Web rule language combining OWL and RuleML* (Internet Report). Retrieved 2008, from http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/.

Rosati, R. (2005). Semantic and computational advantages of the safe integration of ontologies and rules. In *Principles and Practice of Semantic Web Reasoning, Third International Workshop, PPSWR 2005, Dagstuhl Castle, Germany* (LNCS 3703, pp. 50-64).

Rosati, R. (2006). DL+log: Tight integration of description logics and disjunctive datalog. In *Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning* (pp. 68-78). AAAI Press.

Sagonas, K., Swift, T., & Warren, D. (1993). *XSB: An overview of its use and implementation* (Tech. Rep). Retrieved 2006, from http://citeseer.ist.psu.edu/195039.html

Sanderson, M., & Kohler, J. (2004). Analyzing geographic queries. In *Workshop in Geographic Information Retrieval*.

Schmidt-Schauß, M. (1989). Subsumption in KL-ONE is undecidable. In *Proceedings of the 1st International Conference on Principles of Knowledge Representation and Reasoning* (pp. 421-431). Toronto, Canada.

Schockaert, S., & Cock, M. (2007). Reasoning about vague topological information. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management* (pp. 593-602).

Schockaert, S., Smart, P. D., Abdelmoty, A. I., & Jones, C. B. (2008). Mining topological relations from the Web. In *DEXA Workshops* (pp. 652-656). IEEE Computer Society.

Servigne, S., Ubeda, T., Puricelli, A., & Laurini, R. (2000). Methodology for spatial consistency improvement of geographic databases. *GeoInformatica*, 7–34. doi:10.1023/A:1009824308542

Smart, P. D., Abdelmoty, A. I., Jones, C. B., & El-Geresy, B. A. (2007). A framework for combining rules and geo-ontologies. In *Web Reasoning and Rule Systems: First International Conference, RR 2007, Innsbruck, Austria* (LNCS 4524, pp. 133-147).

Smith, T. R., & Frew, J. (1995). Alexandria digital library. *Communications of the ACM*, *38*(4), 61–62. doi:10.1145/205323.205340

Volz, R. (2004). *Web Ontology reasoning with logic databases*. PhD thesis, AIFB, University of Karlsruhe.

Vretanos, P. (2005). *Open geospatial consortium filter encoding implementation specification* (Tech. Rep.). Retrieved 2008, from http://www.opengis.org/techno/specs.htm.

Wessel, M. (2001). Obstacles on the way to spatial reasoning with description logics – Some undecidability results. *Working Notes of the 2001 International Description Logics Workshop (DL-2001) Vol. 49*, Stanford, CA, USA.

Wessel, M. (2003). Some practical issues in building a hybrid deductive geographic information system with a DL component. In *Proceedings of the 10th International Workshop on Knowledge Representation meets Databases (KRDB 2003)* (Vol. 79). Hamburg, Germany.

KEY TERMS AND DEFINITIONS

Description Logics: Successors to Semantic Networks and Frame Based Languages which can represent both asserted and structural knowledge. Modern description logics stem from KL-ONE which formalized the ideas of Semantic Networks and Frames. A description logic can describe the world in terms of properties or constraints that specific individuals have to satisfy.

Geospatial Ontologies: A specilisation of ontology that represent only knowledge from the geographic and spatial domain.

Geospatial Rules: Rules or logical rules of inference represent dynamic, relational knowledge as opposed to static structural knowledge. A rule is an inference of the form PREMISE implies CONCLUSION, mimicking human cognitive reasoning processes. If the PREMISE condition holds, then the CONCLUSION condition is deducible. Geospatial rules are those that represent and reason with geographical and or spatial knowledge.

Integration of Rules and Ontologies: How to integrate description logic ontologies with classical rule based logic programs, while preserving semantics and maintaining decidability and tractability. Approaches are either based on the union or the intersection of the two languages.

Interleaved Mixed Mode Reasoning: Allowing knowledge querying using backward rule sets during the course of forward inferencing. Facts in the premise of forward rules can be found directly from explicit facts or implicitly through additional inferences.

Ontology: Those things that *exist* are those things that have a formal representation within the context of a machine. Knowledge commits to an ontology if it adheres to the structure, vocabulary and semantics intrinsic to a particular ontology i.e. it conforms to the ontology definition. A formal ontology in computer science is a logical theory that represents a conceptualization of real world concepts.

Qualitative Spatial Reasoning: Representation of continuous properties of the world by discrete symbols, and then reasoning over such symbols without recourse to more expensive (computationall) quantitative knowledge. Qualitative knowledge and reasoning better mimics human spatial reasoning processes.

This work was previously published in Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches, edited by A. Giurca; D. Gasevic & K. Taveter, pp. 149-169, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Section III
Tools and Technologies

This section presents extensive coverage of the technology that informs and impacts Web technologies. These chapters provide an in-depth analysis of the use and development of innumerable devices and tools, while also providing insight into new and upcoming technologies, theories, and instruments that will soon be commonplace. Within these rigorously researched chapters, readers are presented with examples of the tools that facilitate and support the emergence and advancement of Web technologies. In addition, the successful implementation and resulting impact of these various tools and technologies are discussed within this collection of chapters.

Chapter 3.1
New Paradigms:
A Collaborative Web-Based Research Tool

Hamish Holewa
International Program of Psycho-Social Health Research, Central Queensland University, Australia

ABSTRACT

The chapter aims to document the challenges associated with the management of an international research program and to look at innovative, information technology (IT) based ways of tackling these. Through the medium of a case study, insights gained from practical experience developing and implementing an original Web based collaborative research management tool are discussed. This tool is based on a centralised model of information distribution and access. It was designed following a reductionist analysis of existing research processes and procedures. The ways in which the integration of responsive IT processes into the management of a large international research program have removed redundancies and increased automation and research efficiency are also discussed.

DOI: 10.4018/978-1-59904-970-0.ch005

INTRODUCTION

This chapter presents, through the medium of a case study, insights gained from practical experience developing and implementing an original web based collaborative research tool to assist and enhance the management of an existing, qualitative research program. The case example used is that of the International Program of Psycho- Social Health Research (IPP-SHR). This case study provides the reader with insights into the ways in which information technology (IT) processes can be used to overcome problems associated with the post-modern research environment. Within this context the major challenges are to address the fragmented nature of research locations, staff and project administration within a global setting.

Technological advances have paved the way for global research, enabling it to transcend physical, geographical and cultural boundaries. However, there are still great challenges to be overcome in conducting a truly international research program. The chapter aims to document the challenges associated with the process and management of a large international research program and to look at innovative, IT based ways of tackling these.

The International Program of Psycho-Social Health Research (IPP-SHR) provides international leadership through research, publication, education, media, newsletters and podcasting activities in the area of psycho-social health research. This program, explores a broad range of psycho-social health issues including: the lived experience of serious and terminal illness; haematology and oncology; palliative care; indigenous health; rural and remote health; mental health; obstetrics; bioethics; and the interface between patients and the health care system. The core aim of IPP-SHR is to 'make a difference' by informing policy and service delivery in the real world of health care.

This program utilises qualitative, or naturalistic, research methodology, which seeks to document the voice of the research participant from their own world view (Streubert & Carpenter 1995). Such methodologies are underpinned by a philosophical perspective that listens to, rather than imposes on the experience of others and has a sensitivity to the disempowered and marginalized (Latimer et al. 2003) The large and diverse amount of the data gained from using such methodologies, coupled with diverse and geographical isolated data collection sites of an International program, necessitated the design and construction of a central based management system.

After extensive literature searches in major databases, consultation with software and project management vendors, collaboration and discussion with international leaders in qualitative methodologies, it was evident that no such program existed to meet the specific requirements of IPP-SHR or collaborative multi-site qualitative research projects. As such, to meet the challenges and technical difficulties associated with IPP-SHR's methodology and operation, an internet based research tool was designed. Server side technologies were utilized to achieve a central research portal for IPP-SHR practitioners to use and collaborate through, independent of their physical location. The maturing of server side and Internet connectivity and speed are major contributors to the success of such a system. The system uses a central website, where users with appropriate security credentials like correct user name, password and encryption key can deposit files related to the research processes, implement automatic work flow processes for dictation, transcription and coding processes, view work and project flows and progress, schedule appointments and stipulate task for other users or team members. The system improves research efficiency and lowers research costs. This is achieved through a streamlined website portal offering best practice security, enhanced ethics compliance, limiting or reducing redundancy between processes and team members, and providing accurate information on the process and state of each particular research project. The software also provides team building and mentoring activities through the use of project reporting, a bulletin board, discussion forums and team feedback.

BACKGROUND AND CHALLENGES

The research paradigm and context within which IPP-SHR operates presents unique challenges. Although the program has developed gradually over the last decade, it has only recently evolved to the level of national and international research data collection and collaboration. As a qualitative research program with a focus on the human interface of health care, the challenge is collecting and managing the magnitude and complexity of data gained from naturalistic methodologies over extensive geographical areas. This section details the challenges and problems associated with running a decentralized, location unspecific international research program. It also introduces the equity and ethical considerations associated with research.

IPP-SHR operates in an environment, characterised by the fragmentation of location, staff skills and expertise, participant groups, disciplinary focus and broad topic interests. Translated to the practicalities of research activities, this means, data collection and analysis occurs in many geographical locations and time zones and focuses on a multiplicity of research topics. Additionally staff management needs to address a multiplicity of duties and responsibilities, and access and control of project information, some of which is confidential. Also posing challenges are the practical necessities of enabling simultaneous access by multiple team members to a broad range of specific documents, and creating processes for multi-site data entry and analysis.

As a core component of IPP-SHR's philosophy is to 'make a difference,' and to, 'document the human experience of human illness,' IPP-SHR's qualitative methodology focuses on a phenomenological perspective using exploratory, iterative and open–ended interviews. A phenomenological perspective is used with data analysis and process as its inherent aim is to document and record the particular phenomena or appearance of things as a lived experience (Streubert & Carpenter 1995). Data gathered during the interviewing process is then transcribed verbatim and analysed from the view point of the participiant (McGrath & Holewa 2006). Such analyses and exploration is undertaken without imposing specific theoretical or conceptual frameworks on the interview or data analyses to ensure that the individual experience is recorded (Polit & Hungler, 1995). As such, it is methodologically important for the data management system to not only store data correctly and without corruption but to ensure that research staff can engage in a rigorous data analysis process. IPP-SHR has a particular pride and a documented history of ensuring a rigorous data analysis process by which the findings are driven by a meticulous coding of all statement by participants.

Due to the rigour of IPP-SHR's application of qualitative methodologies, even relatively small research projects produce large amounts of data. For example, a small project in which ten participants are interviewed will on average produce over two hundred pages of language texts, excluding supplementary data such as descriptive statistics. The sheer quantity of data produced by IPP-SHR's qualitative methodologies necessitates that the discrete processes of qualitative research (i.e. verbatim recording and transcription of interviews, managing coding processes) be streamlined.

Although, the data gained in IPP-SHR projects is usually qualitative, occasionally descriptive data is also included. Thus, any system that is to facilitate and streamline IPP-SHR's research processes also needs to be scalable and flexible so as not to disadvantage collaboration and research efficiency if projects required support for different methodologies. Additionally, each project has differing qualitative methodological requirements, which vary according to project size, participant numbers, interviews per participant, and timeframes. As such, any software implementation needs to support IPP-SHR's diverse projects which requires flexibility, scalability and durability, plus continuity of access over extended periods of time.

Additional requirements and challenges posed to the development of a software system are stipulated by the regulatory and policy frameworks within which IPP-SHR operates. Human Research Ethic Councils (HREC) stipulate privacy, informed consent, confidentiality, and audit requirements for research approval involving humans (Australian Federal Government 1988; NHMRC 2003; AIATS 2004). Additionally, audit requirements necessitate data be stored in a confidential and secure location for a period of time from five to seven years. The challenges imposed by such policy and regulatory frameworks require IPP-SHR to store and be legally liable for any information which is gathered throughout the research process. This is particularly important within IPP-SHR's operating paradigm due to the decentralized composition of IPP-SHR research projects and staff. Document control, ethical requirements and privacy issues represent a major concern and challenge for the research process and for any software system designed to support such research.

IPP-SHR research practitioners have not previously been exposed to a high level of IT involvement and had a reticence based on lack of familiarity. Consequently, an additional challenge for incorporation and implementation of the system was the development of comprehensive and supported training packages. The case study highlights the need for mutual understanding from both academic and IT disciplines in developing the system and the positive outcomes that can be gained from incorporating such diverse professional viewpoints. This case study profiles the importance of IT leadership and innovation in meeting the outlined challenges.

DESIGN AND IMPLEMENTATION

Design and implementation of the technological interface created for the research program required a detailed understanding and high level analysis of the challenges, backgrounds, procedures, needs and desires of proposed users of the system. This required both research practitioners and IT consultants to have a high level of cooperation and effective discourse focussed on the needs and wishes of the research practitioners. Although research practitioners bring to their work an understanding of the role and function of IT, this was insufficient for translating their work into an innovative and original incorporated system. What became evident in this experience was the need to provide IT leadership, collaboration and experience in translating the research work components and wishes into a sustainable and useable system.

Design and implementation of the system needed to be in a bottom up fashion. Effective translation, education and implementation assisted in establishing a self perpetuating and learning experience for users. Users are able to see advances that technology can make and can suggest, modify and drive new innovations and uses. Understanding this user determined innovation variable allows users and IT designers to implement user friendly software with practical outcomes which correctly and efficiently operationalises research processes. That is, by a process of continual feedback and discussion between research practitioners and IT consultants, both parties were able to understand and learn the constraints, practicalities and possibilities of automating the research process. Development and specification of the software requirements was partly informed by the Institute of Electrical and Electronics Engineers (IEEE) document (1998) referring to recommended practice for software requirement specifications.

Such in-depth discourse between the two professional groups and the development of a thorough understanding of the research processes allowed for extra innovative features to be created and other unnecessary ones to be omitted. This assisted in producing an effective program with an efficient design. For example, features such as user control and public and private files where not proposed in the original design, however after extensive consultation this feature was warranted as crucial. This was in particular reference to identifiable participant data which, in line with HREC agreements, only authorised persons should be able to view. Without consultation and ongoing discussion between both professional groups, this feature would have been omitted from the final product. Additionally, features proposed by the IT professionals were omitted as they did not add value and could potentially complicate use of the final software interface.

All parties agreed on final software specifications before commencement of programming and software construction. The use of specification allowed for accurate budgeting and timeframe projections. It also allowed for a medium in which users from both professional groups could suggest and implement changes before programming.

This is particularly important in relation to the ethical imperative of efficient and effective use of resources, as changes made to the system once programmed are costly and expensive compared to alterations made in the pre–programming stage. (Diaz & King, 2002)

This considered, laborious and collaborative consultative period before manufacture of the software began is an essential step for any project which bridges two disparate disciplines. Continual communication between research practitioners and IT consultants, coupled with a referable specification documents, allowed for greater flexibility, increased innovation and design features, producing a software package that is of direct use and benefit to its intended users.

WORKFLOW ANALYSIS, PROGRAM SPECIFICATIONS AND IMPLEMENTATION

Extensive research and market searches indicated that no suitable, scalable and location-independent software was available to serve the unique demands and constraints faced by the research group. Although there are numerous research software packages specifically designed for qualitative methodologies (e.g. NUD*IST or Atlas/ti) (Barry 1998) there was an absence of project management software which incorporated qualitative methodologies and associated workloads. In order to fill this gap, a software program was designed to handle such programmatic demands and incorporate data analysed within the major qualitative data analyses programs. A full working commercial version of this program can be found at www.quadrant-pm.com

A reductionist analysis of the workflow and procedures that IPP-SHR uses to operationalise its research was critical for successful design and implementation of the software program. This process involved detailing and referencing every activity conducted by researchers. All components produced by the research process were discretely assigned a unique identifier. Components were then tracked in order to ascertain a step by step understanding of the research process. For example, the interviewing of a participant produced a component, "audio voice file," this file was then forwarded to the transcriber, who by transcribing the audio voice file produced another component, "text document of interview verbatim". The flow and process of information creation (arrangement of participant interview) to research output (publishing and dissemination of information) was tracked and compiled into a flow diagram. (See Figure 1)

After construction of the flow diagram each process was analysed to identify redundant processes and avenues through which automation and efficiency could be increased. For example a redundant process found was the duplication of communication occurring between team members in the activity of scheduling an interview. Although, in its simplest form, scheduling can be achieved through one person (enrols the participant, schedules and conducts the interview, compiles and stores documents relating to the interview and transcribes the interview) this is a rarity due to the aforementioned operating challenges of the research program. It is IPP-SHR's experience that there can be up to four people involved in the process of arranging, conducting and transcribing an interview. Similar issues of redundancy and repetition of tasks were found within the interface between transcription, coding and recording keeping.

Analysis of data flow and information creation in Figure 1 suggested that many tasks undertaken within a project only need be completed once if every member involved had access to such information. As such, a centralised distribution model of information storage, provision and access was developed to enable this streamlining to occur. Using the centralised model, users of the system, with appropriate rights and security credentials, have access to each component of information that is relevant to the specific needs of their task. This enables users to maintain up-to-date information and accurate data repositories, which assists in smooth and efficient running of research processes. Furthermore, duplication, redundancy and errors are reduced as only one record exists per each research instance per project. Instead, for example, of four team members keeping individual records of each particular research instance and co-ordinating changes between each other, only one record is kept and only one change is made. Using the model in relation to interviewing, scheduling would be provided to the centralised data source (such as a web based calendar) and information given to other members associated with such interview will be provided from such source.

Figure 1. Procedural, sequential information creation and flow (©2007, Hamish Holewa. Used with permission)

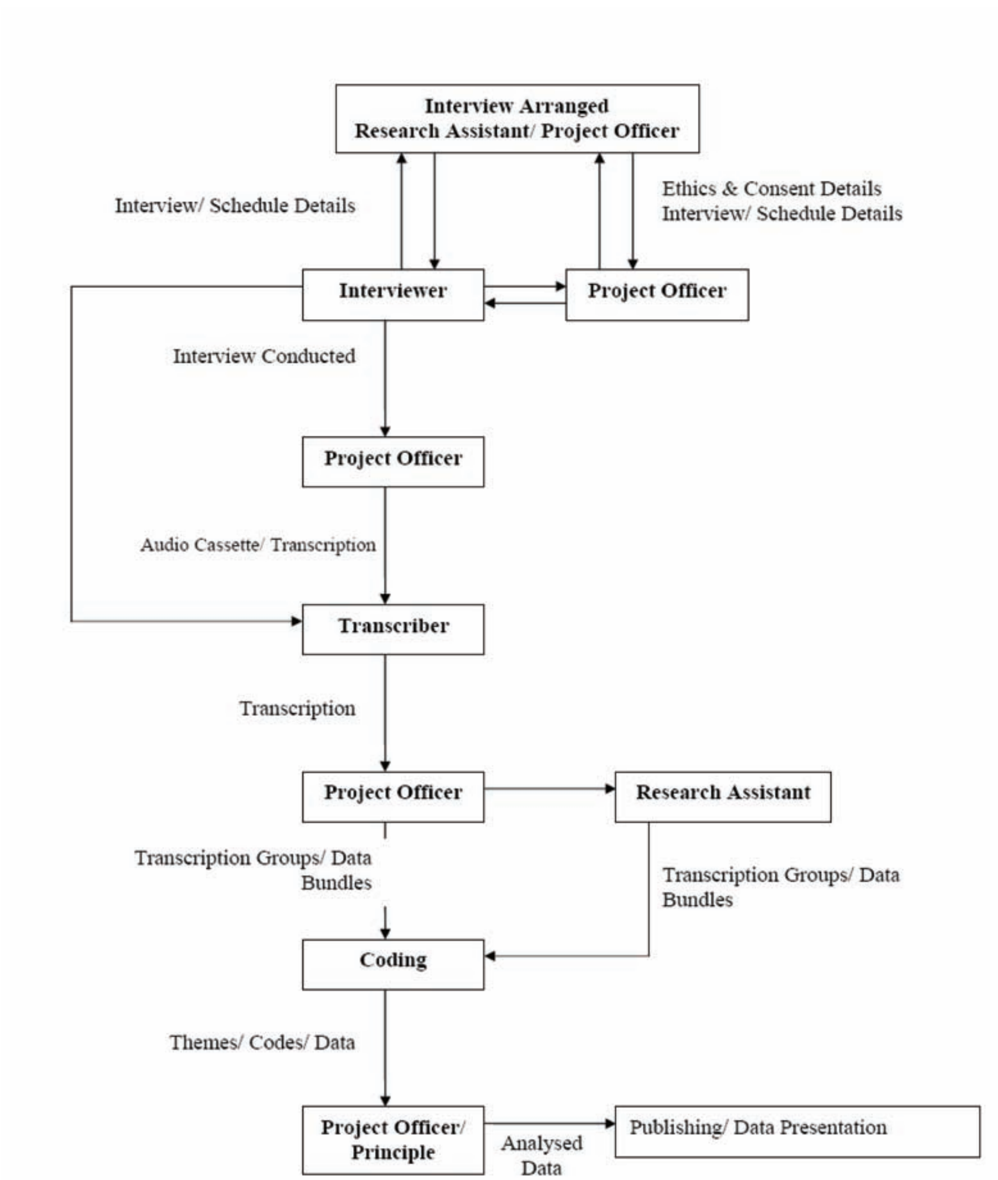
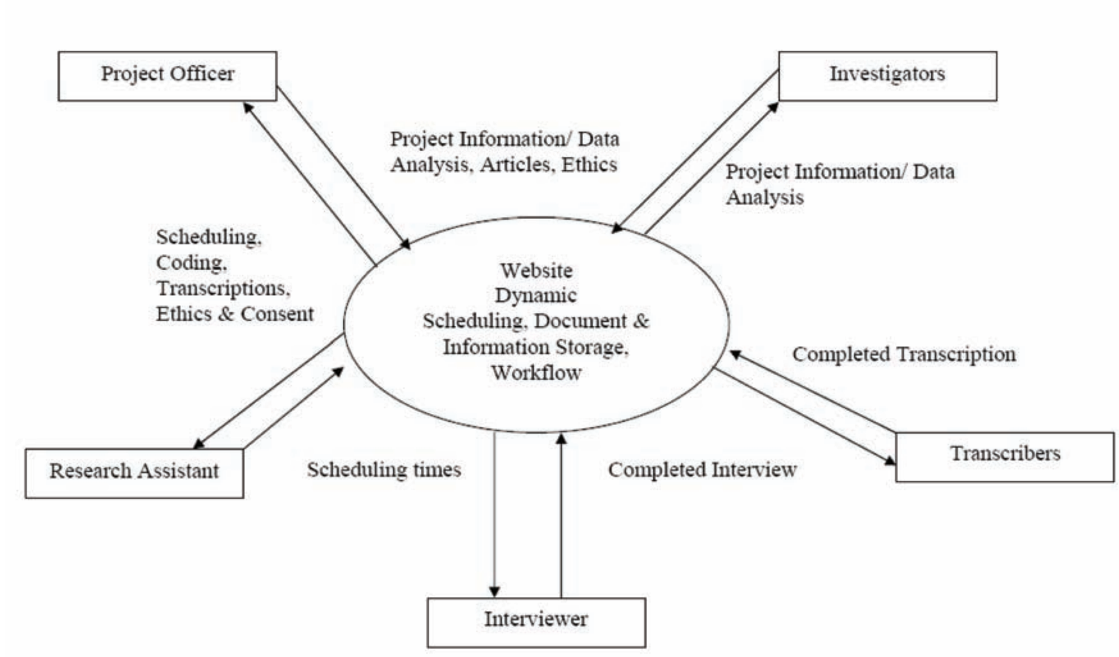


Figure 2. Centralised and distributed information flow (©2007, Leena Hiltunen. Used with permission)



As shown in Figure 2, research staff interacts individually with the central dynamic website and information flow occurs via a centralised, distributed model. This distinctly contrasts with the processes formerly used, whereby information flow was procedural, sequential and task oriented. Once information has been created, research team members with appropriate security credentials can access the most up-to-date source of information through a website portal without needing to ask other team members for the status and latest version of such information.

The depicted system in Figure 2 also removes unnecessary and redundant tasks with the motivation of increasing automation and realising efficiencies through the process. Once, implemented and configured, simple procedural and reporting tasks are completed automatically. In previous work flow the researchers were involved in a six step processes which included: (1) research assistant (RA) enrols participant; (2) schedules an interview time; (3) organises HREC compliance such as informed consent and project description procedures; (4) the interviewer (INT) completes the interview and produces a digital recording; (5) the recording is sent to the transcriber (TRS) who completes transcription processes; and, (6) send completed transcription to RA or project officer (PO), who initiates coding and qualitative analysis. At any point in the process, each member is reliant on another member to complete their scheduled task, record appropriate details and ensure HREC and other policy guidelines are complied with. Furthermore, the PO or investigator may have to contact up to four people to receive an accurate report on the progress of the research project.

A major advance has been achieved on the automation and reduction of the described processes. This is done through the automatic storage and referencing of documents and information (participant details, consent forms, digital interview files, transcriptions) which is completed at each stage and by each individual person responsible for such tasks. Once each task is completed by the assigned research team member, the team member interfaces with the website portal and flags its completion and uploads the accompanied file or information. The website will then contact the next research team member within the process. For example, once an interviewer has completed the interview, this is flagged within the system and the corresponding sound file is uploaded to the website portal. Once uploaded, the website sends an alert to the transcriber of the need for an interview to be transcribed. This process reduces record keeping times and ensures an up-to-date report on progress.

The distributed, centralised model of project work flow implementation has been realised through the use of four discrete online modules operating within the project management portal. These four modules are as follows: calendar and scheduling; project workflow and progress; document and version control; and administration. Included are different levels of user access and control and differing levels of information categories, ranging from public to private. Access to the website portal occurs through a computer with internet connectivity and browser support. The website server runs on a combination of client and server side code attached to a centralised database. Most computation and processing is completed on the remote server and, apart from internet connectivity, there is little computational demand on client computers.

SELF DIRECTED LEARNING COMMUNITIES

The development of a community, centred upon online collaborative applications, is a well regarded method for increasing the use and accessibility of a system and for encouraging users to engage in self directed interaction and learning (Neus, 2007). The implementation and use of community based initiatives has been successful within this case study. The provision of a message board, collaborative document sharing applications, version control and work processes has encouraged community building and increased collaboration and use of the service. It has also facilitated self directed learning and enquiry and has supported a forum for public discussion.

Evidence associated with the development of an online community and self directed learning stemmed from online dialogue evident within the bulletin board feature and the minimum formal training needed for new users. Through anecdotal records and conversions with users of the system, it was noted that software training and learning was largely conducted within the system. Although, software training was required for initial users of the system, the historical dialogue between such users provided a self directed learning environment for new users, and assisted in development of an environment conducive to posting public comments and asking questions to the group of users. However, it should be noted, that the development of such a community of self- directed learning may be further successful due to the separate and disparate physical location of the users. It is unclear whether such development of a bulletin board and self- directed learning would have developed if user where located in the same physical space, e.g. in the same office.

The central place assumed by the software within the research program creates a responsive medium through which staff concerns, ideas and issues can be raised. This enables research staff to pose research-related questions and comments, and provide direct feedback on the program's usability. The facilitation of such open communication allowed for quick development of ideas and creative problem solving. It also helped to foster team culture and shared identity, whilst acting as an effective training tool.

The software program's ability to dynamically provide users with up-to-date project information has encouraged individual and team ownership of the research projects. As the program allows for instant viewing of research project status and progress, research members are also able to identify any aspects of the research process hindering overall progress. Additionally, staff training and supervision, early identification of potential problems, and feedback were facilitated by the availability and ease of gauging project progress.

FUTURE ADVANCES AND ETHICAL IMPLICATIONS

It is anticipated the further development of the system will occur through a user driven, evolutionary response. Increased and synchronised communication avenues are predicted with a decreasing need of physical presence. This is particularly evident with the advent of increased access and data throughput, as a result of increasing bandwidths and the emergence of new and maturing technologies (Choudrie & Dwivedi, 2007). It is envisaged that access to the research collaboration tool will occur through more diverse means, and not be limited to a computer with Internet connectivity. This has strong positive implications for the processes of qualitative research, as many research activities are conducted in the setting of the participant and away from traditional computing equipment. There is potential to limit the redundant activities presently associated with scheduling, interviewing and recording by allowing handheld or portable devices access to the software.

Additionally, the immediacy of information access between research staff and the potential for real-time communication is an important avenue for future advances. Instant Messaging and Voice-Over-Internet-Protocol advances have the potential to contribute to stronger group collaboration, communication and a sense of community. The use of such technologies also has the potential to lower communication costs and providing greater access to research opportunities for marginalised or disadvantaged groups. The adoption and inclusion of such technologies, has the opportunity to lower research costs, providing greater potential outputs from fewer resources.

Increased security represents another domain in which future advancements may occur. This aligns with the ethical imperative of maintaining rigorous privacy and confidentiality protocols within the research process, particularly when using centralised models of information storage collaboration and project management work. The transfer, storage and processing of information within the aforementioned system is governed by strict HREC and other policy requirements. Effective system security, regulating access and use, is paramount to fulfilling the ethical and policy requirements of privacy, confidentiality and cost effective use of research resources. Although the system currently uses best practice security practices, such as encryption, user access and password, security certificates and user training, the advent of increased processor and computational capabilities and increased internet users and connectivity means that security issues are a priority area for ongoing future advancement.

The software program has the potential to increase user engagement between researchers, health professionals and stakeholders in qualitative research processes, especially for groups whose lack of access to traditional research institutions may have formerly acted as a barrier to participation. This use of IT solutions within qualitative research has the potential to bridge geographical and social communication gaps, opening up the potential for increased collaboration and information sharing between researchers worldwide. From IPP-SHR's experience implementing and using such a program, it is evident that the integration of IT management structures into a qualitative research program can be instrumental in bridging geographical constraints and fostering global research collaboration. The potential of this new technology is clearly evidenced by the way in which it has enabled IPP-SHR to effectively engage in research collaboration that spans national boundaries and multiple time zones.

CONCLUSION

This chapter has presented, through the medium of a case study, insights gained from the practical experience of developing and implementing an original web based collaborative research tool to assist and enhance the management of an existing qualitative research program. This tool is based on a centralised model of information distribution and access and was designed following a reductionist analysis of existing research processes and procedures. The ways in which the integration of responsive IT processes into the management of a large international research program have removed redundancies and increased automation and research efficiency has also been discussed. The program opens up avenues for increased research participation and collaboration and makes an important contribution to overcoming the challenges that a fragmented, globalised environment poses.

REFERENCES

AIATS. (2000). Ethical guidelines for research. Australian Institute of Aboriginal and Torres Strait Islander Research. Retrieved March 18,2007, from http://aiatsis.gov.au/research_program/gratns/gratns_assests/ ethics_guidelines.pdf

Australian Federal Government. (1988). *Privacy Act 1988* Australia Privacy Policy. Australian Federal Government. Retrieved March 18,2007, from http://www.comlaw.gov.au/comlaw/management.nsf/lookupindexpagesbyid/IP200401860?OpenDocument

Barry, C. A. (1998). Choosing Qualitative Data Analysis Software: Atlas/ti and Nudist Compared *Sociological Research Online*, 3(3). Retrieved March 17, 2007, from http://www.socresonline.org.uk/socresonline/3/3/4.html

Choudrie, J., & Dwivedi, Y.,K. (2007). Broadband impact on households consumers: online habits and time allocation patterns on daily life activities. *International Journal of Mobile Communications*, *5*(2), 225–241. doi:10.1504/IJMC.2007.011817

Diaz, M., & King, J. (2002). How CMM impacts quality, productivity, rework and the bottom line. *CrossTalk, 15*(3), 9–16.

IEEE – Software Engineering Standards Committee of the IEEE Computer Society. (1998). *IEEE recommended practice for software requirements specifications.* USA

Latimer, J., et al. (Eds.). (2003). Advanced Qualitative Research for Nursing. Oxford, Malden, MA: Blackwell Publishing.

McGrath, P., & Holewa, H. (2006). Missed Opportunities: Nursing insights on end-of-life care for haematology patients. *International Journal of Nursing Practice*, *12*(5), 295–301. doi:10.1111/j.1440-172X.2006.00585.x

Neus, A. (2007). Managing Information Quality in Virtual Communities of Practice. Lessons learned for a decade's experience with exploding Internet communication. *The 6th International Conference on Information Quality at MIT.* Retrieved March 18, 2007, at http://opensource.mit.edu/papers/neus.pdf

NHMRC. (2003). Values & ethics: guidelines for ethical conduct in Aboriginal and Torres Strait Islander health research. National Health and Medical Research Council. Retrieved March 18, 2007, http://www.nhmrc.gov.au/publications/synopses/e52syn.htm

Polit, D., & Hungler, B. (1995). Nursing research: principles and methods. 5th Ed. Philadelphia: Lippincott.

Streubert, K., & Carpenter, D. (1995). Qualitative research in nursing: Advancing the humanistic imperative. Philadelphia: Lippincott.

KEY TERMS

Information Technology (IT): As defined by the Information Technology Association of America, IT is "the study, design, development, implementation, support or management of computer-based information systems, particularly software applications and computer hardware."

Phenomenology: A method of inquiry based around the exploration, description and analysis of a particular phenomenon, untainted by presupposed theories, beliefs and assumptions.

Psycho-Social Research: Research processes which aim to explore and document social and psychological aspects of the human experience.

Qualitative Research: A non-numerical research methodology which aims to describe and understand, rather than explain, human behaviours and experiences.

Redundant Processes: Unnecessary process or tasks that through analyses of such process can be removed without affect the output of such tasks.

Server Side Technology: A form of web server technology in which users' requests are fulfilled by running a script directly on the web server to generate dynamic HTML pages. It is used to provide interactive web sites capable of interfacing with databases and other data stores.

This work was previously published in Handbook of Research on Digital Information Technologies: Innovations, Methods, and Ethical Issues, edited by T. Hansson, pp. 57-67, copyright 2008 by Information Science Reference (an imprint of IGI Global).

Chapter 3.2
Adaptability and Adaptivity in The Generation of Web Applications

Raoudha Ben Djemaa
MIRACL, Tunisie

Ikram Amous
MIRACL, Tunisie

Abdelmajid Ben Hamadou
MIRACL, Tunisie

ABSTRACT

This article proposes a generator for adaptive Web applications called GIWA. GIWA's objective is to facilitate the automatic execution of the design and the generation of Adaptable Web Applications (AWA). Characteristically, the effort in this work has to be pursued with special attention to both issues applied to AWA: adaptability and adaptivity. The architecture of GIWA is based on three levels: the semantic level, the conceptual level and the generation one. Using GIWA, designers specifies, at the semantic level the features of Web application. The conceptual level focuses on the creation of diagrams in WA-UML language; the extended UML by our new concepts and new design elements for adaptation. At the generation level, GIWA acquires all information about users' preferences and their access condition. Consequently, the generated pages are adaptable to all these information. An evaluation and a validation of GIWA are given in this article to prove our adaptation.

INTRODUCTION

The growing demand for data-driven Web applications has led to the need for a structured and controlled approach to the engineering of such applications. Both designers and developers need a framework that in all stages of the engineering process allows them to specify the relevant aspects of the application. The engineering becomes even more complicated when we include notions of adaptation. Here, we address both adaptations during the presentation generation, for example to reflect user preferences or platform used, as well as adaptation inside the generated presentation.

The need for adaptation arises from different aspects of the interaction between users and Web applications. Users' categories which deal with these systems are increasingly heterogeneous due to their different interests, preferences, and the use of number of devices (PC, WebTV, PDA, WAP phone, etc...). User's preferences and interests can be deduced from his and browsing history.

Adaptive Web engineering is meant to provide a systematic and disciplined approach for designing, generating and maintaining adaptive Web applications (Cingil, 2000). For this reason, recently several models and methodologies have been proposed for supporting the development of adaptive Web applications. The main goal of such models is to help designers to reason in a structured way about aspects that are specific to hypermedia, such as links, structure and navigation, and to express adaptation in the design process. Moreover, such models and methodologies should help engineers to manage the overall complexity of Web development which requires a variety of activities, such as organizing the structure, choosing the contents and the presentation modality, some of them involving automated generation of Web page (Brusilovsky, 1998). So, methodologies usually provide guidelines for performing such activities and suitable models for expressing the results of such operations.

In our previous works (Ben Djemaa, 2006a; 2006b, 2006c; Ben Djemaa, 2007; Ben Djemaa, 2008) we have presented a methodology for AWA which guides the designer through different steps of the design process, each of them yielding a specific model that is being interpreted by the GIWA tools. The GIWA methodology is based on several following steps: requirement analysis, conceptual design, adaptation design and generation.

The requirement analysis step (Ben Djemaa, 2005) represents the application domain. This step expresses the purpose and the subjects of the Web application through the functionality model and defines the target audience through the audience model. The result of these two models is a set of audience classes together with an informal description of their functional space. In GIWA, the functional space is determined by a semi automatic algorithm called AGCA.

In the Conceptual Design step (Ben Djemaa, 2008), the functional space for each audience class is represented using traditional conceptual modeling: use case diagram, sequence diagram, class diagram, etc. In GIWA, conceptual model is represented in a specific notation called Web Adaptive Unified Modelling Language (WA-UML) (Ben Djemaa, 2008). This new notation increases the expressivity of UML while adding labels and graphic annotations to UML diagrams. This extension of UML defines a set of stereotypes and constraints, which make possible the design of conceptual model. These models are translated and exported in XML files in a data repository.

The adaptation design level (Ben Djemaa, 2007) is based on the profile model, which takes into account the user's devices capabilities (hardware and software), Users' preferences presentation (desired layout, navigation patterns, etc.) and personal information (eg. Age, sex, language, etc...).

In this article we concentrate on the generation level. At this level the designer is invited to instantiate previous models using the specific interfaces offered by GIWA. Only the aspects related to the two first levels (requirement analysis and conceptual design) are instantiated by the designer. Information related to the devices' capabilities are dynamically captured by the system (using Logs files) and then stored in the profile model. At the end of the step of instantiation, the GIWA deployment can be launched.

Characteristically, the effort in this work has to be pursued with special attention to both issues applied to AWA: adaptability and adaptivity. Adaptability can be defined as the facility of an application to be configurable according to a set of decisions taken by the user, which usually define his preferences and/or background. Whereas adaptivity denotes the capacity of the application to alter the profile model according to the user's behaviour during the application run and adapt dynamically to the current state of the user model to any user.

The article is structured as follows. In Section 2 we provide an overview of related works. In section 3 we present an overview of the different models of GIWA. Section 4 presents the architecture of GIWA and some examples of interfaces which illustrated the prototype. In section 5 we present our experimental design and the carrying out of the experiment using the design. Finally, section 6 concludes the article and suggests future research directions.

RELATED WORKS

For a long time, Web application engineering has been synonymous with ad hoc development and not supports a systematic process. Aspects like adaptation and generation process complicate the design process of Web application engineering and bring its complexity beyond the level that is easily handled by a single human developer. Therefore, to support a systematic development process, a strong methodology (supported by a suite of tools) can help to keep the design process at a practical level. Recently, different approaches for modeling and engineering adaptive hypermedia system have emerged. Approved hypermedia design principles, such as those defined in OOHDM (Schwabe, 1998) or in RMM (Isakowitz, 1995; Isakowitz, 1998) have been enhanced with the notions of adaptation and personalization in a further extension of OOHDM (Rossi, 2001) or the RMM-based Hera methodology (Frasincar, 2001; Frasincar, Houben, 2002). UWE (Koch, 2000, 2001) included a design methodology for adaptive hypermedia applications (AHDM) and a development process for such applications (AHDP). In the AMACONT project authors have introduced a component-based XML document format (Fiala, 2003). This project enables to compose adaptive Web applications by the aggregation of reusable document components.

All these methodologies were originally designed for Adaptive Hypermedia Application (AHA) and do not deal comfortably with Adaptive Web Application (AWA). These methodologies are very much data-driven or implementation oriented and do not covers the lifecycle of adaptive Web applications. Still, most solutions have been originally developed for a manual hypermedia design process and are not particularly well-suited in the context of automated hypermedia design.

Methodologies like RMM, UWE or AMACONT are not specifically targeted to support dynamic adaptation. For these methodologies, personalization means that the application acknowledges the user's situation and its information delivery are adapted. They may be able to solve adaptability problems to some extent but they do not address the real problem of adaptivity relating to the devices' capability (hardware and/or software). On the other hand, most of the currently existing methodologies lack a profile model that would allow for the design of truly adaptive Web applications. In fact, this model can play a significant role in such applications: our aim is to include this aspect in the personalization of the hypermedia presentations that get generated. Generating adaptive presentations requires a clean separation of concerns, as is advocated in (Frasincar, 2002).

OVERVIEW OF GIWA MODELS

In GIWA, Web applications data are defined by different models such functionalities model, audience model and profile model.

The Functionalities Model

Met with the increasing needs of Web applications users, we propose a functionalities model, gathering users' informational and functional needs. In this model, functionalities are classified into three functional classes: Static Informational (SI), Dynamic Informational (DI) and Professional (P) ones. The Static Informational functional class gathers all functionalities enabling the users to have access to the system to acquire static information being in a specific URL Web page. This class translates the set of the static informational users' needs. Using hypertext and hypermedia links, the user is capable to exploit a functionality of information's consultation.

The Dynamic Informational functional class regroups all functionalities enabling the users to have access to the system to acquire dynamic information. This class translates the set of the dynamic informational users' needs. Using techniques of research by a search engine, the user is capable to exploit a research functionality of information.

The Professional functional class is devoted to the representation of the users' functional needs. Indeed, Web applications are based on technologies which make their contents dynamic, enabling, thus, the user to modify the applicative state of the server by carrying out a set of functionalities.

Each of these functional classes, represented above, will be decomposed by the designer into a set of elementary functionalities representing both informational and functional needs. These functionalities, describing every functional class, are three:

- Static Informational Functionality (SIF) displaying a static Web page being in a specific URL.
- Dynamic Informational Functionality (DIF) displaying a dynamic Web page constructed from one (or several) "SELECT" server query (s). It is about a selection from the database without affecting the applicative server's state.
- Profession Functionality (PF) displaying a dynamic Web page constructed from one (or several) server query: UPDATE, ADD or DELETE. The execution of this functionality affects the applicative server's state.

At the lowest level of the Functionalities Model, the concept "Functional Space" provides each actor with a list of authorized functionalities.

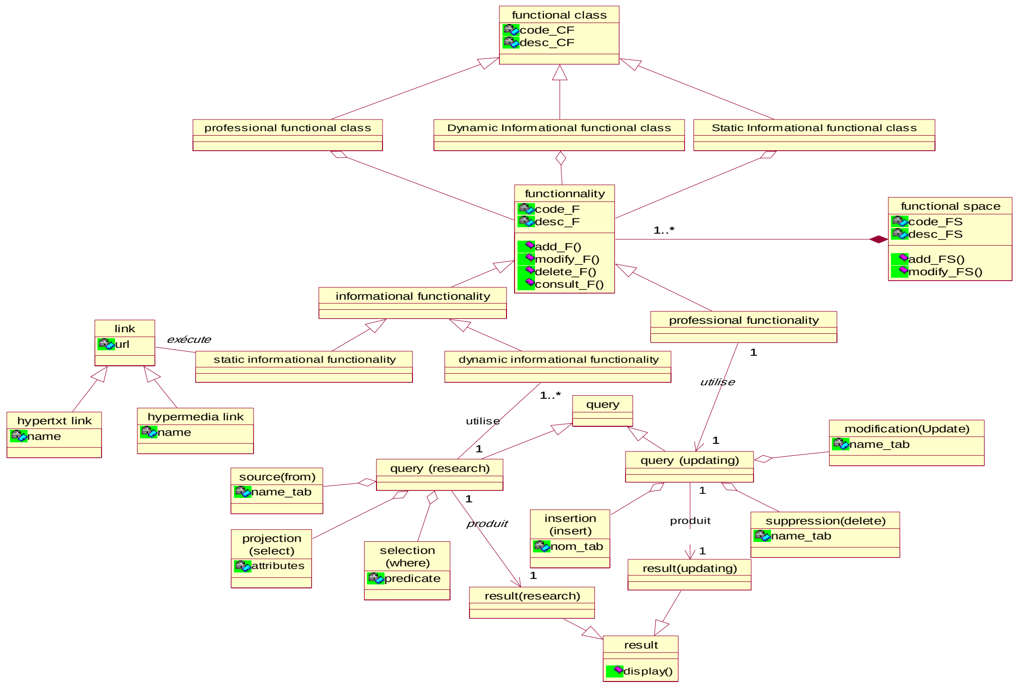
In Figure 1 we present the Functionalities' meta-model (using the UML notation). This figure illustrates the inheritance and composition links existing between the concepts constituting the Functionalities Model (Functional Class, Functionality and Functional Space). We notice that the Informational Functionality class executes link (hypertext and hypermedia link classes) and the Informational Functionality class uses a query of consultation (Query research class) which makes it possible to select and to organize the presented information. In this case, the Query class is composed of several classes representing various clauses of a simple query like in SQL. In this meta-model we present:

- The SQL Select clause by a Projection class which defines the expected structure of the query result;
- The SQL From clause by the Source class which introduces the collections from which the result is built;
- The Where clause by the Selection class which specifies a predicate allowing to filter the collections;

A query produces a result modelled by a Result class. A relation of dependence exists between the classes Query and Result: any modification made to a query has a direct influence on the result. In addition, the Profession functional class uses another type of query which updates the database system (Query Updating class). It also implements the composition of the classes representing the following clauses: The Insert clause represented by the class Insertion; the Delete clause represented by the class Delete and the Update clause represented by the class Modification.

In our approach, the Functionalities Model implementation is carried out in two stages. During the first stage, the designer is invited to define

Figure 1. Representation of the meta-model of the model of the functionalities



the first three levels of the model. The fourth level, relating to the functional space of the actors, is approached only after having defined the list of the actors involved in the application. However, this list is defined in the following section.

The Audience Model

The goal of this model is to define the actors' list of a Web application. Indeed, for this type of application, besides the human users (defines as of the physical actor) that exploits the Web system, we can distinguish services (representing roles played by the human users) or systems (devices, data processing system, Web service,…). In this context, to take account these distinctions, we propose three categories of actors classified as follows: physical actor, logical actor and system actor.

Physical actor represents a human user (or human user group) who visits the application Web. This actor interacts with the system to search or consult information (to execute an informational functionality) and possibly to modify the state of the system (to execute a professional functionality). For example, in the case of an application of library in line; Subscribers and Visitors are some physical actors.

Logical actor represents a role played by a human user (or a human user group) to assure the maintenance of the Web application. This actor assures all actions and functionalities that participate in the configuration and the administration of the system (to execute a professional functionality). For example, in the case of a Web application of a library the Webmaster and the Bookseller are some logical actors.

System Actor represents a computer system, an access device or a Web service, etc. These systems are connected generally to the application to provide news to the system or to update data. It is the external sources that will be charged automatically in the system.

In our approach, we have proposed an algorithm which generates the list of these different actors of the Web application starting from the concept of Functionality. We have presented in (Ben Djemaa, 2007), the process of operations of this algorithm.

The hierarchy of actors generated by this algorithm cannot describe a model for users because the definite actors are not all human actors. Therefore, in the goal to define an audience model (that leans logically on the human users), we propose to differentiate the human users by the concept "audience class" to inhuman one (systems) and, therefore to distinguish between "logical audience class" and "physical audience class".

By definition, an audience class is a potential user group which belongs to the application target audience and which has the same informational and functional needs. These classes are not necessarily disjoined (a user can belong to several classes of audience).

In Figure 2 we present the meta-model of our audience model. This figure illustrates the inheritance and composition links existing between the concepts constituting the Functionalities Model (Functional Class, Functionality and Functional Space).

To evaluate the audience model, we have proposed in (Ben Djemaa, 2005) an EPMA (Evaluation Process of the Audience Model) which evaluates the result generated by the algorithm of actor generation. This process is based on several mathematical symbols and formulas follow to check the distribution of the informational and functional needs between the actors for the application.

Profile Model

In our approach we proposed a new model specific for adaptive Web application called Profile Model (Ben Abdallah, 2008). In Figure 3 we show the different dimensions treated in this model to generate adaptive Web applications.

This model represents an abstract specification of the presentation in terms of users' profiles. In

Figure 2. Meta-model of the audience model

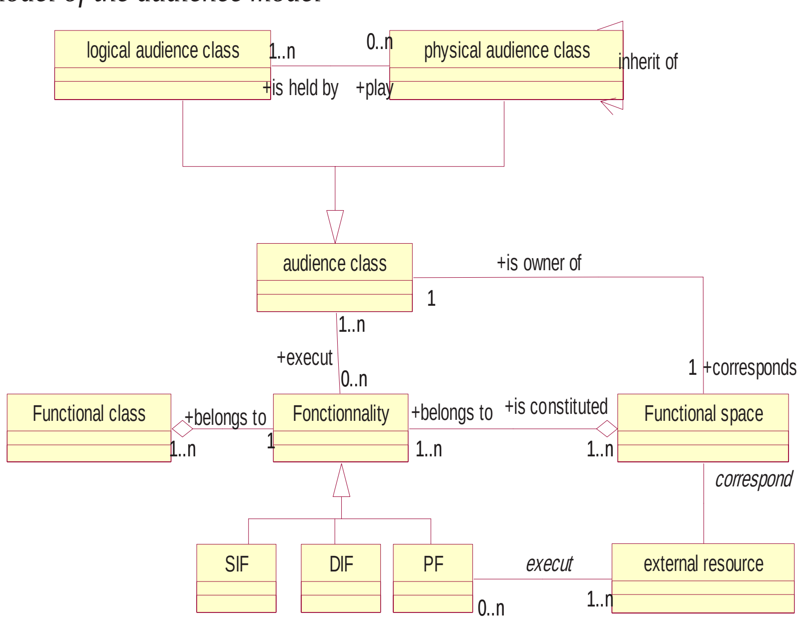
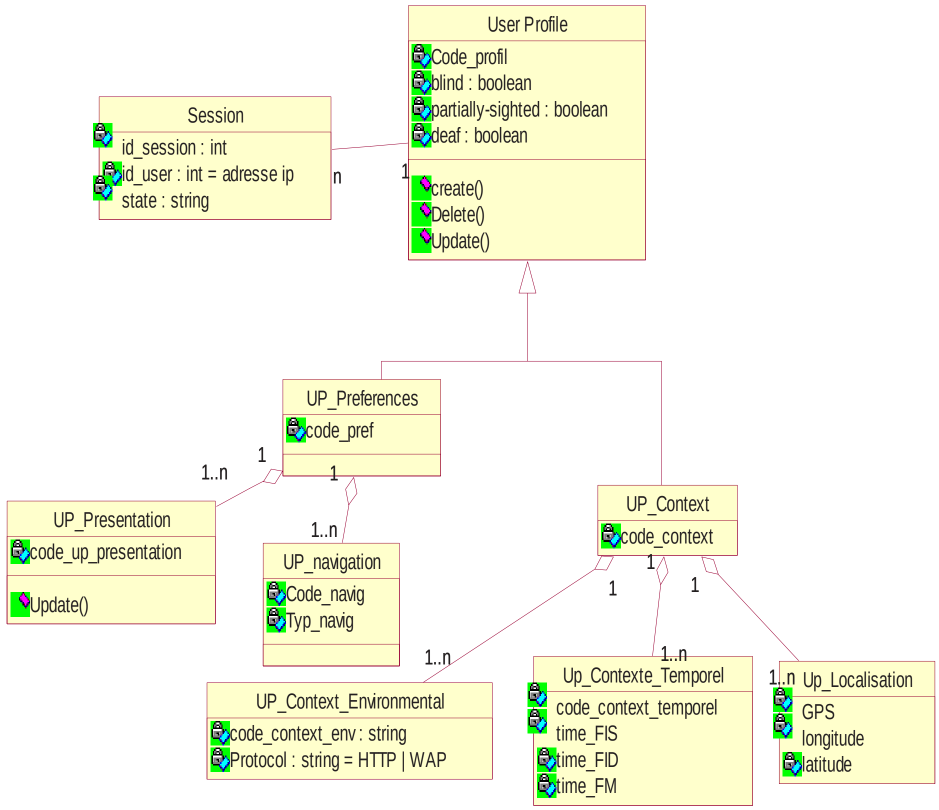


Figure 3. Profile model



fact, this model needs to take into account the users' preferences (Up_preferences) and the user's context (UP_context). The first aspect is composed of preferences of presentation and preferences of navigation. The second aspect dealt with the environmental context (hardware and software), the temporal context (the preferred time to execute a FIS, FID or FM functionality) and the localisation (GPS information's) context.

Users' preferences presentation will be defined through specific techniques of data presentation and different media in the Web page. In fact, to take into account the adaptation of the different media in a Web application we have defined a model of media adaptation which presents preferred choices for users for each type of media: visual, video and audio one. (cf. Figure 5).

UP_context_Environmental has a number of components, each component grouping a number of attributes. In Figure 4 we defined three profiles components (UP_Network, UP_Hardware and UP_software). UP_Hardware component has a number of attributes (eg. Support_image specifies if the device is able to display images and ScreenSize defines the dimensions of the device display).

Both adaptability and adaptivity are considered in our approach. These concepts are treated differently in the profile model. Users' preferences presentation will be defined through specific charters (composition and graphical charters). Each of these charters can be choice by the user after the generation of his Web application (adaptability) and he can also modify some information at his system use (adaptivity). This dimension of the profile model will be detailed in the following section. Users' preferences navigation and information about the context of user is a

Figure 4. Environnemental context / user profile

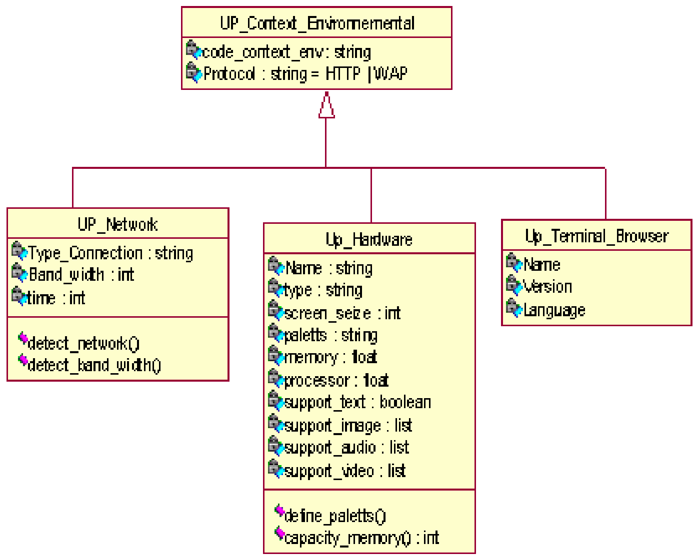
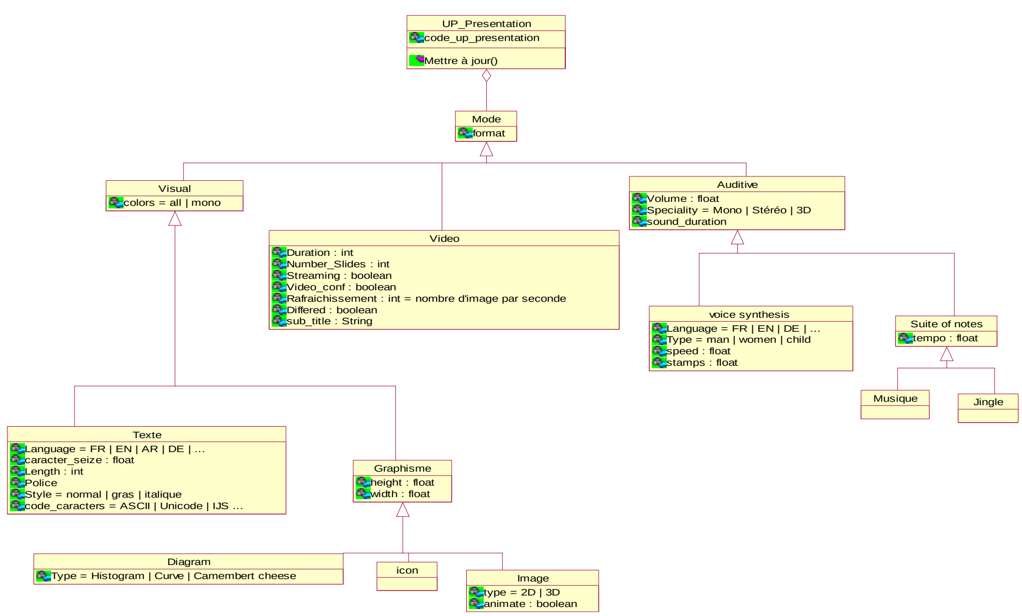


Figure 5. Model of media adaptation in GIWA



dimension which is automatically captured by GIWA (adaptivity).

To generate Web interfaces, the user does not need all dimensions or sub-dimensions or all information characterizing a dimension. A profile is thus an instantiation partial of this model according to the user's needs, to the type of application and to the execution environment.

Different engines and processes are defined in GIWA to instantiate these models and to generate adaptable interfaces Web. The architecture of GIWA is detailed in the following section.

ARCHITECTURE OF GIWA

GIWA's target is to facilitate the automatic execution of the design and the automatic generation of adaptive Web interfaces. It should be possible to program the Web applications in such a way that it can automatically execute the process specified by the design. Figure 6 depicts three different activities of the proposed generator GIWA: Semantic level, Conceptual level and Generation level.

The semantic level instantiates specific data contents of the Web application defined by different semantic model us functionality model, audience model and profile model.

The conceptual level focuses on the creation of diagrams in WA-UML. In fact, in this level we propose an AGL which supports the new design elements that we proposed. This AGL is based on Argo_UML. Thus, in this extension of Argo_UML, we introduce new types of diagrams to represent the different diagrams of our extension, namely, WA-UML.

The generation level focuses on the process of Web page generation and describes how the generator GIWA dynamically adjusts to varying user preferences into chosen implementation platform (HTML, WML, SMIL, etc.).

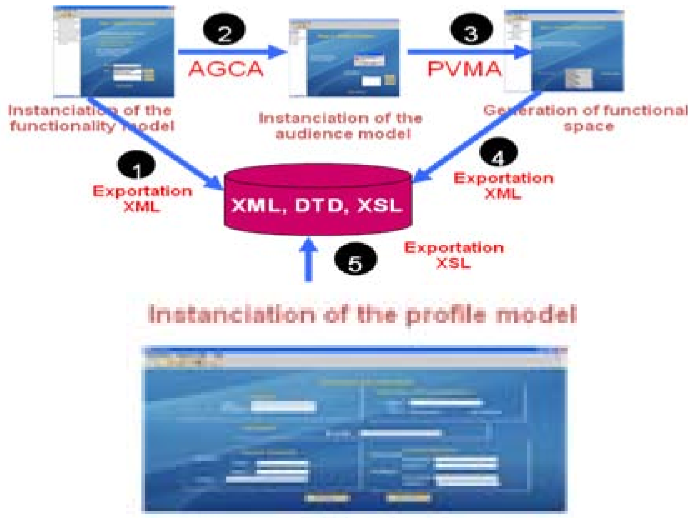
Semantic Level of GIWA

Figure 7 depicts different steps of GIWA in the semantic level. For a designer modelling an adaptive Web application using GIWA consists firstly, in instantiating the functionalities' model.

Once instantiated, this model is translated into XML files to be stored in the data repository ❶ and the system execute the algorithm of generation of audience classes ❷ which built the audience model. Then this model is validating by a specific process called PVMA❸. The last is also translated into XML file ❹ which contains the functional space for each audience class. At this stage, the content is adapted to each audience class. But to adapt the user's presentation preferences, the designer is invited to instantiate the profile model using specific interfaces offered by GIWA that treat respectively by following axes:

Figure 6. Architecture of GIWA

Figure 7. The semantic level of GIWA



- Personal information about the user like name, age, sex, language.
- Users' presentation preferences defined in by two charters called composition and graphical charters (defined in the model of media adaptation (cf. Figure 4)).

After instantiation, the profile model is exported in XSL files in the data repository❺. At this stage GIWA treat the aspect of adaptability which appears through the choice of a graphical charter and through the composition of page after generated users' application. Figure 8 presents some interfaces of GIWA to instantiated the functionality model, the audience model and the profile model.

Conceptual Level

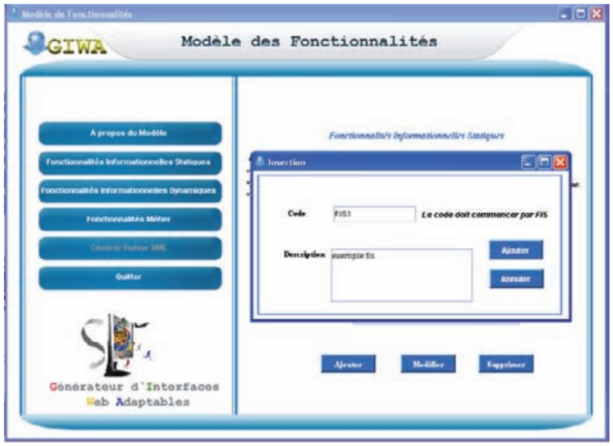
In the Conceptual level, the functional space for each audience class is represented using different conceptual modeling (use case diagram, sequence diagram, class diagram, etc) which are represented in WA-UML (Ben Djemaa 2006c; Djemaa, 2008). While this article is not solely devoted to the conceptual issue, we do want to shortly demonstrate how the functional space for each user, is modeling. In fact, diagram and meta-model of WA-UML are detailed in our work in (Ben Djemaa, 2008).

In this article we propose an AGL which supports the new design elements that we proposed in WA-UML. The last is based on Argo_UML because it permits to guide the user in the use of the UML notation through a mechanism of critiques and help messages. In addition, the source code of Argo_UML is available on the Web making it possible to analyze its inner workings. Thus, in this extension of Argo_UML, we introduce new types of diagrams to represent the different diagrams of our extension. In Figure 9 we present the different steps of the conceptual level in GIWA.

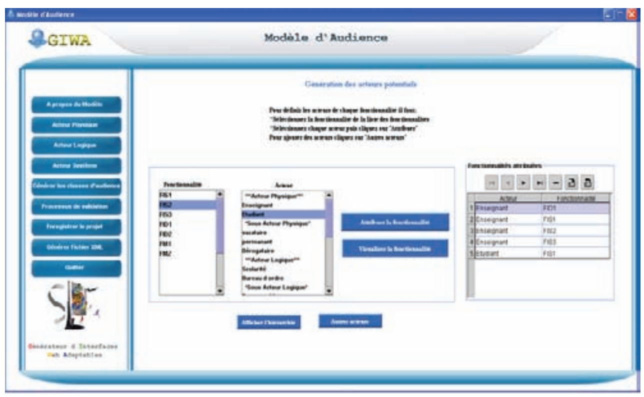
The XML files generated by the semantic level are extracted from the data repository ❻ to be imported into the AGL supporting the new design elements that we proposed.

This AGL is based on Argo_UML. So, in our use of Argo_UML, we introduce new types of diagrams to represent the new diagrams of our extension WA-UML (Ben Djemaa, 2008). In these diagrams, the user can add, displace and copy the different design elements as well as replace faces

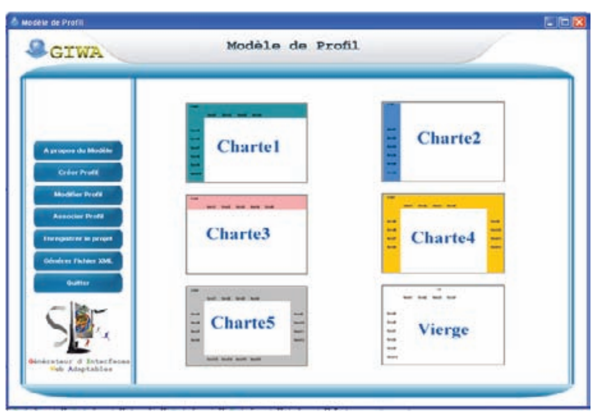
Figure 8. Interfaces of the semantic level in GIWA



The instantiation of the functionality model



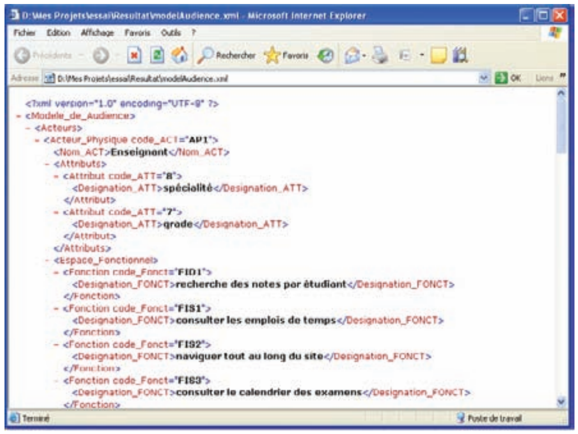
The application of the PVMA



Composition charter of GIWA

continued on following page

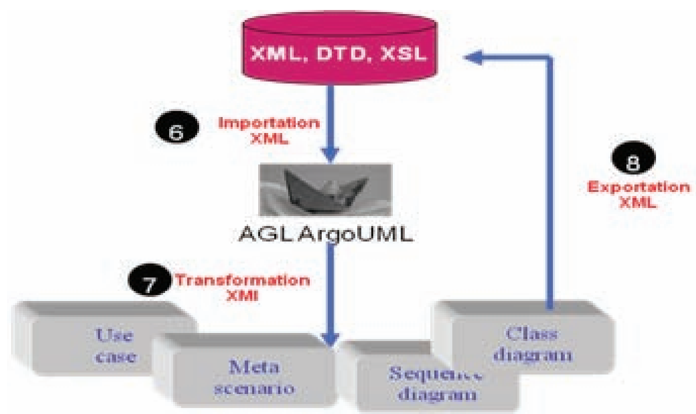
Figure 8. continued



```
<?xml version="1.0" encoding="UTF-8" ?>
- <Modele_de_Audience>
  - <Acteurs>
    - <Acteur_Physique code_ACT="AP1">
        <Nom_ACT>Enseignont</Nom_ACT>
      - <Attributs>
        - <Attribut code_ATT="8">
            <Designation_ATT>spécialité</Designation_ATT>
          </Attribut>
        - <Attribut code_ATT="7">
            <Designation_ATT>grade</Designation_ATT>
          </Attribut>
        </Attributs>
      - <Espace_Fonctionnel>
        - <Fonction code_Fonct="FID1">
            <Designation_FONCT>recherche des notes por étudiant</Designation_FONCT>
          </Fonction>
        - <Fonction code_Fonct="FIS1">
            <Designation_FONCT>consulter les emplois de temps</Designation_FONCT>
          </Fonction>
        - <Fonction code_Fonct="FIS2">
            <Designation_FONCT>naviguer tout au long du site</Designation_FONCT>
          </Fonction>
        - <Fonction code_Fonct="FIS3">
            <Designation_FONCT>consulter le calendrier des examens</Designation_FONCT>
          </Fonction>
```

XML file of the audience model

Figure 9. Conceptual level of GIWA



and publish their properties as used in Argo_UML. Figure 10 shows some examples of new icons of WA-UML presented by Argo_UML.

All conceptual diagram of WA-UML can be described with the new AGL based on Argo_UML❼. The last diagram is translated in XML files ❽.

In Figure 11 we present some diagrams of GIWA at the conceptual level. Us an example we present conceptual diagrams related to E-commerce application.

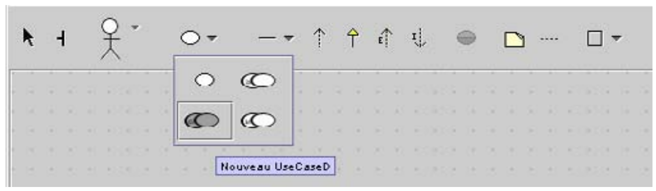
Generation Level

The previous sections dealt with the engineering process of GIWA. This section focuses on the process of generation of adaptive Web applications and describes how the system is dynamically adjusted to varying audience classes.

The generation level focuses on the process of Web page generation and describes how the generator GIWA dynamically adjusts to varying user

Figure 10. New icons in WA-UML

(a) New actors in WA-UML



(b) New use cases in WA-UML

(c) New classes in WA-UML

preferences into chosen implementation platform (HTML, WML, SMIL, etc.). The target of this step is to facilitate the automatic generation of adaptive Web interfaces. It should be possible to program the Web applications in such a way that it can automatically execute the process specified by the design. The tool is based on a collection of engines, which interpret the models provided by the designer during the generation process. (cf. Figure 12).

According to the user/devices profile (refers aspect of adaptivity) is captured by the GIWA using data from logs files to be stored on the server according to a RDF vocabulary (W3C, 2002) ❾ and then to instantiate the profile model (user/devices profile) by specific capabilities (e.g. bandwidth, display resolution,…). Finally, XML, RDF and XSL files are extracted from the data repository ❿ and they are sent to the PARSER in order to apply some adaptation rules for each media (text, image, sound and video) and finally to publish the HTML page corresponding to the devices user (PC, PDA, cell phone or desktop browse) ⓫.

In adaptation media rules (Abdallah, 2008), we use the media properties defined in profile model, e.g. we can test if the dimensions of a particular image fit the size of the screen. Note that the adaptation based on user preferences can be treated in the same way as the adaptation based on device capabilities.

We proposed several adaptation rules for different Medias (text, image, sound and video).

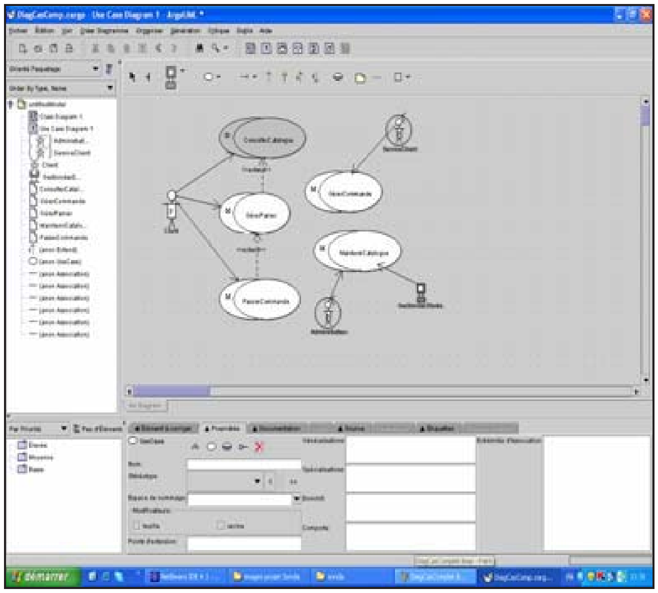
R1 is an example of an adaptation rule for text media. This rule dimensions the size of the screen to adapt the size of the text to the hardware device or to user who is partially-sighted person or blind person. In appendix 1 we present the rest of adaptation rules for text media.

R1: ClientPage.Media.Texte.size_caracter >

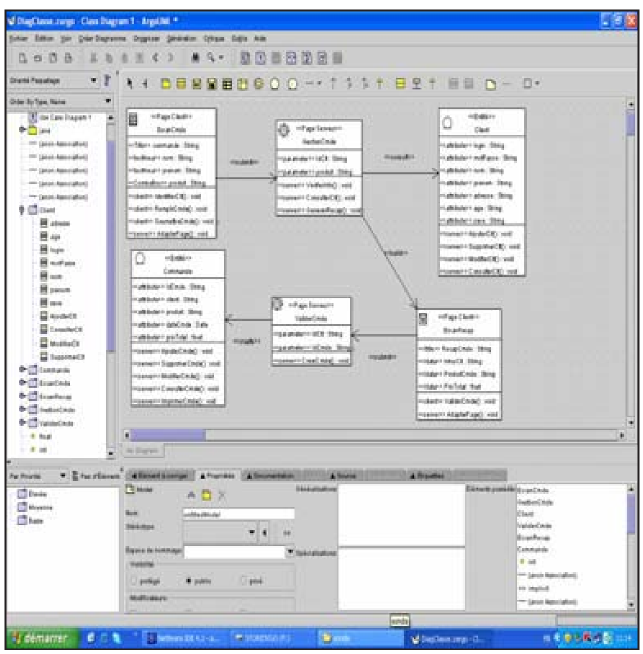
$$\sqrt{\frac{height\_screen * width\_screen}{length\_text}})$$

$\wedge$ ($\neg$ UserProfile. partially-sighted person

Figure 11. Examples of WA-UML diagrams



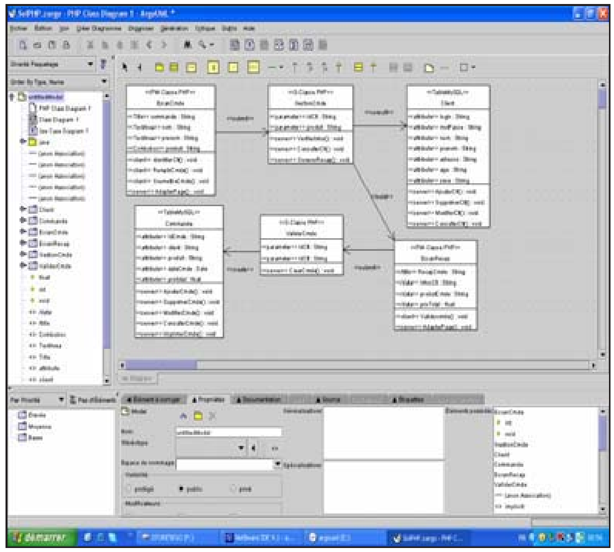
An example of WA-UML use case diagram



An example of WA-UML class diagram

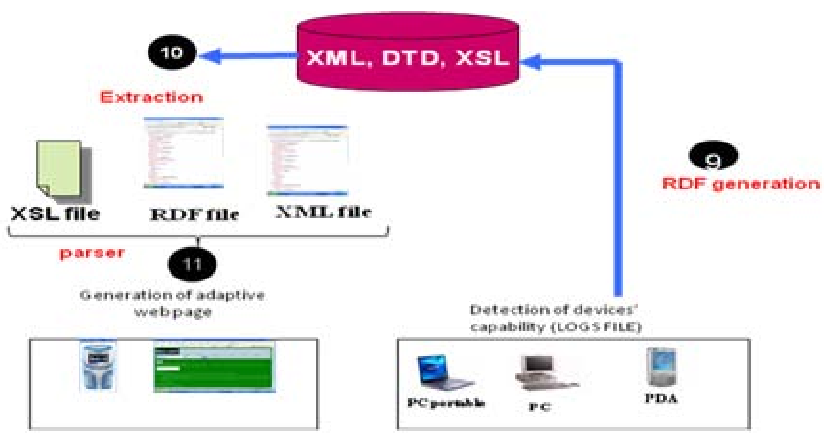
continued on following page

Figure 11. continued



An example of WA-UML PHP class diagram

Figure 12. Generation level of GIWA



$\vee \neg$ UserProfile. blind person)

$\wedge$(UserProfile.Up_Context.Up_Context_Environmental.DeviceCaracteristic.Up_Hardware.support_text))

$\rightarrow t_1$:Resize_text(ClientPage.Media.Text.size_caracter, E(

$$\sqrt{\frac{height\_screen * width\_screen}{length\_text}}))$$

R 2 presents an example of adaptation rules for an image media. This rule adapts the size of the image to be displayed by the device. In appendix 2 we present the rest of adaptation rules for image media.

R2: ClientPage.Media.image.Hauteur > UserProfile.Up_Context.Up_Hardware.hauteur $\vee$ ClientPage.Media.image.largeur > UserProfile.Up_Context.DeviceCharacteristic.Up_Hardware.Largeur →t2 Resize_image(ClientPage.Media.image,UserProfile.Up_Context.Up_Context_Environmental.Up_Hardware.hauteur,UserProfile.Up_Context.Up_Context_Environmental.DeviceCharacteristic .Up_Hardware.largeur)

R 3 presents an example of adaptation rules for sound media. This rule deletes the sound in the Web page where the device doesn't support it or where the user is a deaf person. In appendix 3 we present the rest of adaptation rules for sound media.

R3: UserProfile.deaf ∨ ¬ UserProfile.Up_Context.Up_Context_Environmental.DeviceCharacteristic .Up_Hardware.support_sound ∨ ClientPage.Media.audio.sound.jingle →t3 delete_sound(ClientPage.Media.audio).

R 4 presents an example of adaptation rules for video media. This rule is applied to substitute a video by an image if the device is enabling to display this video. In appendix 4 we present the rest of adaptation rules for video media.

r18:¬UserProfile.Up_Context.Up_Context_Environmental.DeviceCharacteristic .Up_Hardware.support_Video ∧UserProfile.Up_Context.Up_Context_Environmental.Up_Hardware.support_image →t18Convert_video_to_image(ClientPage.Media.video,UserProfile.Up_preference.Up_presentation.Charte.Modalite.Visual.graphic.format,number_image:1 default)

In Figure 13 we present the simulation of GIWA for a user who connected with a PC.

In Figure 14 we present other interfaces of GIWA related to user who connected to the system using cell phone (Nokia SDK 5100).

EXPERIMENTATION AND EVALUATION

Evaluating systems is a difficult task, and it

Figure 13. Simulation of GIWA using PC

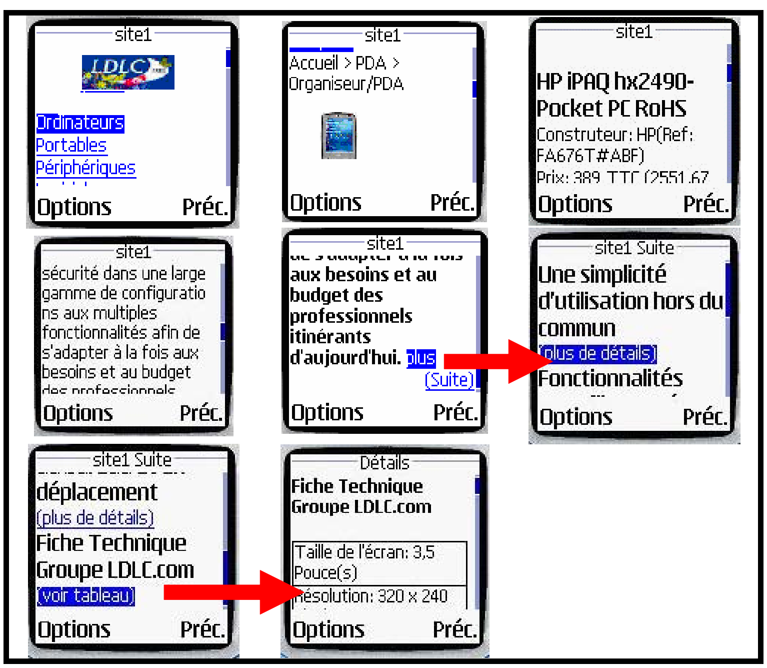
becomes even more difficult when the system is adaptive. It is of crucial importance to be able to distinguish the adaptive features of the system from the general usability of the designed tool.

The evaluation of our GIWA system was designed to address the goals of the system, namely to help users to design and generate adaptive web applications. So firstly, we wanted to measure the usability of GIWA and testing the ability of the process, during its three levels (i.e. semantic, conceptual and generation), to design and generate adaptive interfaces. Secondly, we were also interested in subjects' own evaluation of how well the adaptive system worked compared to the non-adaptive one, and whether they felt in control of the adaptive parts of the system. A last difficulty in making studies of adaptive systems is in the procedure of the study.

Design of the Experiment

The study was done in the laboratory of MIRACL. Subjects were videotaped, and the computer screen was recorded on the same videotape. The subjects' actions were tracked using DRUM (Diagnostic Recorder for Usability Measurement) and descriptive statistics of task completion time, actions performed inefficient use of the system, and others could be easily computed using this tool. There were 8 subjects in the study, 3 female and 5 male, all employed at the laboratory that had developed and used the target domain SDP. All had experience with World Wide Web (WWW) and hypermedia development. Subjects spent in total, approximately three hours in carrying out the steps of the experiment, out of which two hour was spent testing the three level of GIWA. The rest of the time was used for questions on the subject's background, a small diagnostic test of their understanding of certain concepts in the on-line manual before using our system and after using our system, and finally answering some questions about their preferences regarding GIWA and the adaptive versus the non-adaptive

Figure 14. Simulation of GIWA using cell phone (Nokia SDK 5100)



system. Details of the steps of the experiment are presented below.

Each of the subjects followed three steps in this experiment, related to each level of GIWA. We proposed for them an example of web application related to the "teaching gate" of our institute. The two first steps related to the semantic and the conceptual level that were designed to test the explanations provided by the system rather than test the usefulness of the system as such. These two steps also served as a means to introduce subjects to the development of Web applications. The third step concerned the generation level that was realized for both of the two devices: the PC and the cell phone.

In the first step, a subject was given a set of functionalities related to the "teaching gate" with different types (FIS, FID, and FM), and was asked to use the generator to complete the following three tasks:

- **Task 1:** Instantiate the functionality model by adding, for each functionality, the code and the description, and generate the XML file for this model.
- **Task 2:** Apply the AGAC (Algorithm of Generation of Audience Classes).
- **Task 3:** Apply the PVMA and generate the XML file of the audience model.

In the second step the subject was given the XML file generated by the first step and asked to use the ArgoWA-UML to complete the following three tasks:

- **Task 4:** Import the XMI File related to the semantic level and created the use cases of the application.
- **Task 5:** create the sequence and classes diagrams with the new icons of WA-UML.
- **Task 6:** choose a Web language (PHP, JSP, ASP) and generated the technical class diagram.

In the third of the three above mentioned steps the subject was given the XML file generated by the second step and asked to use firstly a Pc and secondly, a cell phone, to complete the following three tasks:

- **Task 7:** connected to GIWA by taping: http://localhost:8080/GIWA/index.html (for PC) http://localhost:8080/GIWA/index.wml (for cell phone).
- **Task 8:** create an account and then choose a profile or create a new profile.
- **Task 9:** choose the XML file related to the conceptual model of the application "teaching gate" and generated the results by PC and cell phone.

The design goal of these nine tasks (1 to 9) was to oblige the users to actively use all the commands provided, while at the same time expose their own patterns of command sequences when performing the various tasks.

Results of Carrying out the Experiment

Usability of the System

Our results are divided into those concerning:

- Some remarks concerning the task completion time, where we can see a weak tendency that users spend some time to achieve the first task.
- The actions in ArgoWA-UML that the subjects have to do (clicking on icons, making menu-choices, clicking on association, etc) to build WA-UML conceptual diagrams are less than the action of building UML diagrams. Thus, this confirms the result that conceptual level of GIWA is a semi automatic one.
- Adaptive Web pages generated by cell

telephone are more interesting than those generated by a PC for most of users.
- The subjects' satisfaction with the system, where they compare the result (page generated) and their preferences described at the semantic level.
- The passage between the three levels of GIWA was not too clear by some of users.
- Combination of design and generation in GIWA is approved by users.

User Satisfaction

After the subjects had used the two variants of GIWA, we asked them to provide their viewpoints on various aspects of the generator. We did this through ten questions, and they were also asked to freely comment on various aspects of the system. For the ten questions the subjects put a cross on a scale grading from 1 to 7 - the interpretation of the scales can be seen from the statements left and right of the Table 1.

In the evaluation column of the Table 1 we present the interpretation of the scales which can be seen from the statements left and right. The x-axes represents the number of user for each scale, the y-axes represents the different interpretation of user for each question.

In Table 1 we see the result of the queries on how the subjects perceived the adaptive system in using GIWA. As we can see, the subjects preferred the adaptive system (mean 5.25); they also like the combination of design and generation in GIWA (mean 5.25); and they felt that the system made good adaptations to their needs (mean 4.75). Also, they claimed that they saw when the system changed the inferred task (mean 4.75).

Evaluating adaptive systems is often done through comparing a non-adaptive version of the system to an adaptive system. Our study is no exception from this approach. Still, an adaptive system should preferably be designed in such a way that the adaptivity is only one instrument in the repertoire of design techniques that together will form the tool that in its whole meets the users' needs and individual differences.

There are few studies of adaptive systems in general and even fewer of adaptive hypermedia. In the studies of adaptive hypermedia by (Boyle 1994; Kaplan 1993) the main evaluation criterion is task completion time. This should obviously be one important criterion by which some systems should be evaluated. In our case, though, the goal of the adaptive hypermedia system is to generate the right needs according to the user's preferences. The time spent in retrieving information is not as relevant as is the quality of the search and the result. Apart from task completion time, Boyle and al. also measured reading comprehension through a diagnostic test put to the subjects after having used their system. Kaplan and al. measured how many nodes the users visited - in their case the more nodes the users visited the better.

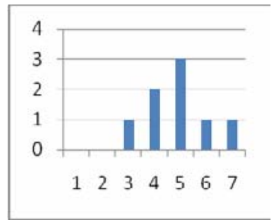
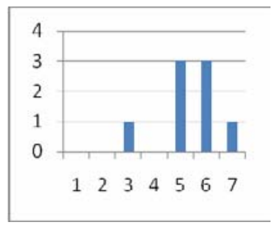
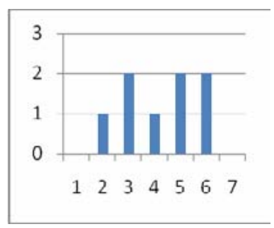
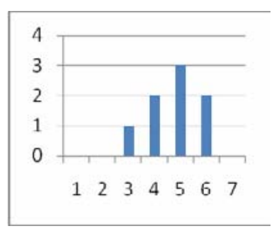
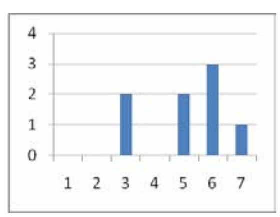
CONCLUSION

The research described in this article targets the support of automated generation Web interfaces in the context of adaptive Web application. Specifically, for Web applications involving various functionalities (informational or professional one), the implementation requires a structured approach.

This article develops a generator called GIWA supporting both design aspect and generation one. The architecture of GIWA is based on three levels: the semantic level, the conceptual level and the generation level.

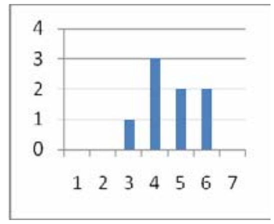
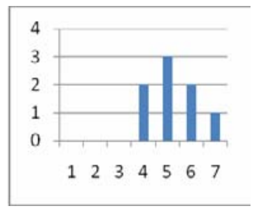
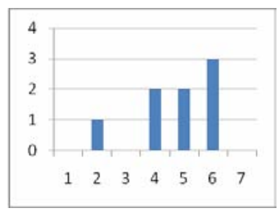
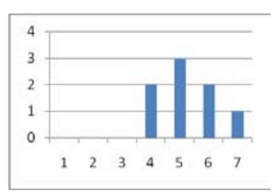
The primary focus of the GIWA is to provide engineering support for adaptive Web applications that automatically generate hypermedia presentations in response for each ad hoc user's requirements. GIWA guides the designer through the different steps of the generation process, each of them yielding a specific model that is being interpreted by the GIWA tools to achieve the objective of automatic presentation generation.

Table 1.

NUM	Questions	Means	Evaluation		
1	How efficiently would you be able to Work with GIWA?	4,87	Badly, the generator gets in the way		Good, the work would be very efficient
2	Did you like using GIWA?	5,37	No, it is very demanding and unpleasant to use.		Yes, I really liked using it.
3	Do you feel in control while using GIWA?	4,25	No, it feels as if the generator controls me.		Yes, I can make the generator do what I want.
4	Did you easily get lost in the information space?	4,75	I got lost several times and did not know where I was.		I knew all along exactly where I was.
5	Did you find it easy to get started?	5,12	No, in the beginning it was very difficult.		Yes, it is possible to get started right away.

continued on following page

Table 1. Continued

6	Are the different Steps of GIWA easy to understand and use?	4,62	No, it is difficult to find the right icon and use it.		Yes, they are easily understood.
7	Did you like the combination of design and generation in GIWA?	5,25	No, there are too many details and it is confusing.		Yes, the interface of GIWA is very appealing.
8	Did you see when the adaptations happened in GIWA?	4,75	No, I never saw that the system changed.		Yes, it was obvious when the generator changed task and opened new operation.
9	Did the adaptive system make good adaptations to your needs?	4,75	No, I repeatedly had to change the answers I got in order to find the right information.		Yes, it managed to get relevant information.
10	Did you prefer the adaptive or the nodaptive system?	5,25	The noadaptive Was definitely better.		The adaptive Was definitely better.

In GIWA, we have distinguished two kinds of adaptation: adaptability (implemented at the semantic level) and adaptivity (implemented at the generation level). Adaptability is based on information about user preferences presentation (a.g. font color, page layout etc.) and user preferences navigation stored in the profile model before browsing starts. Adaptivity is considered in GIWA to provide a system which is able to automatically adapt a given presentation to the user device capabilities (hardware and software configuration). Information about device capabilities are captured from Logs Files and stored then in the profile model. The prototype of GIWA has been built by different engines and equipments. The prototype uses java interfaces to instantiated models at the semantic level. At the second level GIWA uses Argo_UML to create conceptual diagrams which are translate in XML files to be generated in Web pages at the generation level. In the last, XML and RDF files are used to store the data and XSL files to specify transformations between consequent steps.

We have presented an experiment to evaluate our generator GIWA. Results showed significant performance gains, both in design and in generation step. While our experiment provided several revealing results, there are many issues regarding adaptive web generation worthy of further study. It would be interesting to experiment and evaluate GIWA with blind and sighted users. Also interesting is how the framework generalizes between various users.

The work presented in this article provides different opportunities for future research. The future work includes further developments to facilitate adaptation at all levels of the generation process. Firstly we plan the extension of the ArgoWA-UML tool that is currently built to support the semi automatic transition from design models to a running implementation of adaptive Web applications. Secondly we plan to extend adaptivity in GIWA in order to dynamically elaborate and modify both the functional space and the navigation patterns, learning from the user's behaviour.

ACKNOWLEDGMENT

The authors would like to thank Ben Abdallah Fatma, who helped in the design and the development of a part of media adaptation.

REFERENCES

Abdallah, F., Amous, I., Ben Djemaa, R., & Ben Hamadou, A. (2008). Génération d'Interfaces Web Adaptable au Contexte et aux Préférences des Utilisateurs. Les huitième journées scientifiques des jeunes chercheurs en Génie Electrique et Informatique (GEI'2008) (pp. 103-112), Sousse, Tunisie.

Ben Djemaa, R., Amous I., & Ben Hamadou, A. (2005, September). Towards an approach for adaptive Web applications' requirements engineering. The 7th International conference on Information integration and Web-based Application and Services (IIWAS05) (pp. 19-21). Kuala Lumpur.

Ben Djemaa, R., Amous I., & Ben Hamadou A. (2006a, May). Design and implementation of adaptive Web application. *In 8th International Conference on Enterprise Information Systems ICEIS'06 (pp. 23- 27), Paphos Cyprus.*

Ben Djemaa, R., Amous I., & Ben Hamadou A. (2006b, February). GIWA: A generator for adaptive Web applications. In *Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services (AICT/ICIW 2006) IEEE (pp. 23-25), Guadeloupe, French Carabeen*.

Ben Djemaa, R., Amous I., & Ben Hamadou A. (2006c). Use case and meta-scenarios for modelling Adaptive Web Applications. In *proceeding of IEEE 1st International Conference on Digital Information Management (ICDIM).* 06-08 December, 2006. Christ College, Bangalore, India, pp. 283-288.

Ben Djemaa, R., Amous I., & Ben Hamadou A. (2007, May). Adaptable and adaptive Web applications: From design to generation. *International Review on Computers and Software (IRECOS), 2(3), 198-207.*

Ben Djemaa, R., Amous I., & Ben Hamadou, A. (2008, April/June). Extending a conceptual modeling language for adaptive Web applications. *International Journal of Intelligent Information Technologies (IJIIT), 4*(2), 37-56.

Boyle, C., & Antonio, O. (1994). MetaDoc: *An Adaptive Hypertext Reading System*. *User Models and User Adapted Interaction*, (UMUAI), 4, 1-19.

Brusilovsky, P. (1998). Methods and Techniques of Adaptive Hypermedia. In P. Brusilovsky, A. Kobsa, & J. Vassileva (Eds.), *Adaptive Hypertext and Hypermedia* (pp. 1-43). Kluwer Academic Publishers.

Cingil, I., Dogac, A., & Azgin, A. (2000). A broader approach to personalization. Communication of the ACM, 43(8), 136-141.

Fiala, Z., Hinz,, M., Meissner K., & Wehneer, F. (2003). A component-based component architecture for adaptive dynamic Web documents. *Journal of Web engineering, 2(*1), 58-73.

Fons, J., Pelechano, V., Pastor, O., Albert, M., & Valderas, P. (2003). *Extending an OO Method to Develop Web Applications*. The Twelfth International World Wide Web Conference (pp. 20-24), Budapest, Hungary.

Frasincar, F., & Houben, G.-J. (2002). Hypermedia presentation adaptation on the semantic Web. In P. de Bra, P. Brusilovsky, & R. Conejo (Eds.), *Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2002)* (pp. 133-142), Malaga, Spain.

Frasincar, F., Houben, G.-J., & Vdovjak, R. *(2001).* An RMM-Based Methodology for Hypermedia Presentation Design. *In Proceedings of the 5th East European Conference on Advances in Databases and Information Systems (ADBIS 2001), LNCS 2151 (pp.* 323-337), Vilnius, Lithuania.

Isakowitz, T., Stohr, A., & Balasubramanian, E. (1995). RMM: A methodology for structured hypermedia design. *Communications of the ACM, 38*(8), 34-44.

Isakowitz, T., Kamis, A., & Koufaris, M. (1998). *The Extended RMM Methodology for Web Publishing.* Center for Research on Information Systems.

Kapla, C., Justine, F., & James, C. (1993). Adaptive Hypertext Navigation Based On User Goals and Context. *User Modeling and User-Adapted Interaction, 3,* 193-220.

Koch, N. (2000). Software Engineering for Adaptative Hypermedia Systems-Reference Model, Modelling Techniques and Development Process. Ph.D Thesis, Fakultät der Mathematik und Informatik, Ludwig-Maximilians Universität München.

Koch, N. (2001). The Authoring Process of the UMLbased Web Engineering Approach. 1*st International Workshop on Web-Oriented Software Technology. The Tenth International Conference on the World Wide Web,* Hong Kong.

Rossi, G., Schwabe, D., & Guimaraes, R. (2001) Designing Personalized Web applications. *The Tenth International Conference on the World Wide Web, Hong Kong.*

Schwabe, D., & Rossi, G. (1998*). D*eveloping hypermedia applications using OOHDM. *In Proceedings of Workshop on Hypermedia development Process, Methods and Models* (pp. 85-94).

Villanova-Oliver, M., Gensel, J., Martin, H., & Erb, C. (2002). *Design and generation of adaptable Web information systems with KIWIS*. In Proceedings of the 3rd IEEE Conference on Information Technology TCC-2002, Las Vegas, USA World Wide Web Consortium.

Argo UML. http://www.tigris.org. eXtensible Stylesheet Language, XSL Specification – version 1.0, W3C Recommendation. Retrieved October 15, 2001 from http://www.w3.org/TR/xsl/ World Wide Web Consortium, 2002.

RDF Vocabulary Description Language 1.0: RDF Schema, W3C Working Draft. Retrieved April 30, 2002 from http://www.w3.org/TR/rdf-schemadix.

APPENDIX

Appendix A: Adaptation rules for Text Media

R5: (ClientPage.Media.Text.color $\not\subset$ UserProfile.Up_preference.Up_presentation.Charte.Modalite. Visual.colors) $\vee$ (ClientPage.Media.Text.color$\notin$UserProfile.Up_Context.Up_Context_Environmental. DeviceCharacteristic.Up_Hardware.palette) → t5: Modify_color(ClientPage.Media.Text.color, color)

R6 : ClientPage.Media.Text.police≠UserProfile.Up_preference.Up_presentation.Charte.Modalite. Visual.Text.police $\wedge$ (UserProfile.partially-sighted person $\vee$ UserProfile.blind_man $\vee$ UserProfile. Up_Context.Up_Context_Environmental.DeviceCharacteristic.Up_Hardware.support_Text$\wedge$UserProfile. Up_Context.Up_Context_Environmental.DeviceCharacteristic.Up_Hardware.support_image) →t6 : Convert_text_to_image(ClientPage.Media.Text)

Appendix B: Adaptation rules for Image Media

R7: (ClientPage.Media.image.color $\not\subset$ UserProfile.Up_preference.Up_presentation.Charte.Modalite. Visual.colors) $\vee$ (ClientPage.Media.image.color $\notin$ UserProfile.Up_Context.Up_Context_Environmental. DeviceCharacteristic . Up_Hardware.palette)→t7 : color_image(ClientPage.Media.image.color, color)

R8: UserProfile.Up_Context.Up_Context_Environmental.Protocole_de_transfert="HTTP" $\wedge$ ClientPage.Media.image.size_File > UserProfile.Up_Context.Up_Context_Environmental.DeviceCharacteristic.Up_Hardware.memory → t8 $\wedge/\vee$ t_{10} $\vee/\wedge$ t_{13} convert_gif (ClientPage.Media.image)

R9: $\neg$ UserProfile.Up_Context.Up_Context_Environmental.DeviceCharacteristic .Up_Hardware. support_image $\wedge$ $\neg$ UserProfile.Up_Context.Up_Context_Environmental.DeviceCharacteristic .Up_ Hardware.support_Text $\wedge$ UserProfile.blind_man →t9 Delete (ClientPage.Media.image

Appendix C: Adaptation rules of Sound Media

R10: UserProfile.mal_entendant $\vee$ $\neg$ UserProfile.Up_Context.Up_Context_Environmental.DeviceCharacteristic .Up_Hardware.support_sound $\vee$ $\exists$ClientPage.Media.audio.sound.jingle →t10 Delete_sound(ClientPage.Media.audio).

R11: ClientPage.Media.audio.size_File > UserProfile.Up_Context.Up_Context_Environmental. DeviceCharacteristic .Up_Hardware.memory →t11 Stereo_to_mono(ClientPage.Media..audio) $\wedge$ / $\vee$ Reduct_sampling(ClientPage.Media..audio) $\wedge$ / $\vee$ Convert_sound(ClientPage.Media..audio, UserProfile. Up_preference. Up_presentation. Charte .Modalite.auditive.format)

Appendix D: Adaptation rules of Video Media

R12: $\neg$ UserProfile.Up_Context.Up_Context_Environmental.DeviceCharacteristic .Up_Hardware. support_Video$\wedge\neg$UserProfile.Up_Context.Up_Context_Environmental.DeviceCharacteristic .Up_Hardware.support_image $\wedge\neg$ UserProfile.Up_Context.Up_Context_Environmental.DeviceCharacteristic. Up_Hardware.support_sound →t12 Delete_video(ClientPage.Media.video)

This work was previously published in International Journal of Information Technology and Web Engineering, Vol. 4, Issue 2, edited byG. I. Alkhatib and D. C. Rine, pp. 20-44, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 3.3
Migrating Web Services in Mobile and Wireless Environments

Myung-Woo Park
Yonsei University, South Korea

Yeon-Seok Kim
Yonsei University, South Korea

Kyong-Ho Lee
Yonsei University, South Korea

ABSTRACT

Mobile devices enabled with Web services are being considered as equal participants of the Web services environment. The frequent mobility of devices and the intermittent disconnection of wireless network require migrating or replicating Web services onto adjacent devices appropriately. This article proposes an efficient method for migrating and replicating Web services among mobile devices through code splitting. Specifically, the proposed method splits the source code of a Web service into subcodes based on users' preferences for its constituent operations. The subcode with a higher preference is migrated earlier than others. The proposed method also replicates a Web service to other devices to enhance its performance by considering context information such as network traffic or the parameter size of its operations. To evaluate the performance of the proposed method, the effect of the code splitting on migration was analyzed. Furthermore, to show the feasibility of the proposed migration method, three application scenarios were devised and implemented.

INTRODUCTION

Web services (Huhns & Singh 2005; Stal, 2006), which are independent from operating systems and programming languages, have gained momentum as an enabling technology to realize business processes on distributed network environments such as the Web. Additionally, the technology of mobile devices is continually developing and thus allows for a new form of Web services, that is, mobile Web services (Schall, Aiello, & Dustdar, 2006; Sirirama, Jarke, & Prinz, 2006). However, it is difficult to provide Web services on mobile devices seamlessly, since wireless and mobile environments still involve unstable connectivity, unlike the typical client-server environment.

If Web services autonomously migrate among mobile devices in this unstable wireless environment, seamless provisioning of services would be possible. When a service cannot be provided during movement of a device, it can be migrated to an adjacent mobile device and provide its functionality continuously. Additionally, requests can be distributed by replicating the service to other devices when the requests are concentrated on one device. Moreover, in the case of a client's request for a service that takes large parameters such as bitmap image files as input, the service itself can be replicated and executed on the client side, resulting in saving resources.

Recently, research on Web service migration has been performed. However, most of the research targets desktop and wired environments or does not consider constraints such as low bandwidth of wireless and mobile environments. Therefore, the research approaches might take much longer time to migrate.

To resolve this issue, this article proposes a method for migrating Web services through code splitting. Specifically, an original code, which implements the functionality of a service, is split into subcodes based on users' preferences to its constituent operations. The subcodes of higher preference are migrated earlier to minimize the latency of the operations of high priority and raise the efficiency of Web services migration and replication in wireless and mobile environments. To evaluate the performance of the proposed method, the effect of the code splitting on migration was analyzed. Furthermore, to show the feasibility of the proposed migration method, three application scenarios were devised and implemented.

Meanwhile, how to determine when and where to migrate services is an important issue. The migration of a service may be carried out by the request of a service provider or the change of context information, such as the shortage of battery level and the location change of a device. It involves developing the context model and strategies or policies relevant to the migration of Web services. If a migration of a service is requested, the proposed framework collects the context information of neighboring devices based on the migration policy of the service. It computes the suitability values of candidate devices and determines a target host. To establish a migration policy, the proposed method is based on our previous approach (Kim & Lee, 2007), which is based on WS-Policy (2006). In this article, we do not discuss the context model and migration strategies, but focus on describing the migration method itself.

Meanwhile, the process of identifying when and where to migrate services is also an important issue. The migration of a service may be caused by context changes such as the battery shortage and location change of a device. For the seamless provisioning of a service, we have to determine which device is the most suitable target host. This process involves a mechanism to describe context models and migration strategies, which are relevant to the migration of Web services in mobile environments. For example, a service provider should be able to specify that if the CPU usage-ratio of an origin host is over 80%, a service should be migrated to a new device, which has enough processing power and supports J2ME. For this purpose, we propose a method to establish the context model and migration policy to determine when and where to migrate services in mobile environments (Kim & Lee, 2007). The method determines a target host based on the migration policy of a service as well as the information collected from devices in the neighborhood of the origin host that is hosting the service. In this article, we do not discuss the process of determining when to migrate a service to a specific target host, but focus on how to migrate a service to a target host efficiently.

The organization of this article is as follows. First, a brief survey of previous works is presented. Next, the methods of splitting Web services codes and migrating and replicating them are described. Next, the effect of splitting codes on migration is analyzed through experiments, and three application scenarios are implemented to show the feasibility of the proposed migration method. Finally, we summarize the conclusions and further studies.

RELATED WORK

We summarize the conventional methods about Web service migration and also discuss previous works concerning code mobility, which do not target Web services but are related with migration. Pratitha and Zaslavsky (2004) propose a framework and strategies for migrating Web services. Based on context information such as network bandwidth, their method selects a target server, to which a service should be migrated. Moreover, their method defines service modules in advance according to context information and selects a proper module to the corresponding context.

Hammerschmidt and Linnemann (2005) propose a migration framework, which supports Web services based on Tomcat-Axis. Particularly, since instances are supported, connections do not need to be restarted after migration. Ishikawa, Yoshioka, Tahara, and Honiden (2004) propose the migration of mobile Web services, which are defined as composite Web services. Particularly, agents that execute several services in a composite Web service can migrate to an appropriate host. An endpoint of each service in a composite Web service can be changed; however, the endpoint of a composite Web service does not change. The method of Mifsud and Stanski (2002) monitors context information of available target hosts and migrates Web services to a specific target host.

Meanwhile, Kern, Braun, Fensch, and Rossak (2004) propose a more rapid execution of mobile agents through splitting and arranging codes which implement the functionality of agents. A shortcoming of this method is that the execution order of agent functions in each host must be fixed. The middleware for an ad-hoc network proposed by Bellavista, Corradi, and Magistretti (2004) can receive and execute a list of binary files from neighboring devices. Montanari, Lupu, and Stefanelli (2004) propose a middleware framework which can implement migration policies application independently.

Previous works are mostly based on desktop and wired environments, and therefore the latency of Web services are too long to apply to wireless and mobile environments. Therefore, it is difficult to deal with a service migration efficiently in mobile and wireless environments. In this article, we aim at presenting how to migrate services quickly and to reduce the latency of service requests during migration.

SPLITTING WEB SERVICE CODES

We describe a splitting method of Web service codes. For the splitting and migration method, the framework of Figure 1 is proposed.

The code splitter splits a Web service code and saves them in a code repository in a component form. The migration manager encapsulates components and instances in a transportable form, and unpacks the encapsulated form and saves it at the resource space. The service execution manager manages the execution of a service, which stops and restarts the service execution. The context manager collects context information from the current and adjacent mobile devices and decides when and where to migrate services. The channel manager takes charge of the establishment of network connections. The logger records the execution history of the modules of the framework.

How to split a Web service code by the code splitter in the proposed framework is shown in Figure 2. In our method, it is assumed that the original source codes of Web services are proportionate to their compiled codes.

Figure 1. The proposed framework

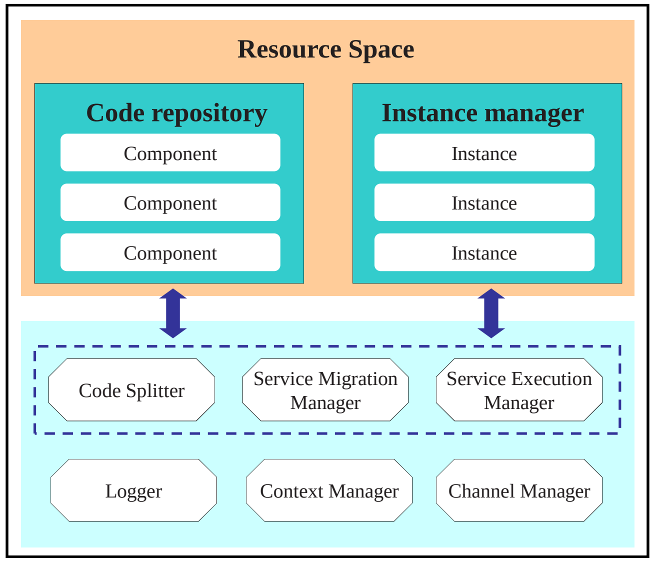
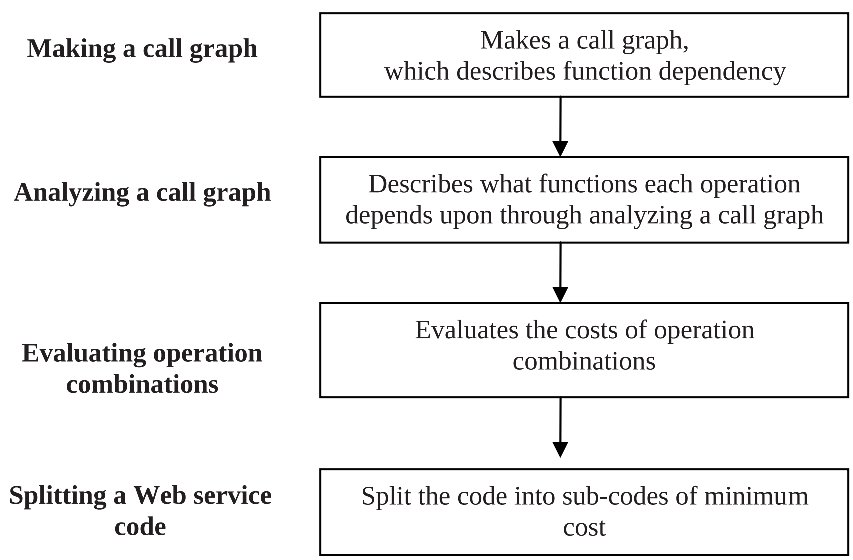


Figure 2. The process of splitting codes



Making and Analyzing a Call Graph

A code that implements a Web service is analyzed and represented with a call graph, and the dependency among functions (or methods) is derived from the call graph. The proposed method for computing the dependency among functions is based on the method of Grove and Chambers (2001). The functions exposed as operations in a WSDL interface are recorded particularly and are used in the next analyzing step. After making the call graph, it is analyzed in order to determine functions, upon which each operation depends.

The analysis progresses following nodes, to which each operation node is connected. Figure 3 illustrates the making and analyzing of a call graph.

Evaluating Operation Combinations and Splitting a Web Service Code

A condition required in code splitting is that each split code should contain at least one operation and exists in a class form so that it can be compiled. The code split must have all the functions upon which its operations depend. However, if operations depend on functions in common, code splitting is impossible or inefficient. Copying appropriate codes of functions can solve this problem. In the call graph of Figure 3, Operations 1, 2, and 3 commonly depend on Function E, and therefore disjoint code splitting is impossible. However, if Function E is copied, it is possible to split the code, as shown in Figure 4.

It is possible but somewhat inefficient to copy all the function codes that are used in common. In Figure 4, Functions B, C, D, and E need to be copied in order to split Operations 2 and 3, and those four functions are almost a half of all the function code, including operations. Therefore, whether functions would be copied should be determined for efficient code splitting. Alternately, an original code is split based on which operations will exist in a class, and then the combination of operations can decide which functions need to be copied.

The selection of the optimal code split is related to both the total size of the codes split and users' preferences to operations. If a code segment has not arrived in a target server, it is impossible to invoke its operation. In this case, users' dissatisfaction with the service would be raised. The proposed method transmits a code segment of higher priority earlier to minimize its service discontinuance. So, the dissatisfaction rate of a code segment at a certain time is proportionate to the time to migrate the total codes that include it and the code segments with a higher priority. The dissatisfaction would be also proportionate to user preferences. User preferences can be statistically calculated from users' previous requests by the logger. In this article, the dissatisfaction with a service migration is formulated and computed by a cost function as follows:

$$Cost = \sum_{i=1}^{n} (p_i \times \sum_{j=1}^{i} S_j) \qquad (1)$$

Figure 3. An example of making and analyzing a call graph

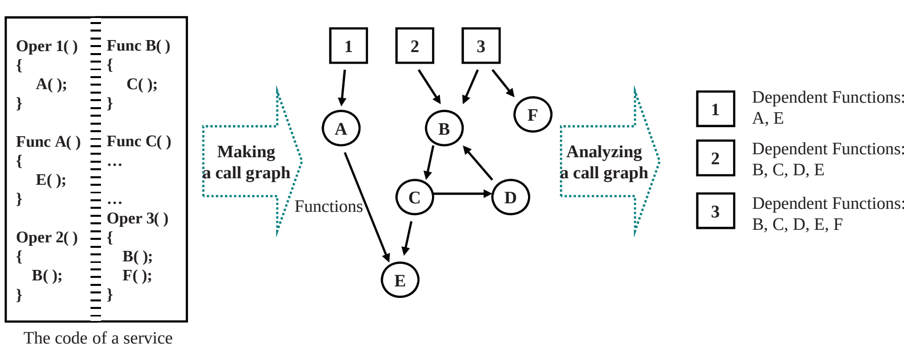
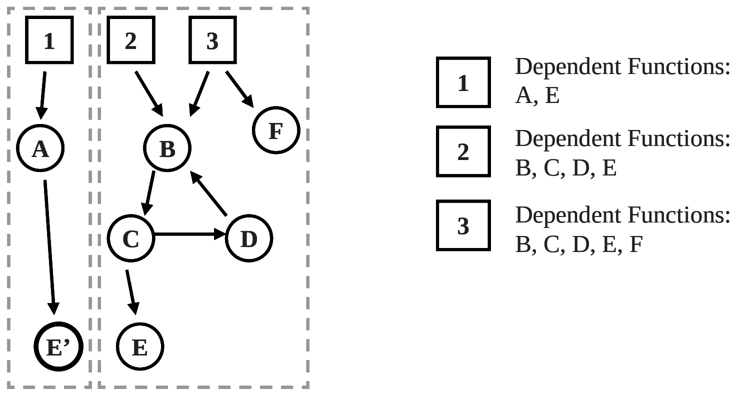


Figure 4. An example of copying functions



where $i < j \rightarrow p_i \geq p_j$, $p_i = p_j \rightarrow S_i \leq S_j$, n is the number of partitions, p_i $(0 \leq p_i \leq 1)$ is the sum of user preferences to operations in the ith partition, and S_j $(0 \leq S_j \leq 1)$ is the total code size of jth partition.

The proposed method computes the costs of every possible combination of code split. It finds the combination of minimum cost that corresponds to the optimal split of operations. The original service code is split according to the optimal combination. Each partition of operations and their functions is compiled and saved as a component in the code repository.

MIGRATING AND REPLICATING WEB SERVICES

The proposed method handles both migration and replication. The former occurs when an origin server cannot provide a service any more because of some reasons (e.g., battery shortage and a non-service area). The replication of a service is done when the service needs to be copied to a different server temporarily to maintain its quality. Example cases are as follows: 1) when network bandwidth becomes more crowded and traffic slows down; 2) when the number of requests from a particular location grows rapidly; and 3) when a client wants to use a service with large parameters.

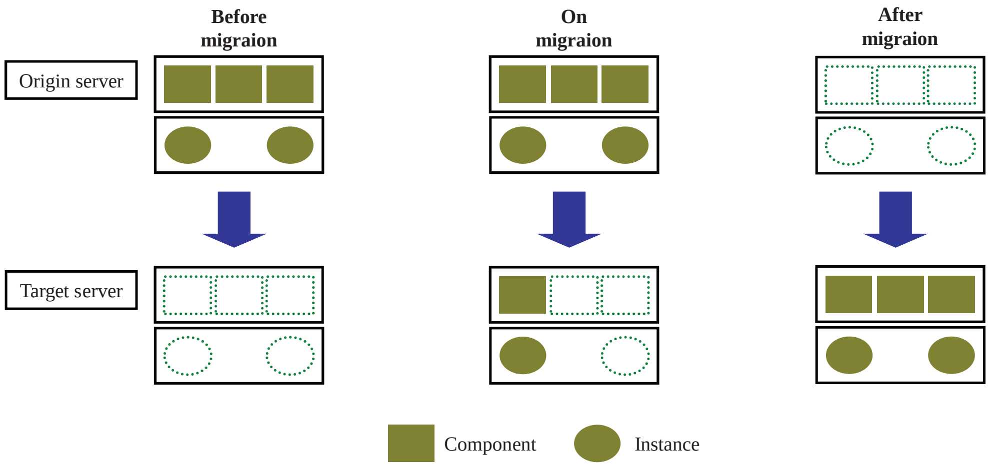
When the migration of a service occurs, its endpoint should be changed as the corresponding endpoint of the target host. In the case of replication, a new endpoint is added to the existing list of endpoints of an origin server. The context manager selects between migration and replication depending on context information.

Migrating Web Services

When the battery level of an origin server becomes very low or a service cannot be provided any more due to the location change of its device into a nonservice area, the service should be migrated to an appropriate target device. If a target server accepts the migration request, the components and instances in a transportable form are transmitted, as shown in Figure 5. Specifically, existing requests on the execution queue of the service execution manager are processed on the server, selected between the origin server and the target server for faster execution. This is basically the same with the case of a service request during migration. At the same time, instances and their associated components on the resource space are transmitted in order of the user preference.

Whenever a component and/or its instance arrive, the service execution manager of the target server deploys the operations of the component.

Figure 5. An example procedure of Web services migration



As soon as all the components and instances have arrived at the target server, they are deleted on the origin server. If the components and instances of a service have not arrived at the target server due to any reason, the process of migration halts. Additionally, the origin server searches again for a new target server. The former target server maintains the components and instances already arrived for a certain period of time. After the time period, the target server deletes them.

If a service is requested during its migration, the proposed method determines which server between the origin and target servers executes the request faster. Figure 6 describes the control flow of the process. In the case where the component of the operation called has already transmitted to the target server, t_o is the time to receive the instance from the target server, execute it on the origin server, respond its result to a client, and transmit the instance updated to the target server, that is, synchronize two instances between the two servers. t'_n includes the time to notify a client that the request should be made to the target server and the time for the client to request the operation to the target server and receive its response. Figure 7(a) describes the case of Figure 6.

In the case where the code has not been transmitted yet, t'_o corresponds to the time required to respond to the client request and transmit the instance and component of the operation to the target server. t''_n includes the time to forward the component and client request to the target server and execute the operation on the target server. It also contains the time for the origin server to notify a client that the result should be received from the target server and for a client to receive the result from the target server. Figure 7(b) illustrates the case of Figure 6.

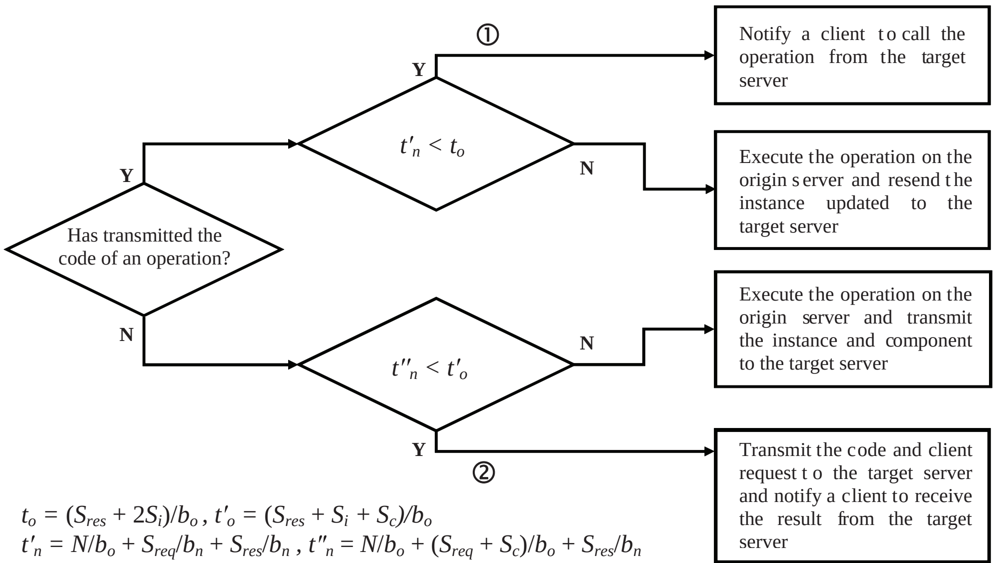
Replicating Web Services

The replication of a Web service is classified into two cases: the replications of a service to another server, and to the client requesting it.

Replication to Another Server

The method of replication is similar to that of migration. If a target server accepts a replication request, the corresponding components and instances are transmitted. However, all the instances

Figure 6. The flow of processing in case of a service call on migration



$t_o = (S_{res} + 2S_i)/b_o$, $t'_o = (S_{res} + S_i + S_c)/b_o$

$t'_n = N/b_o + S_{req}/b_n + S_{res}/b_n$, $t''_n = N/b_o + (S_{req} + S_c)/b_o + S_{res}/b_n$

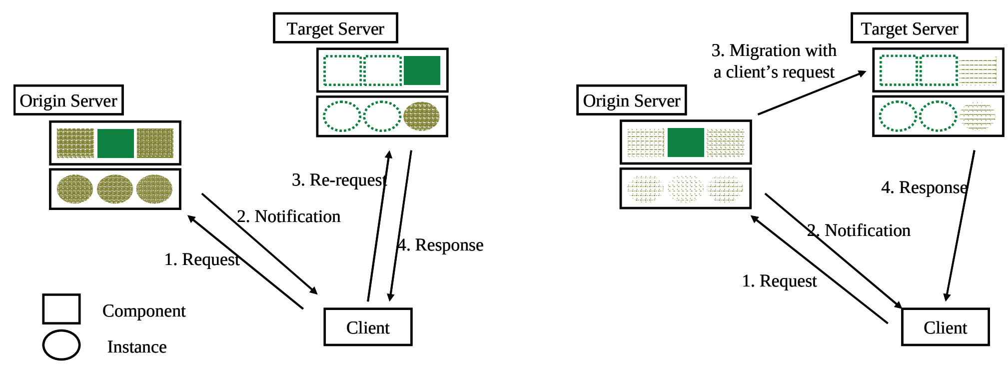
S_{req}: the size of a request message, S_{res}: the size of a result message

S_i: the size of an instance, S_c: the size of a component,

N: the size of a notification, b_o: the bandwidth of an origin server

b_n: the bandwidth of a target server

Figure 7. Service requests during the migration process



(a) A component request after migration *(b) A component request before migration*

do not need to be transmitted, and after finishing the replication process the components are not deleted from its origin server. The instances that have been transmitted to the target server are deleted on the origin server. When a replication occurs, the origin server records the target servers on the list of replications. Additionally, the origin server may request the target server to delete the replicated service. If a service is requested during replication, the same processes with the case of migration are applied.

Figure 8 is an example of a service replication, where an origin server with three components and five instances communicates with five clients. The context manager decides which server should be selected as the target server and which instances

Figure 8. An example of Web services replication

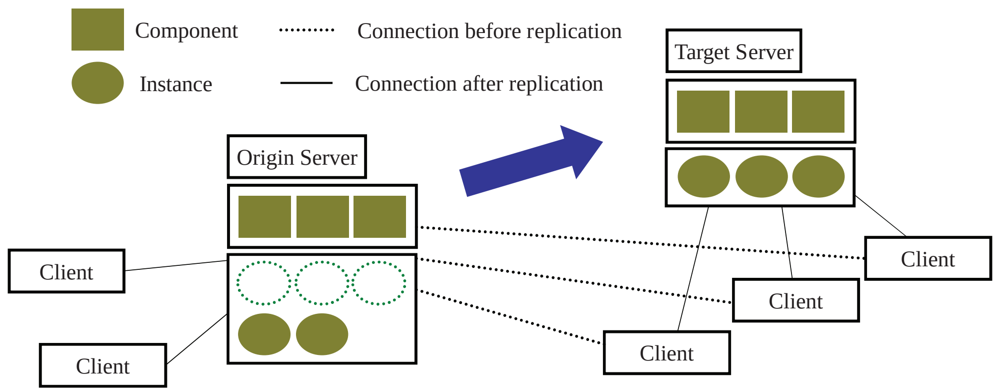
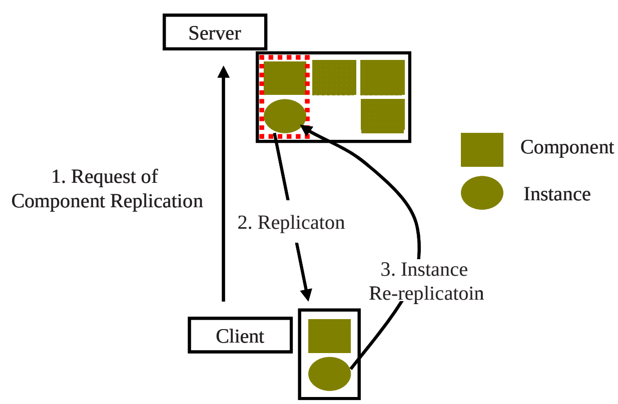


Figure 9. Component replication



should be migrated. For example, after copying three components and three instances, the target server and three clients would reconfigure their connections. The three instances would be removed from the origin server, resulting in reducing traffic on the origin server.

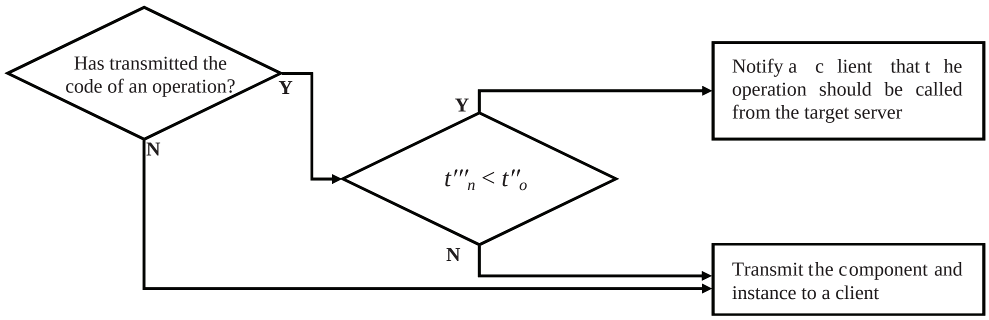
Replication to a Client

If a service requires a large message as an input parameter, a client may want to download the service code of a smaller size and execute it on his or her device, as illustrated in Figure 9.

To do this, a client should examine the component information about an operation and compare the sizes of the parameter and component. If a component were downloaded, an origin server would record the client on the replication list. Likewise, the origin server may request the client to delete the component.

If a replication is requested during migration or replication, a server, from which the component would be downloaded more quickly, should be determined, as shown in Figure 10. While t''_o indicates the download time of the component and instance from the origin server, t'''_n corresponds

Figure 10. The flowchart for a replication request during migration or replication



$t''_o = (S_c + S_i)/b_o$
$t'''_n = N/b_o + S_{req}/b_n + (S_c + S_i)/b_n$
S_c: the size of a component, S_i: the size of an instance
N: the size of a notification, S_{req}: the size of a request message
b_o: the bandwidth of an original server, b_n: the bandwidth of a target server

to the time to notify the client that the component and instance would be downloaded from a new server and when the client would receive them from the new server.

EXPERIMENTAL RESULTS

We have experimented evaluating the effect of the proposed code splitting approach on migration. Additionally, to show the feasibility of the proposed migration and replication, three application scenarios were implemented and tested. The tests were conducted on devices, which have an Intel(R) PXA270 processor, a memory of 62.28M, and the Windows Mobile 2003 operating system.

Code Splitting

To evaluate the performance of the proposed code splitting method, we measured the time to split codes from three different experiments. First, to investigate how much the method is affected by the dependency among operations, we experimented with test data, where Web services consist of 10 operations and each operation calls 3 functions. Second, concerning the number of functions called by an operation, the test was done with services, which contain 10 operations and 20 functions. Here we varied the number of functions, which are called by an operation, from 2 to 20. The third test is about the number of operations in a service. Web services contain operations that call 10 functions. Here the number of operations was varied from 1 to 10. In our experiments, the dependencies between operations and functions were randomly selected. In each experiment, the size of the functions also varied from 1 KB (kilobyte) to 100 KB.

Evaluation in Terms of the Dependency between Operations

If operations in a service share functions in common, we can say that they depend on each other. We define the dependency of operation a in a service as:

operation_dependency(a) = the number of operations sharing functions in common with a / the total number of operations in a service (2)

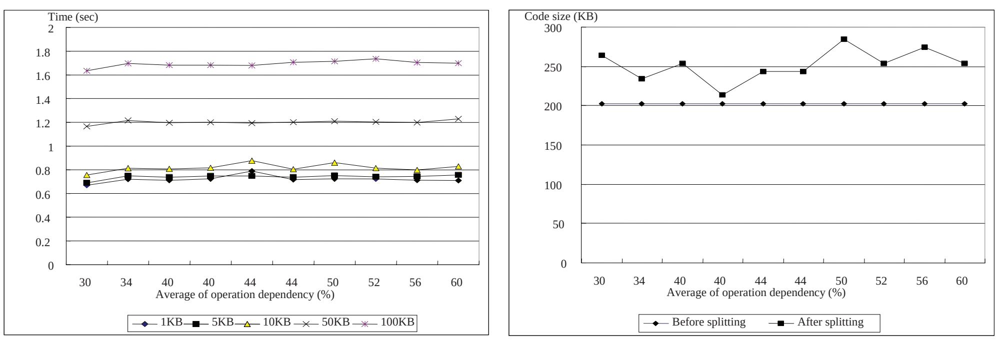
Figure 11(a) shows the experimental results of the relation between the average value of operation dependencies and the time taken to split codes. The figure shows that operation dependency and splitting time are scarcely related. If a Web service has many functions shared by its operations, the possibility of splitting would be low. Otherwise, the possibility of splitting would be high. Since only the functions shared by operations are copied, the number of copied functions is limited, resulting in increasing a limited amount of time. Consequently, the operation dependency is shown to be unrelated to the splitting time. Moreover, Figure 11(b) shows that the dependency is unrelated to the total size of the codes split. The test was done with services, whose operations and functions have the size of about 10 KB.

Evaluation in Terms of the Number of Functions Called by Operations

The experimental result of Figure 12(a) shows that as the number of functions called by operations grows, splitting time increases. Nodes in a call graph represent functions or operations. The number of functions called by operations determines the number of edges among function nodes. As the number of connections among nodes, we need more time to analyze and split the graph.

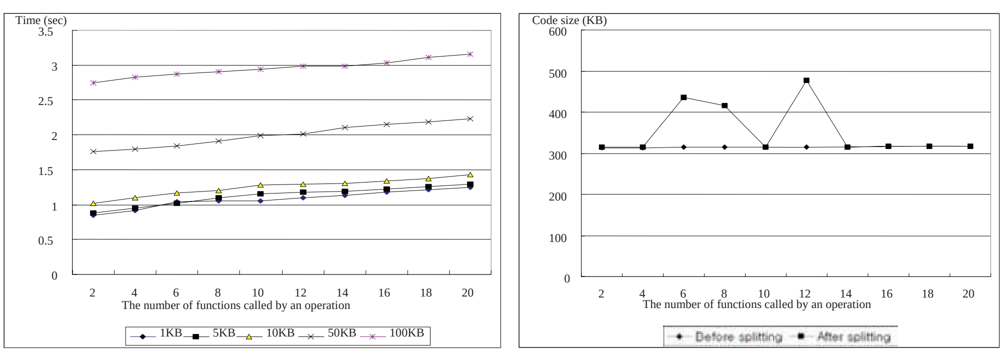
Figure 12(b) shows the relationship between the number of functions called by an operation

Figure 11. An experimental result on operation dependency



(a) Splitting time *(b) Total size of the codes split (function size 10KB)*

Figure 12. An experimental result on the number of function calls



(a) Splitting time *(b) Total size of the codes split (function size 10KB)*

and the size of the codes split. The test was done with services, whose operations and functions have the size of about 10 KB. In the cases where the numbers of the function called are 2 or 4—since the numbers are relatively small and each operation did not share functions—there was no difference in the size of the codes split before and after splitting. No regularities were found while increasing the number of functions called by an operation until it reached 12. In the case of 14 or more functions, operations called more than 70% of functions in this test. Splitting did not happen since too many functions needed to be copied otherwise.

Evaluation in Terms of the Number of Operations

As shown in Figure 13, there is a close correlation between the number of operations and the splitting time. An increase in the number of operations results in an exponential growth of the splitting time since all possible combinations of operations need to be considered to determine the code splitting of minimum cost.

As a result, we find that the splitting time of a service is related with its code size, the number of functions called by operations, and the number of operations, while the dependency among operations seldom affects the splitting time. In particular, the splitting time grows exponentially as the number of operations increases. We have a plan to consider this matter to enhance the processing time of the proposed method.

Migration and Replication

To show the feasibility of the proposed migration method, we implemented three application scenarios.

Migrating Web Services to a Mobile Device

The first scenario is about the migration of a Web service, which provides traffic information about a particular bridge. As described briefly in Figure 14, we assume that a service offers information about the traffic and weather situation of bridge continuously by migrating from a car leaving the bridge to a car arriving at the bridge. The service should be migrated to a new server because its origin server is leaving the bridge and so it cannot be provided any more. Additionally, the connections from clients should be reconfigured to the new car.

Figure 15(a) shows the log information of an origin server, which hosts the service. When the car is about to leave the bridge, a client calls the service to get information on traffic speed (getAveVel) and temperature (getTemperature). The service is being migrated to a new server

Figure 13. An experimental result on the number of operations

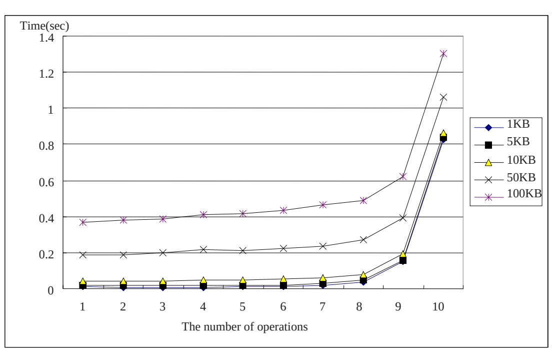


Figure 14. A scenario of service migration

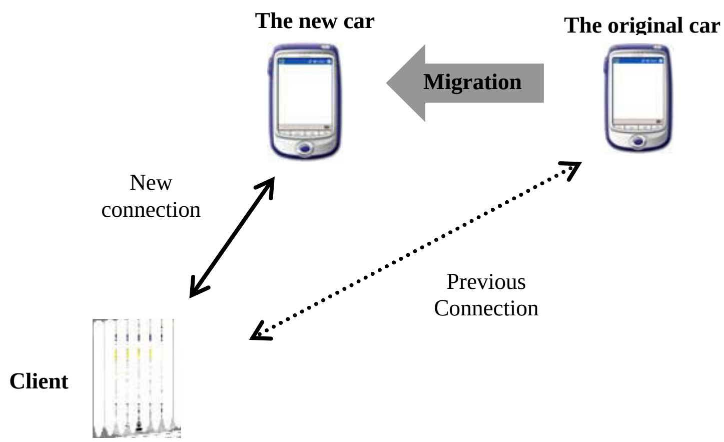
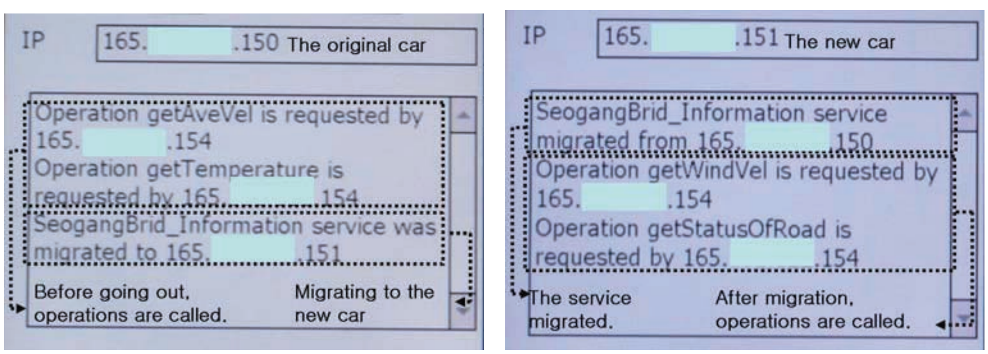


Figure 15. Log information of servers



arriving at the bridge. Figure 15(b) shows the log information of a car, to which the service has arrived and its two operations, that is, getWindVel and getStatusOfRoad, have been requested by the client

Replicating Web Services to a Mobile Device

This scenario is about a service replication for providing art information seamlessly in a museum. The museum in the scenario provides a Web service, which offers information about artists and their artworks, and if necessary, replicates the service to guiding devices, resulting in reduced network traffic.

Figure 16 provides a brief description of this scenario. Visitors or clients are served from the museum server, which collects the information and adjacent guiding devices and maintains their list. If the requests to the origin server increase suddenly, the origin server will an appropriate device, to which its service should be replicated. If a target device accepts the replication request from the origin server, the service code is replicated to the target device and the connections between the origin server and clients are reestablished in order to provide the service seamlessly.

While Figures 17(a) and 17(c) illustrate the interaction between the museum and its visitor before replication, Figures 17(b) and 17(d) show the interaction between the device of a guide and

Figure 16. A scenario of service replication

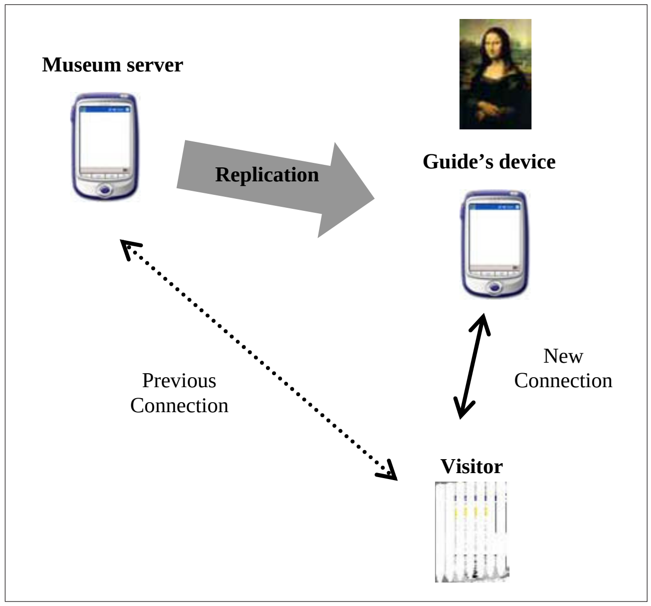
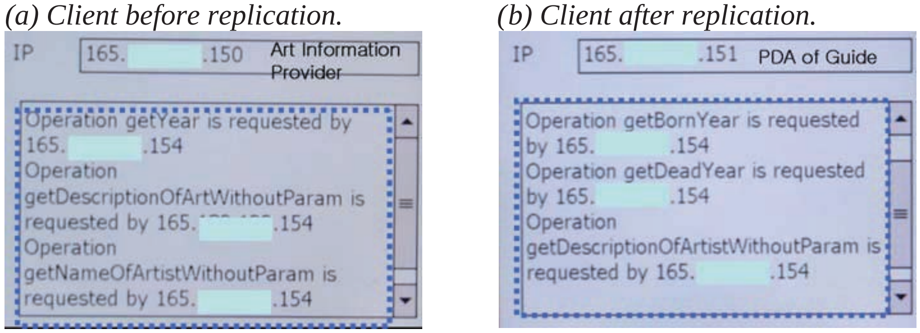


Figure 17. Screenshots of the server and client

(a) Client before replication. *(b) Client after replication.*

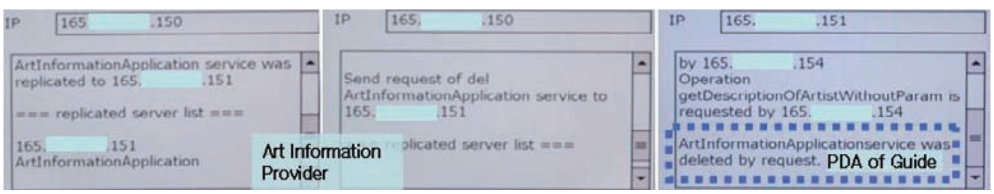


(c) Museum server before replication. *(d) Guide device after replication.*

the visitor after replication. Once a client calls an operation, the service provider maintains the instance and thus enables the client to call other operations of the same service without parameters. After the client calls operation getYear with parameter Monalisa, he or she could call two operations, such as getDescriptionOfArtWithoutParam and getArtistWithoutParam, without parameters. As shown in Figures 17(b) and 17(d), we find that the client could call three operations with no parameters (e.g., getBornYear, getDeadYear, and getDescriptionOfArtistWithoutParam). This is due to the fact that the instance of the service was replicated to the target server.

As described before, an origin server maintains a list of replicated services and may request a target server to delete the service replicated. Figure 18(a) shows that an origin server replicates a Web service to a target server and records the service and its target server. As shown in Figure 18(b), after replication, the origin server can request to the target server to delete the service replicated. Figure 18(c) indicates that the service was deleted from the target server according to the request of the origin server.

Figure 18. Deletion of the service replicated



(a) Before requesting deletion *(b) After requesting deletion* *(c) After deletion (target)*

Figure 19. Replicating a component to a client

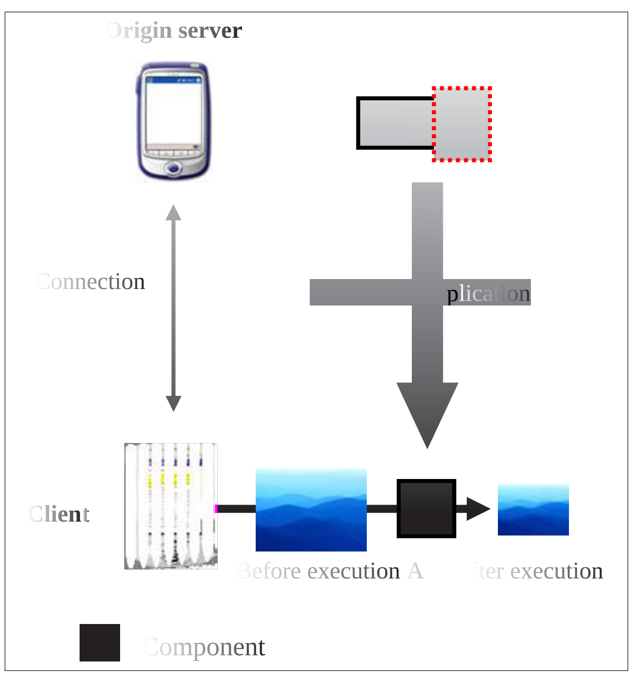
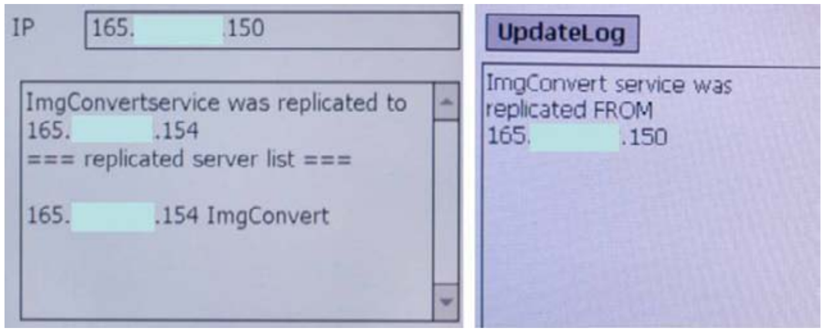


Figure 20. An example of component replication



(a) Log information of a server *(b) Log information of a client*

Replicating Components to a Client

There may be a case where if a service requires a large amount of input and output messages, it is more efficient to download and execute the service at a client side. In the scenario of Figure 19, an origin server provides a service of converting images and a client wants to resize an image. The client does not transmit a large image file as a parameter; instead, the client resizes the image by executing the service component replicated from the origin server.

Figure 20(a) shows the log information of the origin server, which provides the image conversion service. Figure 20(b) is the log information of the client that uses the service replicated. Likewise, the origin server records the service replicated and the client. It may also request the client to delete the service replicated.

Comparison with Previous Works

Table 1 shows the comparison result of the proposed method with previous works. The method of Pratitha and Zaslavsky (2004) migrates services to a target server, which is selected based on context information. Services are composed of service modules, which have been defined in advance depending on context. However, as the size of a service increases, it takes a lot of resources to migrate the service. They do not handle the migration of service instances and the replication of services. Hammerschmidt and Linnemann (2005) support the migration of instances to reduce network traffic and resource consumption on an origin server. Among the methods concerning code mobility, Bellavista et al. (2004) and Montanari et al. (2004) support the migration of binary codes, but do not offer the instance migration.

In this article, we present a method for migrating and replicating Web services on mobile and wireless environments. Based on the context information of mobile devices, the proposed method selects between migration and replication. Additionally, service instances are preserved during migration and replication. The proposed framework does not consider the reconfiguration of services after migration. The reconfiguration involves increases in code size since it requires codes—which are relevant to each context—to be implemented. Therefore, it may be not suitable for wireless and mobile environments.

Table 1. Comparison with previous works

Methods \ Features	Web services	Instance preservation	Reconfiguration	Supporting mobile and wireless environments	Context –awareness
Pratitha and Zaslavsky (2004)	O*	X	O	X	O
Hammerschmidt and Linnemann (2005)	O	O	X	X	X
Ishikawa, Yoshioka Tahara, and Honiden (2004)	O	X	X	X	X
Mifsud and Stanski (2002)	O	X	X	X	X
Kern and Braun (2005)	X	O	X	X	X
Bellavista, Corradi, and Magistretti (2004)	X	X	X	O	O
Montanari, Lupu, and Stefanelli (2004)	X	X	X	X	O
The proposed method	O	O	X	O	O

* *O: supported, X: not supported*

CONCLUSION AND FUTURE STUDY

This article presents an efficient method for migrating and replicating Web services through code splitting. Specifically, a service is split into subcodes based on users' preferences to its constituent operations. The subcodes of higher preference are migrated earlier to minimize the discontinuance of the operations of high priority and raise the efficiency of Web services migration and replication in wireless and mobile environments. From experimental results, we found that the proposed splitting method depended on the size of a Web service and the number of functions called by its operations. In particular, as the number of operations increased, the splitting time grew at an exponential rate. Moreover, the proposed migration method was experimented under three application scenarios. Experimental results show that the proposed migration method helps services to be provided continuously.

The proposed method selects an optimal combination of code split, on which the source code is split and compiled into components. However, it may be necessary to merge and split the codes split repeatedly since user preferences may change rapidly and frequently. Therefore, to reflect the changing desires and preferences of users dynamically at runtime, we have a plan to enhance the proposed migration method to make it possible to accommodate user feedback at runtime.

ACKNOWLEDGMENT

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2005-041-D00690).

REFERENCES

Bellavista, P., Corradi, A., & Magistretti, E. (2004, September 20-22). Lightweight code mobility for proxy-based service rebinding in MANET. In Proceedings of the 1st International Symposium on Wireless Communication Systems, Port Louis, Mauritius (pp. 208-214).

Grove, D., & Chambers, C. (2001). A framework for call graph construction algorithms. ACM Transactions on Programming Languages and Systems, 23(6), 685-746.

Grove, D., DeFouw, G., Dean, J., & Chambers, C. (1997, October 5-9). Call graph construction in object-oriented languages. In Proceedings of the ACM SIGPLAN Conference on Object-Oriented Programming Systems Languages and Applications, Atlanta (pp. 108-124).

Hammerschmidt, B. C., & Linnemann, V. (2005, February 22-25). Migrating stateful Web services using Apache AXIS and P2P. In Proceedings of the IADIS International Conference on Applied Computing, Algarve, Portugal (pp. 433-441).

Huhns, M. N., & Singh, M. P. (2005). Service-oriented computing: Key concepts and principles. IEEE Internet Computing, 9(1), 75-81.

Ishikawa, F., Yoshioka, N., Tahara, Y., & Honiden, S. (2004, July 19-20). Toward synthesis of Web services and mobile agents. In Proceedings of AAMAS Workshop on Web Services and Agent-based Engineering (WSABE), New York (pp. 227-245).

Kern, S., & Braun, P. (2005, September 20-22). Towards adaptive migration strategies for mobile agents. In Proceedings of the Second GSFC/IEEE Workshop on Radical Agent Concepts (WRAC) (pp. 334-345). NASA Goddard Space Flight Center Visitor's Center Greenbelt, MD.

Kern, S., Braun, P., Fensch, C., & Rossak, W. (2004, October 25-29). Class splitting as a method to reduce migration overhead of mobile agents. In Proceedings of the International Conference on Ontologies, Databases and Application of Semantics (ODBASE'04), Agia Napa, Cyprus (pp. 1358-1375).

Kim, Y.-S., & Lee, K.-H. (2007). An efficient policy establishment scheme for Web services migration. In Proceedings of the International Conference on Convergence Information Technology (ICCIT'07) (pp. 595-600).

Mifsud, T., & Stanski, P. (2002, October 8-11). Measuring performance of dynamic Web service migration using LAMS. In Proceedings of the 10th International Conference on Software, Telecommunications and Computer Networks (SoftCOM), Venice, Italy (pp. 214-218).

Montanari, R., Lupu, E., & Stefanelli, C. (2004). Policy-based dynamic reconfiguration of mobile-code applications. Computer, 37(7), 73-80.

Pratistha, I. M., & Zaslavsky, A. B. (2004, March 14-17). Fluid: Supporting a transportable and adaptive Web service. In Proceedings of the 2004 ACM Symposium on Applied Computing, Nicosia, Cyprus (pp. 1600-1606).

Schall, D., Aiello, M., & Dustdar, S. (2006). Web services on embedded devices. Web Information System, 1(1), 1-6.

Sirirama, S. N., Jarke, M., & Prinz, W. (2006, February 19-22). Mobile Web services provisioning. In Proceedings of the Advanced International Conference on Telecommunicatons and International Conference on Internet and Web Application and Services, Guadeloupe, France (pp. 120-126).

Stal, M. (2006). Using architectural patterns and blueprints for service-oriented architecture. IEEE Software, 23(2), 54-61.

This work was previously published in International Journal of Web Services Research, Vol. 6, Issue 2, edited byL. Zhang, pp. 1-19, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 3.4

Applying Web-Based Collaborative Decision-Making in Reverse Logistics: The Case of Mobile Phones

Giannis T. Tsoulfas
University of Piraeus, Greece

Costas P. Pappis
University of Piraeus, Greece

Nikos I. Karacapilidis
University of Patras, Greece

ABSTRACT

The increasing environmental concerns and the technological advances have boosted the post-use treatment of nearly all kinds of products and a new area for research and application has emerged described by the term "reverse logistics." In this chapter, parameters that may affect reverse logistics operation are discussed from a decision-making perspective, so that alternative design options may be proposed and evaluated. In particular, these parameters are used for the qualitative evaluation of the reverse supply chain of mobile phones in Greece. For this purpose, we present an illustrative application of a Web-based decision support tool that may assist collaborative decision-making in conflicting environments, where diverse views, perspectives, and priorities shared among stakeholders have to be considered.

DOI: 10.4018/978-1-60566-114-8.ch018

INTRODUCTION

The increasing environmental concerns and the technological advances have boosted the post-use treatment of nearly all kinds of products, regardless of their size, composition, and initial value. Relevant legislative frameworks have been enforced in developed countries aiming at apportioning the responsibilities related to the recovery of end-of-life products. In addition, specific targets regarding product design and recovery rates are set, networks' requirements are suggested and, last but not least, voluntary schemes are applauded. As a result, further extensions in research and applications of supply chain management have emerged described by the term "reverse logistics." De Brito and Dekker (2004) defined reverse logistics as "the process of planning, implementing and controlling backward flows of raw materials, in process inventory packaging and finished goods, from a manufacturing, distribution or use point, to a point of recovery or point of proper disposal." In this definition both economic and environmental dimensions of reverse logistics are implied, indicating the potential benefits that companies would have by adopting such practices.

Reverse logistics is a multidisciplinary area of research. For example, operations research, environmental analysis, marketing, and informatics have all a significant role to play in order to assist decision-making regarding the design and operation of reverse supply chains. Moreover, reverse logistics is often regarded in conjunction with forward logistics, since they are interrelated. However, the distinguishing characteristics of reverse supply chains introduce new dimensions in decision-making aspects. In particular, the main differences between forward and reverse supply chains, as stated by Fleischmann, Krikke, Dekker, and Flapper (1999) and Krikke, Pappis, Tsoulfas, and Bloemhof-Ruwaard (2002), are the following:

- In contrast to forward supply chains, in reverse supply chains there are a lot of sources of "raw materials" (used products), which may enter the reverse flow at low or no cost at all, and significantly fewer "customers" (recyclers, remanufacturers, etc.).
- The economic value of inputs in reverse supply chains is lower than the one in the case of forward supply chains.
- In the case of reverse supply chains, offer does not follow demand.
- The economic efficiency of reverse supply chains is precarious, since it is not sure that there will be markets to exploit their outputs.
- Reverse supply chains are characterized by higher uncertainty regarding issues like quality, volumes, and composition of reverse flows.

From this perspective, it is important to identify the parameters that may affect reverse logistics operation so that alternative design options are proposed and evaluated. In Tsoulfas, Dasaklis, and Pappis (2007), a first attempt to define and categorize them is presented. Given these parameters, in this chapter we discuss a qualitative evaluation of the reverse supply chain of mobile phones in Greece, as presented by Pappis, Tsoulfas, and Dasaklis (2006). For this purpose, we make use of a Web-based decision support tool that may assist collaborative decision-making (CDM) in conflicting environments, where diverse views, perspectives, and priorities shared among stakeholders have to be considered.

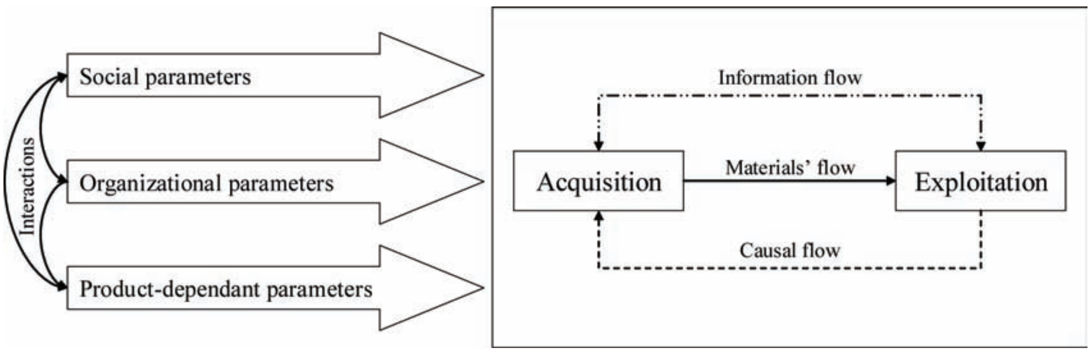
The remainder of the chapter is structured as follows: First, the parameters affecting reverse logistics operation are discussed. Then, the reverse supply chain of mobile phones in Greece is briefly presented. Next, the CDM tool is presented, followed by its illustrative application regarding the reverse supply chain of mobile phones in Greece. Finally, some concluding remarks are outlined.

PARAMETERS AFFECTING REVERSE LOGISTICS OPERATION

Three major categories of parameters that may affect reverse logistics operation are identified: product-dependant, organizational, and social. These parameters, which cannot be addressed independently since they may interact with each other, may form a nonexhaustive basis for analysis in the following decision-making situations:

a. When assessing the current situation regarding the operation of reverse supply chains;
b. When exploring alternative options for the reverse supply chain activities, as well as their interaction with the external environment.

Figure 1. The various flows between the two major sessions of reverse supply activities



Generally speaking, reverse supply chains may be considered as the conjunction of two major sessions of activities: acquisition and exploitation. The first one refers to the activities that aim at the physical transportation of used products and the second one includes the activities targeting final value extraction or environmentally sound management. Although exploitation follows acquisition when the materials' flows are regarded, it may be considered as a necessary condition for the acquisition in the causal chain. To be more specific, the ability to exploit used products may trigger their acquisition. Otherwise, it would be purposeless to acquire used products without having in mind how to treat them. In addition, information exchange between these sessions is bidirectional and repeated. The relationships between the two major sessions of reverse supply activities are illustrated in Figure 1, where the parameters affecting these operations are regarded as influential factors.

Product-Dependent Parameters

Product-dependant parameters refer to the particular characteristics of products that determine their post-use treatment from a technical and an economic point of view. In particular:

- *The weight or the volume of used products and the infrastructure needed* is a decisive criterion for the development of reverse logistics activities, since several operations, such as collection, storage, and transportation, may be affected. Generally speaking, large products may require special machinery and equipment for handling, transportation, and so forth, whereas small products may call for big quantities to be collected before being transported.
- *The composition and the technical characteristics of used products* may be another decisive issue, since they determine the ways used products should be treated in order to preserve their value and to prevent them from harming the environment.
- *The way used products are replaced by new ones* is another important parameter for the implementation of reverse logistics. It is obvious that replacements, which occur in the same place or using the same distribution means for the return of used products, have positive effects on the reverse supply chain operation, both from an economic and an environmental perspective.
- *The remaining value of used products* is considered to be a very significant issue for consumers and manufacturers, since they can both benefit from the post-use treatment of used products. Consumers may achieve reduced prices for new products replacing used ones, while manufacturers

can extract value from used products by refurbishing, reusing, or recycling them.

- *Direct reuse or reuse after minor treatment* is a situation commonly perceived in the case of packaging materials and may offer significant benefits to companies and the environment, as the useful life cycle of used products is extended and the production of new ones is avoided.
- *The capability to change the usage of used products or to provide them to different markets (e.g., second hand)* may be another important criterion. Generally, in such situations no special treatments are necessary and the life cycle of products is extended.

Organizational Parameters

Organizational parameters refer to issues regarding the stakeholders involved in the recovery of used products. In particular:

- *The recovery networks structure* is a decisive element for reverse logistics activities. Generally speaking, companies have two options: either they will handle the recovery processes themselves (even by using outsourcing practices), or they will participate in wider networks, usually by financially supporting them. The first option is commonly adopted when companies can achieve significant return rates, whereas the second scheme is preferred, especially when used products are widely dispersed.
- *Asset control policies* that are adopted by some companies can contribute to the effective operation of reverse logistics. Such cases are often met, for example, in the automotive sector and in the electronics industry. By using such practices, companies actually sell services rather than the product itself. As a result, they can have improved control of their products and, at the same time, fulfill their customers' needs, with whom they can easier communicate.
- *Marketing* is a very important criterion for the implementation of reverse logistics. Companies may participate in campaigns for promoting collection of used products and they can indicate recovery options in the products themselves or in their packages.
- *Economic motivation* is a means used in many industrial sectors in order to involve consumers in the recovery activities. Usually, consumers prepay a certain amount of money as a deposit and they get it back when they return the product or the package to collection facilities.

Social Parameters

Social parameters involve attitudes and values prevailing in societies that may determine practices regarding recovery of used products. In particular:

- *Social habits* may significantly affect the results of reverse logistics activities, since individual attitudes are often affected by mainstreams. Recovery of used products seems to find a more fertile ground in big cities rather than in small communities.
- *Legislation* is a decisive parameter affecting the recovery of used products. In particular, the principle of shared responsibility and the "cradle to grave" perspective have been elevated in legislative frameworks around the world. Furthermore, explicit targets are posed and certain benefits are offered in some cases (e.g., tax relieves, improved financial eligibility, etc.).
- *Social awareness* is a critical issue regarding reverse logistics practices, especially with respect to consumers' attitudes. In developed countries, the environmental standards stemming from social demand are higher. Education is of particular

importance regarding this parameter, not solely in schools, but also in corporate environments.

THE REVERSE SUPPLY CHAIN OF MOBILE PHONES IN GREECE

Facts

In 2002, the total number of mobile phones in use worldwide exceeded the number of land-lines (Donner, 2005). According to the International Telecommunication Union the mobile subscribers in 2006 were more than 2.5 billions (International Telecommunication Union, 2007). According to the same source, the subscribers in Greece were around 11 million.

Typically, mobile phones are used for only 1½ years before being replaced (Fishbein, 2002). These obsolete mobile phones are mainly replaced due to fashion trends and the rapid technological improvements, as new features are added in mobile phones. Other reasons for replacement are the incompatibility with a new provider, or the fact that they no longer function. Less than 1% of mobile phones retired and discarded annually are recycled and the majority is accumulated in consumers' desk drawers, store rooms, or other storage, awaiting disposal (Most, 2003). Of this small percentage recovered, some are refurbished and put into use or used for replacement parts. If these options are not possible, some metals are recycled. The refurbishment process can significantly aid to the prolongation of a mobile phone's life cycle and therefore prevent it from early entry into the waste stream. The recycling process keeps discarded phones out of disposal facilities and reduces the need for raw materials used to make new products.

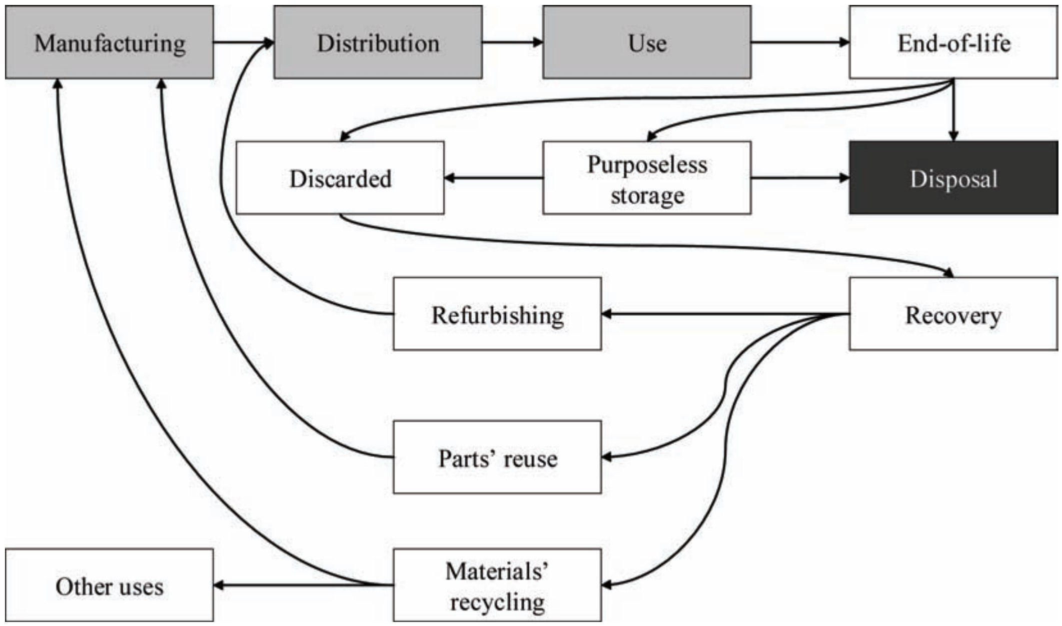
In the case of Greece, Appliances Recycling S.A. is the authorized collective take-back and recycling organisation for all electrical and electronic waste in Greece (Pappis et al., 2006). Actually, all service providers and importers are obliged by law to cooperate with Appliances Recycling S.A., since they are the only authorized take-back organisation in Greece right now. The program relies on in-store collection and special bins have been installed in retail stores. In addition, several bins have been put on central city spots, as the result of cooperation with local municipalities. Recycling Appliances S.A. aimed at covering 67% of Greek mobile phones population by the end of 2006, while their corresponding target for 2008 is 90% (http://www.electrocycle.gr). The used mobile phones collected have been destined only abroad for further treatment, since no appropriate facilities exist in Greece. In the recycling process, the plastic parts of mobile phones are incinerated and utilised as a fuel to melt the metal mixture. Then metals are separated using electrolytic refining and mechanical (e.g., magnetic segregation) procedures.

The possible routes of used mobile phones and the affected activities in the forward supply chain (grey color) are illustrated in Figure 2.

Major Concerns

Mobile phones contain a great number of metals such as copper, aluminium, iron, nickel, silicon, lead, antimony, beryllium, arsenic, silver, tantalum, and zinc. Some of these metals are toxic and hazardous for mankind and the environment. This variety of valuable metals raises very significant issues regarding the gradual exhaustion of natural resources. In addition, the side effects of this exhaustion are also important. For example, the mining of tantalum has been identified as a serious threat to gorillas clinging to survival in the Democratic Republic of Congo (Macey, 2005). Apart from metals, mobile phones contain also brominated flame retardants, which are used in the plastic parts and cables in order to reduce the risk of fire. When burned in incinerators, these substances have the potential to pollute the air and to pose threats for the workers in recycling facilities, since dioxins and furans can be formed. When buried in landfills, they may leach into soil

Figure 2. The possible routes of used mobile phones and the affected activities in the forward supply chain (grey color)



and drinking water.

The environmental impact of the substances mentioned above is of great concern because some of them, like flame retardants and lead, are considered to be persistent, bioaccumulative, and suspected carcinogens. Relative legislation enforcement in the European Union aims at the restriction of the use of certain hazardous substances in electrical and electronic equipment, such as mobile phones (RoHS Directive) (European Union, 2003a). In addition, the WEEE Directive draws the frame regarding the post-use treatment of electrical and electronic equipment (European Union, 2003b).

Apart from the environmental concerns related to the treatment of used mobile phones, there are some important economic issues as well. Indeed, many substances contained in mobile phones are valuable as it is relatively more expensive to acquire them as primary raw materials (e.g., lead, zinc).

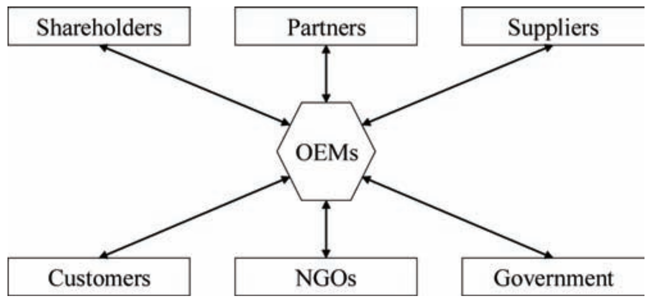
It is obvious that decision-making procedures regarding the operation of reverse supply chains get more complicated, due to the involvement of diverse parties. For example, in the case of mobile phones, manufacturers, distributors, service providers, recovery operators, and recyclers would be the participants of such a decision-making situation. Moreover, even different departments of these stakeholders might have diverse views of the situation.

The Decision-Making Situation

From an OEMs' perspective, reverse logistics implementation in the case of mobile phones is determined by their interaction with several stakeholders, as shown in Figure 3.

Thus, reverse logistics managers are responsible for taking into account and coordinating all stakeholders' requirements. Correspondingly, similar actions are necessary among the different departments within a company. Consequently, a conflicting decision-making environment is formed, where the factors "place" and "time" may pose restrictions. Such decision-making

Figure 3. OEMs and their interactions with stakeholders in reverse logistics activities



situations may be dealt with Web-based CDM tools and a corresponding approach is presented in the sequel.

A WEB-BASED TOOL FOR COLLABORATIVE DECISION-MAKING

Approaching Conflicting Decision-Making Situations

Choices in decision-making cannot generally be addressed by individuals working alone or even by several people working separately and then merging their pieces of work. Instead, they have to be addressed through collaborative work among stakeholders with diverse views, perspectives, and priorities.

Information and Communication Technology (ICT) infrastructure to support people working in teams has been the subject of interest for quite a long time (Fjermestad & Hiltz, 2000). Such systems aim at facilitating group decision-making processes by providing forums for expression of opinions, as well as qualitative and quantitative tools for aggregating proposals and evaluating their impact on the issue in hand. They may exploit intranet or Internet technologies to connect decision-makers in a way that encourages dialogue and, at the same time, stimulates the exchange of knowledge. Accordingly, recent computer-based knowledge management systems (KMS) focus on providing a corporate memory, that is, an explicit, disembodied, and persistent representation of the knowledge and information in an organization, as well as mechanisms that improve the sharing and dissemination of knowledge by facilitating interaction and collaboration among the parties involved (Bolloju, Khalifa, & Turban, 2002).

CDM may provide a means for a well-structured decision-making process. Usually, CDM is performed through debates and negotiations among the parties involved. Conflicts of interest are inevitable and support for achieving consensus and compromise is required. Decision-makers may adopt and suggest their own strategy that fulfils some goals at a specific level and may have arguments supporting or against alternative solutions. In addition, they have to confront the existence of insufficient information. Generally speaking, efficient and effective use of information technology in the collection and dissemination of information and knowledge produced by diverse sources, the evaluation of alternative schemes, the construction of shared meanings, and the associated feedback learning process are critical factors for the decision-making process (Clases & Wehner, 2002).

The Web-Based Tool

Given the above issues, a Web-based tool has been implemented that supports the collaboration conducted in decision-making situations, by facilitating the creation, leveraging, and utilization of the relevant knowledge. This tool is based on an argumentative reasoning approach, where discourses about complex problems are considered as social processes and they may result in the formation of groups whose knowledge is clustered around specific views of the problem (Karacapilidis, Adamides, & Pappis, 2004). In addition to providing a platform for group reflection and capturing of organizational memory, this approach augments teamwork in terms of knowledge elicitation, sharing, and construction, thus enhancing the quality of the overall process. This is due to its structured language for conversation and its mechanism for evaluation of alternatives. Taking into account the input provided by the individual experts, the proposed tool constructs an illustrative discourse-based knowledge graph that is composed of the ideas expressed so far, as well as their supporting documents. Moreover, through the integrated decision support mechanisms, experts are continuously informed about the status of each discourse item asserted so far and reflect further on them according to their beliefs and interests on the outcome of the discussion. In addition, the overall approach aids group sense-making and mutual understanding through the collaborative identification and evaluation of diverse opinions.

The proposed tool builds on a server-client network architecture. It is composed of two basic components, namely the *collaboration visualization module* and the *collaborative decision making module*. The former provides a shared Web-based workspace for storing and retrieving the messages and documents deployed by the discussion participants, using the widely accepted XML document format (http://www.w3.org/XML). This module actually provides the interfaces through which participants get connected with the system via Internet (by using a standard Web browser; there is no need of installation of any specific software in order to use the tool). Exploitation of the Web platform renders, among others, low operational cost and easy access to the system. The knowledge base of the system maintains all the above items (messages and documents), which may be considered, appropriately processed and transformed, or even re-used in future discussions. Storage of documents and messages being asserted in an ongoing discussion takes place in an automatic way, upon their insertion in the discussion. On the other hand, retrieval of knowledge is performed through appropriate interfaces, which aid participants in exploring the contents of the knowledge base and exploit previously stored or generated knowledge for their current needs. In such a way, our approach builds a "collective memory" of a particular community. On the other hand, the collaborative decision-making module is responsible for the reasoning and evaluation purposes of the system. Alternative mechanisms for these purposes can be invoked each time, upon the participants' wish and context under consideration. These mechanisms follow well-defined and broadly accepted algorithms (based on diverse decision making approaches, such as multi-criteria decision-making, argumentation-based reasoning, utility theory, risk assessment, etc.), which are stored in the tool's model base.

The basic discourse elements in the proposed tool are *issues, alternatives, positions,* and *preferences*. In particular, issues correspond to problems to be solved, decisions to be made, or goals to be achieved. They are brought up by users and are open to dispute (the root entity of a discourse-based knowledge graph has to be an issue). For each issue, the users may propose alternatives (i.e., solutions to the problem under consideration) that correspond to potential choices. Nested issues, in cases where some alternatives need to be grouped together, are also allowed. Positions are asserted in order to support the selection of a specific course of action (alternative), or avert the users' interest from it by expressing some objection. A position may also refer to another (previously asserted) position, thus arguing in favor or against it. Finally, preferences provide individuals with a qualitative way to weigh reasons for and against the selection of a certain course of action. A preference is a tuple of the form (position, relation, position), where the relation can be "more important than" or "of equal importance to" or "less important than." The use of preferences results in the assignment of various levels of importance to the alternatives in hand. Like the other discourse elements, they are subject to further argumentative discussion.

These four types of elements enable users to contribute their knowledge on the particular problem (by entering issues, alternatives, and positions) and also to express their relevant values, interests and expectations (by entering positions and preferences). In such a way, the tool supports both the rationality-related dimension and the social dimension of the underlying collaborative decision-making process. Moreover, the tool continuously processes the elements entered by the users (by triggering its reasoning mechanisms each time a new element is entered in the graph), thus enabling users to become aware of the elements for which there is (or there is not) sufficient (positive or negative) evidence, and accordingly conduct the discussion in order to reach consensus.

ASSESSING THE OPERATION OF THE REVERSE SUPPLY CHAIN OF MOBILE PHONES IN THE CASE OF GREECE

An illustrative application of the Web-based tool presented earlier is conducted regarding the qualitative assessment of the operation of the reverse supply chain of mobile phones in the case of Greece. In this application, decision-makers A, B, and C explore interventions in the operation of the chain as well as their possible interaction with the external environment. The parameters that affect reverse logistics operations are used as a basis for the discourse. The decision-making process may reveal flaws of current practices as well as improvement potentials and areas to focus on.

Figures 4 and 5 correspond to instances of collaboration concerning the *"Recovery network structure,"* and *"Marketing,"* respectively. In these instances, the stakeholders participate in an argumentation-based decision-making process. More specifically, in the instance shown in Figure 4, the issue under consideration is *"Priorities in improving the recovery network's structure,"* while three alternatives, namely *"Extended cooperation with local municipalities," "Collection bins in super markets,"* and *"Collection programs in schools,"* have been proposed so far (by C, A, and B, respectively). The three stakeholders have argued about them by expressing positions speaking in favor or against them.

For instance, *"People visit super markets at least once a week"* is a position (asserted by A) that argues in favor of the second alternative, while *"It is time-consuming"* is a position (asserted by B) that argues against the first alternative. As also shown in Figures 4 and 5, argumentation can be conducted in multiple levels. Furthermore, users may also assert preferences about positions already expressed. As shown in Figure 5, user C has expressed a preference concerning the relative importance between the positions *"It is very expensive to initiate a nationwide campaign"* and *"The timing is excellent,"* arguing that the first position is of bigger importance for him. Users may also express their arguments in favor or against a preference.

When clicking on a discourse item, detailed information about it is provided in a dedicated window of the basic interface of the tool. More specifically, this part contains information about the user who submitted the selected discussion element, its submission date, any comments that the user may have inserted, as well as links to related Web pages and multimedia documents that

Figure 4. Instances of collaboration concerning "Priorities in improving the recovery network's structure"

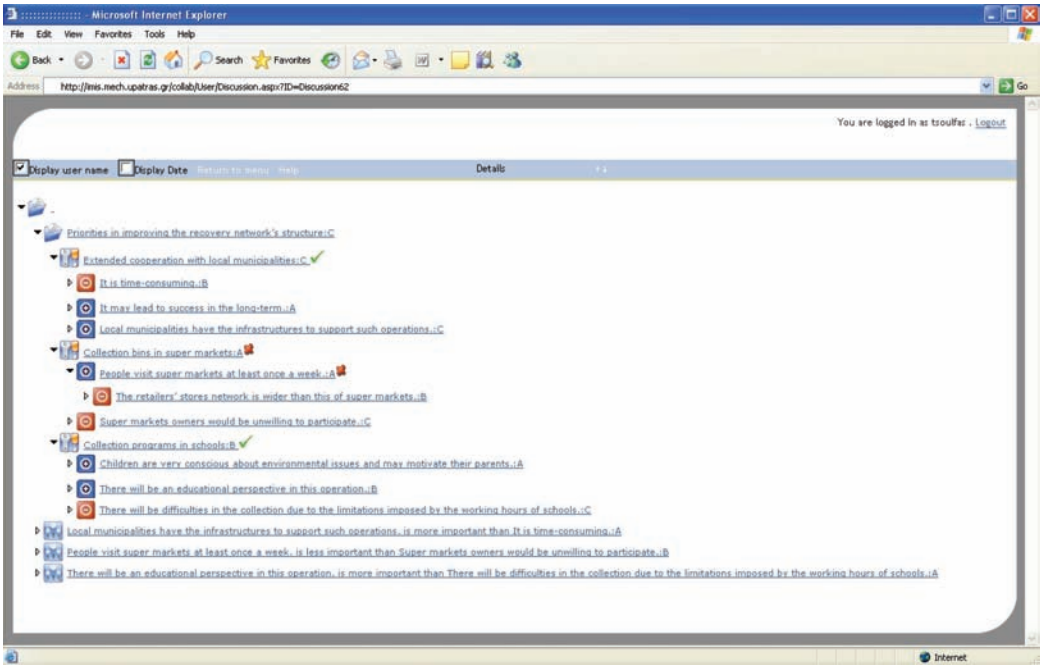
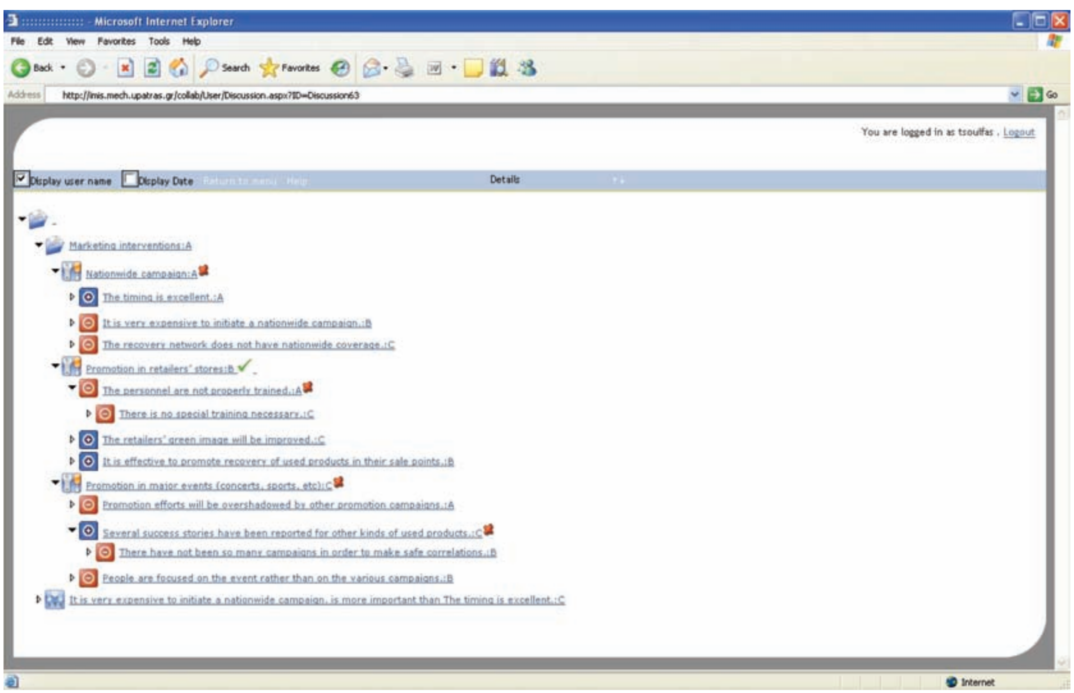


Figure 5. Instances of collaboration concerning "Marketing interventions"



the user may have uploaded to the tool in order to justify this element and aid his/her peers in their contemplation.

Further to the argumentation-based structuring of a discourse, the tool integrates a reasoning mechanism that determines the status of each discourse item in order to keep users aware of the discourse outcome. More specifically, alternatives, positions, and preferences of a graph have an *activation label* (it can be *"active"* or *"inactive"*) indicating their current status (inactive entries are indicated with a red "x"). This label is calculated according to the argumentation underneath and the type of evidence specified for them. In the instance of Figure 4, the position "*People visit super markets at least once a week*" is inactive because, according to the argumentation rule holding for this specific discussion, it has been defeated by the position "*The retailers' stores network is wider than this of super markets.*" Activation in the tool is a recursive procedure; a change of the activation label of an element is propagated upwards in the discourse graph. Depending on the status of positions and preferences, the mechanism goes through a scoring procedure for the alternatives of the issue. A detailed presentation of more technical details concerning the argumentation-based reasoning and scoring mechanisms of the tool can be found in Karacapilidis and Papadias (2001).

At each discourse instance, the tool informs users about what is the most prominent (according to the underlying argumentation) alternative solution (this is shown by a green "tick" sign). In the instance shown in Figure 4, *"Extended cooperation with local municipalities"* and "*Collection programs in schools"* are equally justified as best solutions, while in the instance shown in Figure 5 *"Promotion in retailers' stores"* is the better justified solution so far. However, this may change upon the type of the future argumentation; each time an alternative is affected during the discussion, the issue it belongs to is updated, since another alternative solution may be indicated by the tool.

CONCLUSION

The introduction of reverse logistics in supply chain management has created new decision-making dimensions. Consequently, parameters that may affect the operation of reverse supply chains should be evaluated. In this chapter, a qualitative approach has been discussed with respect to such parameters, aiming at facilitating and augmenting decision-making in reverse supply chains. In such cases, several stakeholders get involved, including governments, producers, distributors, and customers. As a result, decision-making procedures get more complicated due to increased levels of conflicts of interests but also due to practical reasons. For example, it is not always easy to get all stakeholders together in a round table. As it has been illustrated in this chapter, ICT may support decision-making procedures in conflicting environments by providing the means to structure dialogue, disseminate information, and last but not least, facilitate the associated reasoning process.

FUTURE RESEARCH DIRECTIONS

The parameters affecting reverse logistics operation may guide decision-makers towards identifying possible modifications in supply chain activities as well as in other corporate issues, such as marketing and supplier selection. Further research should be devoted to explore the interactions among these parameters and the ways they affect the reverse supply chains' operation and success. In addition, research efforts should also focus on how reverse supply chains may interact with forward supply chains and on relevant expedient strategies that aim at making the extended supply chains more efficient. Moreover, the qualitative evaluation of the reverse supply chains of different products may reveal the determinant parameters for each case, helping to create a body of knowledge based on thorough observations. In particular, it is important to identify the circumstances under which reverse supply chains are impeded and the options to improve their operation. Such knowledge may be exploited in developing organizational memory, a process which, beyond storing individual and collective knowledge, is related to organizational learning, decision-making, and competitive capability issues.

Apart from economic criteria, additional criteria (environmental, social, etc.) get increasingly involved in decision-making problems, leading to more complex decision-making situations. Moreover, such problems may not be usually addressed by formal models or methodologies. Instead, an argumentative practical reasoning approach seems to offer a more convenient solution. Thus, decision-making tools should focus on facilitating the cooperation of the different parties involved towards well-structured decision-making processes. The corresponding technologies should further exploit the advances in ICT in order to deliver applications of enhanced performance to decision-makers, while efficiently and effectively addressing communication and collaboration requirements. In particular, it is important to develop a more human-centric view of the problem, which appropriately structures and manages the underlying human interaction. The CDM tool discussed previously in this chapter may be exploited in order to retrieve useful information and knowledge, as well as to reason according to previous cases or predefined rules. The proposed tool may be enhanced with intelligent agent technologies, which are able to facilitate a variety of decision-makers' tasks and actions by acting on their behalf, as well as to automate system's processes.

ACKNOWLEDGMENT

The project is co-funded by the European Social Fund and national resources - EPEAEK II.

REFERENCES

Bolloju, N., Khalifa, M., & Turban, E. (2002). Integrating knowledge management into enterprise environments for the next generation decision support. *Decision Support Systems*, *33*(2), 163–176. doi:10.1016/S0167-9236(01)00142-7

Clases, C., & Wehner, T. (2002). Steps across the border: Cooperation, knowledge production and systems design. *Computer Supported Cooperative Work*, *11*(1/2), 39–54. doi:10.1023/A:1015207530896

de Brito, M. P., & Dekker, R. (2004). A framework for reverse logistics. In R. Dekker, K. Inderfurth, L. N. van Wassenhove, & M. Fleischmann (Eds.), *Quantitative approaches to reverse logistics* (pp. 3-27). Berlin, Germany: Springer-Verlag.

Donner, J. (2005, June 7-8). *Research approaches to mobile use in the developing world: A review of the literature*. Paper presented at the International Conference on Mobile Communication and Asian Modernities, Hong Kong.

European Union. (2003a). Directive 2002/95/EC of the European Parliament and of the Council of 27 January 2003 on the restriction of the use of certain hazardous substances in electrical and electronic equipment. *Official Journal of the European Union.*

European Union. (2003b). Directive 2002/96/EC of the European Parliament and of the Council of 27 January 2003 on waste electrical and electronic equipment (WEEE). *Official Journal of the European Union.*

Fishbein, B. K. (2002). *Waste in the wireless world: The challenge of cell phones*. INFORM Inc., New York, USA.

Fjermestad, J., & Hiltz, S. R. (2000). Group support systems: A descriptive evaluation of case and field studies. *Journal of Management Information Systems*, *17*(3), 115–159.

Fleischmann, M., Krikke, H. R., Dekker, R., & Flapper, S. D. P. (1999). *Logistics network (re)-design for product recovery and reuse* (Management Report 17). Rotterdam, The Netherlands: Erasmus University of Rotterdam.

International Telecommunication Union. (2007). *Basic indicators: Population, GDP, total telephone subscribers and total telephone subscribers per 100 people*.

Karacapilidis, N., Adamides, E., & Pappis, C. P. (2004, September 20-25). An IS framework to support the collaborative design of supply chains. In M. Negoita, R. J. Howlett, & L. C. Jain (Eds.), *Knowledge-based intelligent information and engineering systems: Proceedings of the 8th International Conference, KES 2004*, Wellington, New Zealand, Part II (pp. 62-70). Berlin/Heidelberg, Germany: Springer-Verlag.

Karacapilidis, N., & Papadias, D. (2001). Computer supported argumentation and collaborative decision making: The HERMES system. *Information Systems*, *26*(4), 259–277. doi:10.1016/S0306-4379(01)00020-5

Krikke, H. R., Pappis, C. P., Tsoulfas, G. T., & Bloemhof-Ruwaard, J. (2002). Extended design principles for closed loop supply chains: Optimising economic, logistic and environmental performance. In A. Klose, M. G. Speranza, & L. N. Van Wassenhove (Eds.), *Quantitative approaches to distribution logistics and supply chain management* (pp. 61-74). Berlin/Heidelberg, Germany: Springer-Verlag.

Macey, R. (2005, March 10). Mobile phones kill great apes. *Sydney Morning Herald*.

Most, E. (2003). *Calling all cell phones: Collection, reuse, and recycling programs in the U.S.* INFORM Inc., New York, USA.

Pappis, C. P., Tsoulfas, G. T., & Dasaklis, T. (2006, July 2-5). *Assessing alternative reverse logistics policies for the recovery of used mobile phones.* Paper presented at the EURO XXI Conference, Reykjavik, Iceland.

Tsoulfas, G. T., Dasaklis, T., & Pappis, C. P. (2007, May 9-11). Identifying success factors for reverse logistics activities. In *Proceedings of the 2nd International Conference ECO-X 2007*, Vienna, Austria (pp. 39-46).

ADDITIONAL READING

Adamides, E., & Karacapilidis, N. (2006). A knowledge centred framework for collaborative business process modelling. *Business Process Management Journal*, *12*(5), 557–575. doi:10.1108/14637150610690993

Ahn, H. J., Lee, H. J., Chob, K., & Park, S. J. (2005). Utilizing knowledge context in virtual collaborative work. *Decision Support Systems*, *39*(4), 563–582. doi:10.1016/j.dss.2004.03.005

Álvarez-Gil, M. J., Berrone, P., Husillos, F. J., & Lado, N. (2007). Reverse logistics, stakeholders' influence, organizational slack, and managers' posture. *Journal of Business Research*, *60*(5), 463–473. doi:10.1016/j.jbusres.2006.12.004

Angell, L. C., & Klassen, R. D. (1999). Integrating environmental issues into the mainstream: An agenda for research in operations management. *Journal of Operations Management*, *17*(5), 575–598. doi:10.1016/S0272-6963(99)00006-6

Boons, F. (2002). Greening products: A framework for product chain management. *Journal of Cleaner Production*, *10*(5), 495–505. doi:10.1016/S0959-6526(02)00017-3

Bowen, F. E., Cousins, P. D., Lamming, R. C., & Faruk, A. C. (2001). The role of supply management capabilities in green supply. *Production and Operations Management*, *10*(2), 174–189.

Chouinard, M., D'Amours, S., & Aït-Kadi, D. (2005). Integration of reverse logistics activities within a supply chain information system. *Computers in Industry*, *56*(1), 105–124. doi:10.1016/j.compind.2004.07.005

Courtney, J. (2001). Decision making and knowledge management in inquiring organizations: Toward a new decision-making paradigm for DSS. *Decision Support Systems*, *31*(1), 17–38. doi:10.1016/S0167-9236(00)00117-2

Dowlatshahi, S. (2000). Developing a theory of reverse logistics. *Interfaces*, *30*(3), 143–155. doi:10.1287/inte.30.3.143.11670

Eichner, T., & Pethig, R. (2001). Product design and efficient management of recycling and waste treatment. *Journal of Environmental Economics and Management*, *41*(1), 109–134. doi:10.1006/jeem.2000.1126

Evangelou, C., Karacapilidis, N., & Tzagarakis, M. (2006). On the development of knowledge management services for collaborative decision making. *Journal of Computers*, *1*(6), 19–28.

Fleischmann, M., Krikke, H. R., Dekker, R., & Flapper, S. D. P. (2000). A characterization of logistics networks for product recovery. *Omega*, *28*(6), 653–666. doi:10.1016/S0305-0483(00)00022-0

Forgionne, G., Gupta, J., & Mora, M. (2002). Decision making support systems: Achievements, challenges and opportunities. In M. Mora, G. Forgionne, & J. Gupta (Eds.), *Decision making support systems: Achievements and challenges for the new decade* (pp. 392-403). Hershey, PA: Idea Group.

Georgiadis, P., & Vlachos, D. (2004). The effect of environmental parameters on product recovery. *European Journal of Operational Research*, *157*(2), 449–464. doi:10.1016/S0377-2217(03)00203-0

Gungor, A., & Gupta, S. M. (1999). Issues in environmentally conscious manufacturing and product recovery: A survey. *Computers & Industrial Engineering*, *36*(4), 811–853. doi:10.1016/S0360-8352(99)00167-9

Handfield, R., Sroufe, R., & Walton, S. (2005). Integrating environmental management and supply chain strategies. *Business Strategy and the Environment*, *14*(1), 1–19. doi:10.1002/bse.422

Karacapilidis, N. (2006). An overview of future challenges of decision support technologies. In J. Gupta, G. Forgionne, & M. Mora (Eds.), *Intelligent decision-making support systems: Foundations, applications and challenges* (pp. 385-399). London: Springer-Verlag.

Khanna, M., & Anton, W. Q. (2002). Corporate environmental management: Regulatory and market-based pressures. *Land Economics*, *78*(4), 539–558. doi:10.2307/3146852

Liebowitz, J., & Megbolugbe, I. (2003). A set of frameworks to aid the project manager in conceptualizing and implementing knowledge management initiatives. *International Journal of Project Management*, *21*(3), 189–198. doi:10.1016/S0263-7863(02)00093-5

Malhotra, Y. (2005). Integrating knowledge management technologies in organizational business processes: Getting real time enterprises to deliver real business performance. *Journal of Knowledge Management*, *9*(1), 7–28. doi:10.1108/13673270510582938

Mora, M., Forgionne, G., Gupta, J., Cervantes, F., & Gelman, O. (2003, September 3-5). A framework to assess intelligent decision-making support systems. In V. Palade, R. J. Howlett, & L. C. Jain (Eds.), *Knowledge-based intelligent information and engineering systems: Proceedings of the 7th International Conference, KES 2003, Oxford, UK, Part II* (pp. 59-65). Berlin/Heidelberg: Springer-Verlag.

Muckstadt, J., Murray, D., Rappold, J., & Collins, D. (2001). Guidelines for collaborative supply chain system design and operation. *Information Systems Frontiers*, *3*(4), 427–453. doi:10.1023/A:1012824820895

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science*, *5*(1), 14–37. doi:10.1287/orsc.5.1.14

Provis, C. (2004). Negotiation, persuasion and argument. *Argumentation*, *18*(1), 95–112. doi:10.1023/B:ARGU.0000014868.08915.2a

Ravi, V., & Shankar, R. (2005). Analysis of interactions among the barriers of reverse logistics. *Technological Forecasting and Social Change*, *72*(8), 1011–1029. doi:10.1016/j.techfore.2004.07.002

Richey, R. G., Chen, H., Genchev, S. E., & Daugherty, P. J. (2005). Developing effective reverse logistics programs. *Industrial Marketing Management*, *34*(8), 830–840. doi:10.1016/j.indmarman.2005.01.003

Rogers, D. S., & Tibben-Lembke, R. S. (1999). *Going backwards: Reverse logistics trends and practices.* Pittsburgh: Reverse Logistics Executive Council.

Shapiro, J. (2001). Modeling and IT perspectives on supply chain integration. *Information Systems Frontiers*, *3*(4), 455–464. doi:10.1023/A:1012876804965

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present and future of decision support technology. *Decision Support Systems*, *33*(2), 111–126. doi:10.1016/S0167-9236(01)00139-7

Tibben-Lembke, R. S., & Rogers, D. S. (2002). Differences between forward and reverse logistics in a retail environment. *Supply Chain Management*, *7*(5), 271–282. doi:10.1108/13598540210447719

Tsoulfas, G. T., & Pappis, C. P. (2006). Environmental principles applicable to supply chains design and operation. *Journal of Cleaner Production*, *14*(18), 1593–1602. doi:10.1016/j.jclepro.2005.05.021

This work was previously published in Web-Based Green Products Life Cycle Management Systems: Reverse Supply Chain Utilization, edited by H. Wang, pp. 401-415, copyright 2005 by Information Science Reference (an imprint of IGI Global).

Chapter 3.5
WSBen:
A Web Services Discovery and Composition Benchmark Toolkit[1]

Seog-Chan Oh
General Motors R&D Center, USA

Dongwon Lee
The Pennsylvania State University, USA

ABSTRACT

In this article, a novel benchmark toolkit, WSBen, for testing web services discovery and composition algorithms is presented. The WSBen includes: (1) a collection of synthetically generated web services files in WSDL format with diverse data and model characteristics; (2) queries for testing discovery and composition algorithms; (3) auxiliary files to do statistical analysis on the WSDL test sets; (4) converted WSDL test sets that conventional AI planners can read; and (5) a graphical interface to control all these behaviors. Users can fine-tune the generated WSDL test files by varying underlying network models. To illustrate the application of the WSBen, in addition, we present case studies from three domains: (1) web service composition; (2) AI planning; and (3) the laws of networks in Physics community. It is our hope that WSBen will provide useful insights in evaluating the performance of web services discovery and composition algorithms. The WSBen toolkit is available at: http://pike.psu.edu/sw/wsben/.

INTRODUCTION

A *Web Service* is a set of related functionalities that can be loosely coupled with other services programmatically through the Web. Examples of web applications using Web services include weather forecasting, credit check, and travel agency programs. As a growing number of Web services are available on the Web and in organizations, finding and composing the right set of Web services become ever more important. As a result, in recent years, a plethora of research work and products on Web-service discovery and composition problems have appeared[2]. In addition, the Web service research community has hosted open competition programs (e.g., ICEBE05[3], EEE06[4]) to solicit algorithms and software to discover pertinent Web services and compose them to make value-added functionality. Despite all this attention, however, there have been very few test environments available for evaluating such algorithms and software. The lack of such a testing environment with flexible features hinders the development of new composition algorithms and validation of the proposed ones. Therefore, the need for a benchmark arises naturally to evaluate and compare algorithms and software for the Web-service discovery and composition problems. As desiderata for such a benchmark, it must have (a large number of) web services in the standard-based WSDL files and test queries that can represent diverse scenarios and situations that emphasize different aspects of various Web-service application domains. Often, however, test environments used in research and evaluation have skewed test cases that do not necessarily capture real scenarios. Consider the following example.

Example 1 (Motivating) *Let us use the following notations: A Web service $w \in W$, specified in a WSDL file, can be viewed as a collection of operations, each of which in turn consists of input and output parameters. When an operation op has input parameters $op^i = \{p_1,...,p_n\}$ and output parameters $op^o = \{q_1,...,q_n\}$, we denote the operation by $op(op^i, op^o)$. Furthermore, each parameter is viewed as a pair of (name, type). We denote the name and type of a parameter p by p.name and p.type, respectively. For the motivating observation, we first downloaded 1,544 raw WSDL files that Fan and Kambhampati (2005) gathered from real-world Web services registries such as XMethods or BindingPoint. We refer to the data set as PUB06. For the purpose of preprocessing PUB06, first, we conducted WSDL validation according to WSDL standard, where 874 invalid WSDL files are removed and 670 files are left out. Second, we removed 101 duplicated WSDL files at operation level, yielding 569 valid WSDL files. Finally, we conducted type flattening and data cleaning processes subsequently. The type flattening process is to extract atomic types from user-defined complex types using type hierarchy of XML schema. This process helps find more compatible parameter faster. Details are found in (Kil, Oh, & Lee, 2006). The final step is the data cleansing to improve the quality of parameters. For instance, substantial number of output parameters (16%) was named "return", "result", or "response" which is too ambiguous for clients. However, often, their more precise underline meaning can be derived from contexts. For instance, if the output parameter named "result" belongs to the operation named "getAddress'", then the "result" is in fact "Address". In addition, often, naming follows apparent pattern such as getFooFromBar or searchFooByBar. Therefore, to replace names of parameters or operations by more meaningful ones, we removed spam tokens like "get" or "by" as much as we could.*

We measured how many distinct parameters each WSDL file contained. Suppose that given a parameter $p \in P$, we denote the number of occurrences of p.name as #(p.name). That is, #("pwd") indicates the number of occurrences of the parameter with name of "pwd". Figure 1 illustrates #(p.name) distributions of PUB06 and the ICEBE05 test set, where the x-axis is #(p.name) and the y-axis is the number of parameters with the same #(p.name) value. The distribution of PUB06 has no humps. We also plotted a power-function, over the #(p.name) distribution, and found that the exponent is 1.1394. Although 1.1394 does not suffice the requirement to be the power law (Denning, 2004), the distribution is skewed enough to be seen as the Zipf-like distribution. Indeed, the parameters such as "license key", "start date", "end date," or "password" have a large #(p.name) value, while most parameters appear just once. This observation also implies the existence of hub parameters, which appear in Web services frequently, and serve important roles *on the inter-connections between Web services. On the contrary, the distribution of ICEBE05 test set has four humps equally at around 1, 100, 200, and 800 with the highest value at third hump. This distribution shape differs considerably from PUB06, the real public Web services. This implies that the test environments of ICEBE05 do not necessarily capture characteristics of real Web services.*

In conclusion, as demonstrated in the example, our claim is that *any Web-services discovery and composition solutions must be evaluated under diverse configurations of Web services networks* including two cases of Figure 1. However, to our best knowledge, there have been no publicly available benchmark tools. To address these needs and shortcomings, therefore, we developed the *WSBen* - a Web-Service discovery and composition Benchmark tool. The main contributions of WSBen is to provide diverse Web service test sets based on three network models such as "random", "small-world", and "scale-free" types. These three network models have been shown to model many real-world networks sufficiently (Albert & Barabasi, 2002). We also present three use cases in different communities to demonstrate the application of WSBen. In addition, we propose a flexible framework, by which we can study real Web service networks, and establish the design

Figure 1. #(p.name) distributions. (left) PUB05. (right) ICEBE05.

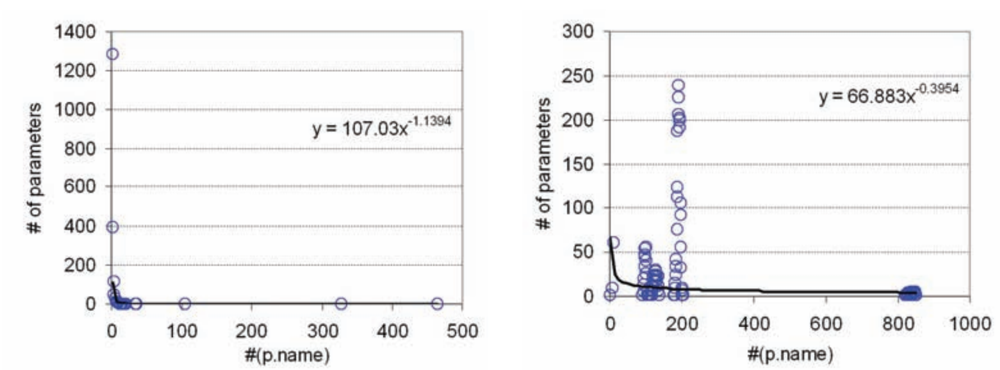


Table 1. Summary of notation

Notation	*Meaning*
w, W	Web service, set of Web services
p, P	Parameter, set of parameters
r, r^i, r^o	Request, initial and goal parameter sets of r
$G_p(V_p, E_p)$	Parameter node network
$G_{op}(V_{op}, E_{op})$	Operation node network
$G_{sw}(V_{sw}, E_{sw})$	Web service node network
$G_{op}^f(V_{op}^f, E_{op}^f)$	Full-matching operation node network
$G_{c1}(V_{c1}, E_{c1})$	Parameter cluster network
$g_{r^i}(p)$	Minimum cost of achieving $p \in P$ from r^i in G_p
xTS	WSBen's 5-tuple input framework (e.g., *baTS*, *erTS* and *nwsTS* are instances)

foundation of WSBen. As a whole, this article is based on two of our earlier works (Kil et al., 2006; Oh, Kil, Lee, & Kumara, 2006). Extended from the previous works, this article describes how WSBen is designed and works to generate test sets from the software architecture perspective, and furthermore introduces three use cases to highlight the practical benefits of WSBen. Table 1 summarizes important notations used in this article.

This article is organized as follows. First, in the background section, we review concepts and techniques required for the WSBen development, especially focusing on the complex network theory. Second, in the related works section, we discuss related studies in the literature as well as surveying existing world-wide challenges with regard to Web services and Semantic Web. Third, in the overview of WSBen section, we present WSBen with its design concept, test set generation mechanism, key functions and characteristics. Fourth, in the use cases of WSBen section, we illustrate how WSBen can be exploited to obtain research benefits, especially by demonstrating three use cases. We expect three use cases enough to provide vigorous experiments and evaluation of our WSBen. Finally, conclusions are provided.

BACKGROUND

In this section, we review prerequisite techniques and concepts required to build WSBen. First, we revisit the definition and complexity of Web-service discovery and composition problems. Second, we introduce three complex network topologies based on which WSBen is designed to populate WSDL test files. Finally, we explain our conceptual methodology to project a bipartite Web-service network consisting of three distinct nodes (parameter, operation, and Web service) and heterogeneous arc types into three distinct Web-service networks, each of which consists of single node and uniform arc. The main benefit of projecting Web-service networks is that it can allow for straightforward analysis on referred network's characteristics. Throughout this article, we will use our conceptual Web-service network concept in order to analyze real public Web-service networks as well as WSDL test file sets generated by WSBen.

Web-Service Discovery and Composition

Suppose that a Web service w has one operation so that a Web service can be considered as an operation, and input and output parameter sets of w are denoted by w^i and w^o, respectively. When one has a request r that has initial input parameters r^i and desired output parameters r^o, one needs to find a Web service w that can fulfill such that (1) $r^i \supseteq w^i$ and (2) $r^o \subseteq w^o$. Finding a Web service that can fulfill r alone is referred to as *Web-service discovery* (WSD) problem. When it is impossible for one Web service to fully satisfy r, on the other hand, one has to compose multiple Web services $\{w_1, w_2,...,w_n\}$, such that (1) for all $w_k \in \{w_1, w_2,...,w_n\}$, w_k^i can be applicable when w_k^i is required at a particular stage in composition, and (2) $(r^i \cup w_1^o \cup w_2^o \cup ... \cup w_n^o) \supseteq r^o$. This problem is often called as *Web-service composition* (WSC) problem. In addition, one can also consider different matching schemes from the operation perspective – "partial" and "full" matching. In general, given w_1 and w_2, if w_1 can be invoked at the current information state and $w_1^o \supseteq w_2^i$, then w_1 can "fully" match w_2. On the other hand, if w_1 cannot fully match w_2, but w_1 can match a subset of w_2, then w_1 can "partially" match w_2. When only full matching is considered in the WSC problem, it can be seen as a single-source shortest path problem whose computational complexity is known as polynomial (Bertsekas, 2000). On the other hand, when both full and partial matching must be considered concurrently, the problem becomes a decision problem to determine the existence of a solution of k operators or less for propositional STRIPS planning, with restrictions on negation in pre- and post-conditions (Nilsson, 2001). Its computational complexity is proved to be NP-complete (Bylander, 1994). Therefore, when the number of Web services to search is not small, finding an optimal solution to the WSC problem (i.e., a chain of Web services to invoke) is prohibitively expensive, leading to approximate algorithms instead.

Complex Network Models

There are many empirical systems to form complex networks such as the scale-free network and the small-world network, in which nodes signify the elements of the system and edges represent the interactions between them.

Definition 1 (Random network) *A network is defined as the random network on N nodes, if each pair of nodes is connected with probability p. As a result, edges are randomly placed among a fixed set of nodes. The random network can be constructed by means of the Erdos-Renyi's random-graph model (Erdos, Graham, & Nesetril, 1996).*

Definition 2 (Regular network) $Rg_{(N,k)}$ *is defined as the regular network on N nodes, if node i is adjacent to nodes* $[(i+j) \bmod N]$ *and* $[(i-j) \bmod N]$ *for* $1 \le j \le k$*, where k is the number of valid edge of each node. If* $k = N - 1$, $Rg_{(N,k)}$ *becomes the complete N-nodes graph, where every node is adjacent to all the other* $N - 1$ *nodes.*

We can define some metrics to quantify the characteristic properties of the complex networks as follows:

- L: The average shortest distance between reachable pairs of nodes, where the distance between nodes refers to the number of hops between the nodes. $L(p)$ is defined as L of the randomly rewired Watts-Strogatz graph (Watts & Strogatz, 1998) with probability p. L_{random} is identical to $L(1)$.
- C: The average clustering coefficient. Suppose that for a node i with v_i neighbor,

  $$C_i = \frac{2E_i}{v_i(v_i - 1)},$$

  where E_i is the number of edges between v_i neighbors of i. C is the average clustering coefficient C_i for a network. $C(p)$ is defined as C of the randomly rewired Watts-Strogatz graph with probability p. C_{random} is identical to $C(1)$.

Definition 3 (Small-world network) *Small-world networks are characterized by a highly clustered topology like regular lattices and the small network diameter, where the network diameter suggests the longest shortest distance between nodes. Specifically, small-world networks are* $C \succ C_{random}$ *and* $L \approx L_{random}$ *(Delgado, 2002).*

By using the Watts-Strogatz model (Watts, 1999; Watts & Strogatz, 1998), we can construct networks that have the small-world properties. The model depends on two parameters, connectivity (k) and randomness (p), given the desired size of the graph (N). The Watts-Strogatz model starts with a $Rg_{(N,k)}$ and then every edge is rewired at random with probability p; for every edge (i, j), we decide whether we change j node (the destination node of (i, j)) with probability p. The Watts-Strogatz model leads to different graphs according to the different p as follows:

- When $p = 0$, an $Rg_{(N,k)}$ is built.
- When $p = 1$, a completely random network is built.

Otherwise, with $0 < p < 1$, each edge (i, j) is reconnected with probability p to a new node k that is chosen at random (no self-links allowed). If the new edge (i, k) is added, the (i, j) is removed from the graph. The long-range connections (short-cuts) generated by this process decrease the distance between the nodes. For intermediate values of p, there is the "small-world" region, where the graph is highly clustered yet has a small average path length.

Definition 4 (Scale-free network) *Networks are called scale-free networks if the number of nodes that have v number of neighbor nodes is proportional to* $P_w(v) \propto v^{(-\gamma)}$, *where* γ *is typically greater than two with no humps.*

Barabasi and Albert provided several extended models (Albert, Jeong, & Barabasi, 1999; Delgado, 2002) to provide the scale-free properties. The extended model uses an algorithm to build graphs that depend on four parameters: m_0 (initial number of nodes), m(number of links added and/or rewired at every step of the algorithm), p (probability of adding links), q (probability of edge rewiring). The procedure starts with m_0 isolated nodes and performs one of the following three actions at every step:

- With the probability of p, $m(\leq m_0)$ new links are added. The two nodes are picked randomly. The starting point of the link is chosen uniformly, and the end point of the new link is chosen according to the following probability distribution:

$$\Pi_i = \frac{v_i + 1}{\sum_j (v_j + 1)} \qquad (1)$$

  where $\prod_i$ is the probability of selecting the i node, and v_i is the number of edges of node i.

  The process is repeated m times.
- With the probability of q, m edges are rewired. For this purpose, i node and its link l_{ij} are chosen at random. The link is deleted. Instead, another node z is selected according to the probabilities of Equation (1), and the new link l_{iz} is added.
- With the probability of $1 - p - q$, a new node with m links is added. These new links connect the new node to m other nodes chosen according to the probabilities of Equation (1).
- Once the desired number N nodes are obtained, the algorithm stops. The graphs generated by this algorithm are scale-free graphs, and the edges of the graphs are constructed such that the correlations among edges do not form. When $p = q$, the algorithm results in a graph, whose connectivity distribution can be approximated by

$$P(v) \propto (v+1)^{-\frac{2m(1-p)+1-2p}{m}+1} \qquad (2)$$

  where v is the number of edges.

Diverse Web Service Network Models

A set of Web services form a network (or graph). Depending on the policy to determine nodes and edges of the network, there are varieties: Web service level (i.e., coarse granularity), operation level, and parameter level (i.e., fine granularity) models. The graph at the middle of Figure 2 has a bipartite graph structure and consists of three distinct kinds of vertices (i.e., parameter, operation, and web-service node) and directed arcs between bipartite nodes (i.e., operation nodes and parameter nodes). An edge incident from a parameter node to an operation node means that the parameter is one of the inputs of the corresponding operation. Reversely, an edge incident from an operation node to a parameter means that the parameter is one of the outputs of the corresponding operation. The graph has two Web services, labeled `WS1` and `WS2`. `WS1` has two operations `Op11` and `Op12`, and `WS2` has one operation, `Op21`, respectively. There are seven parameters, labeled `P1` to `P7`. According to the node granularity, we can project the upper graph into three different Web service networks.

- Parameter-Node Network: A graph $G_p(V_p, E_p)$, where V_p is a set of all parameter nodes and E_p is a set of edges. An edge (p_i, p_j) is directly incident from input parameters $p_i \in V_p$ to output parameters $p_j \in V_p$, where there is an operation that has an input parameter matching p_i and an output parameter matching p_j. For example, `P1→Op11→P3` in the upper graph is projected into `P1→P3` in the parameter node network. Figure 3 shows the parameter node network for PUB06 and the ICEBE05 test set.
- Operation-Node Network: A graph $G_{op}(V_{op}, E_{op})$, where V_{op} is a set of all operation nodes, and E_{op} is a set of edges. An edge (op_i, op_j) is incident from operation $op_i \in V_p$ to operation $op_j \in V_p$, here op_i can fully or partially match op_j. For example, `Op11` partially matches `Op12` which, in turn, fully matches `Op21` in the upper graph. Therefore, `Op11→Op12→Op21` can be shown in the operation node network. In particular, the fully matching operation node network, G^f_{op} has only `Op12→Op21`.
- Web-service Node Network: A graph $G_{ws}(V_{ws}, E_{ws})$, where V_{ws} is a set of all web-service nodes, and E_{ws} is a set of edges. An edge (ws_i, ws_j) is incident from web-service node $ws_i \in V_{ws}$, to $ws_j \in V_{ws}$, where there is at least one edge between any operation in ws_i and any operation in ws_j. For example, since `WS1` possesses `Op12` and `WS2` possesses `Op21` in the upper graph, `WS1→WS2` appears in the Web service node network.

RELATED WORKS

Constantinescu, Faltings, and Binder (2004)

Figure 2. Web services networks: (a) WSDLs, (b) Conceptual Web service network, (c) Web service networks from diverse models, (d) Parameter node network, G_p, (e) Operation node network, G_{op}, (f) Fully invocable operation node network, G^f_{op}, and (g) Web service node network, G_{ws}

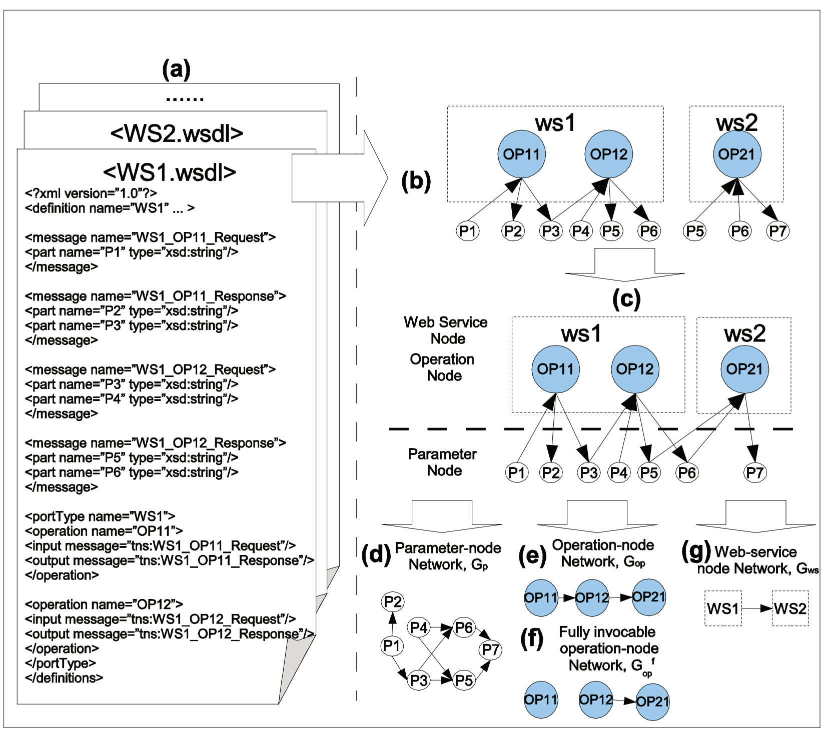
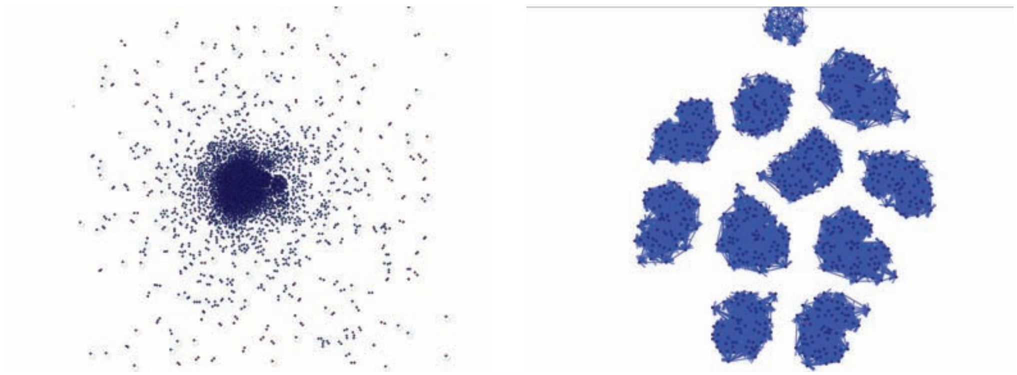


Figure 3. Diverse parameter networks. (left) PUB05. (right) ICEBE05



proposed a scalable syntactic test bed where Web services are generated as transformation between sets of terms in two application domains. For doing that, they first defined parameter sets corresponding to application domains and then, connected those parameter sets randomly and constructed a service graph which structure (i.e., nodes and arcs) is similar to the parameter cluster network of WSBen. However, WSBen takes a significant different approach to construct its parameter cluster networks in that WSBen does not simply connect parameter sets at random but simulates topologies of real Web service networks. WSBen is inspired by extensive studies on real Web services, and therefore is designed to support various Web service network topologies and distributions. As a result, WSBen can present more realistic testing situation for researchers who want to test their Web service discovery or composition algorithms than that of Constantinescu et al. (2004).

XMark (XMark, 2006) is an XML benchmark suite that can help identify the list of functions which an ideal benchmark should support. WSBen uses XMark as a reference model to identify necessary functions to simplify the testing process. One feature that is offered by XMark but not by WSBen is the provision of solutions to queries. In other words, XMark provides queries and their corresponding solutions but WSBen gives requests only because the optimal solution to a Web service composition problem may not be obtained in the reasonable time window due to the problem's inherently high complexity.

There are three unique challenges that have been established to investigate research issues with regard to Web services and Semantic Web. First is the Semantic Web Services Challenge[5]. This venue invites application submissions for demonstrating practical progress towards achieving the vision of the Semantic Web. According to the event, it has the overall goal to advance our understanding of how semantic technologies can be exploited to produce useful applications for the Web. Second is the Web Services Challenge[6]. This venue solicits approaches, methods, and algorithms in the domain of Web-service discovery and composition. This event evaluates participants' approaches based on their quantitative and qualitative performance results on discovery and composition problems. The Web Services Challenge is more driven by common problems, but the Semantic Web Challenge concentrates more on the environment. As such, the Semantic Web Challenge places more focus on semantics while the Web Services Challenges favors applied and short-term solutions (Brian, William, Michael, & Andreas, 2007). Third is the Service Oriented Architecture Contest[7] which asks participants to openly choose particular domain-specific problems and show their best approaches for them. There are unique characteristics for each venue so that they have undoubtedly contributed to advance the state-of-art technologies in Web services and Semantic Web. Among these challenges, WSBen can be exploited especially for the Web Services Challenge to provide various benchmark environments, discovery and composition problems by varying Web-service network topologies.

As for WSC, there are two main approaches, depending on the use of domain knowledge. First, the template-workflow based approach is to use software programs and domain experts to bind manually-generated workflows to the corresponding concrete Web services. METEOR-S (Sivashanmugam, Verma, Sheth, & Miller, 2003) is an example of this approach. Second, various AI planning techniques have been applied to the WSC problem, ranging from simple classical STRIPS-style planning to an extended estimated regression planning (McDermott, 2002). We believe that our WSBen is complementary for AI Planning based tools for the WSC problem. In fact, we demonstrate how WSBen can be used to compare the performance of AI planners for the WSC problems in the illustrative use-cases section. In this article, meanwhile, we do not propose how METEOR-S can make use of WSBen for a test case generation tool. It is because METEOR-S consists of three modules such as process designer, configuration module, and execution environment, where the execution environment requires executable Web services but WSBen can generate only WSDL files without real implementation.

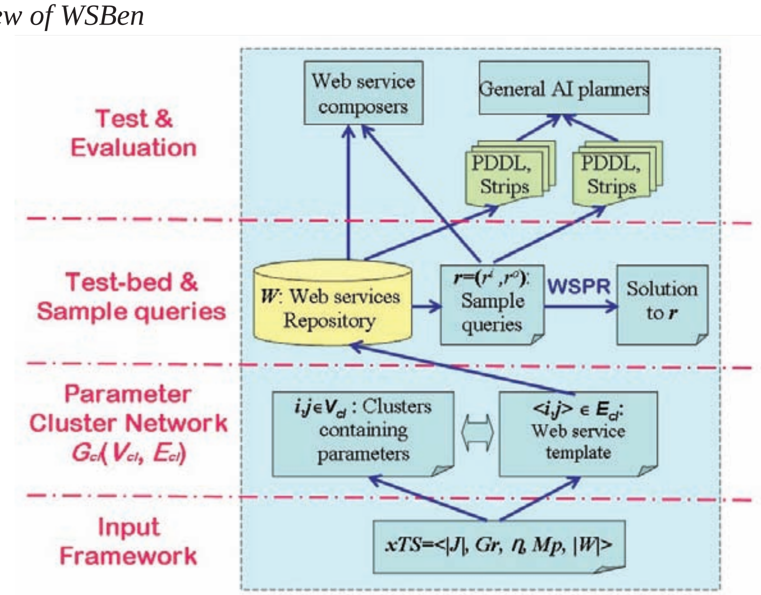
Overview of WSBen

In this section, we present a novel benchmark tool titled WSBen, which provides a set of functions to simplify the generation of test environments for WSD and WSC problems.

Overview of WSBen

At a higher level, a Web service can be assumed as a transformation between two different application domains, and each can be represented by a cluster. This assumption was the basis in developing WSBen. From the perspective of graph theory, WSBen builds *Parameter Cluster Network*, which consists of clusters and directed edges connecting two different clusters. These directed edges become Web service templates from which WSBen generates Web services as users specify. Formally, the parameter cluster

Figure 4. Overview of WSBen



network is defined as follows:

Definition 5 (Parameter Cluster Network) *A directed graph $G_{cl}(V_{cl}, E_{cl})$, where V_{cl} is a set of clusters and E_{cl} is a set of directed edges that are incident from input clusters $i \in V_{cl}$ to output clusters $j \in V_{cl}$. Here, cluster i and j contain a set of non-overlapping parameters denoted by Pa_i and Pa_j, respectively, where $Pa_i \cap Pa_j = \phi$. Each directed edge is also called a Web service template, from which WSDL files are generated.*

Figure 4 shows the overview of WSBen. In detail, WSBen consists of the following functionalities:

- Input framework: Users can specify and control the generated synthetic WSDL files and their characteristics. For this purpose, WSBen provides an input framework *xTS* = <$|J|$, G_r, η, G_p,$|W|$>. *xTS* applies existing complex and random network models to specify G_r. Each element of *xTS* will be discussed in more detail below.
- Parameter cluster network, $G_{cl}(V_{cl}, E_{cl})$: If *xTS* is given by users, based on the first four elements, WSBen generates G_{cl}. Each cluster of G_{cl} is filled with some number of atomic parameters. In this network, Web services are defined as transformations between two different clusters. That is, <*i*, *j*> $\in E_{cl}$ becomes Web service templates. The role of Web service templates in the test set generation will be illustrated.
- Test set and sample requests: By randomly selecting the Web service templates (arcs of the parameter cluster network), WSDL files are generated. Once a test set is generated, users can generate sample test requests r = <r^i, r^o>. The generation process of test sets and test requests will be illustrated.
- Test and evaluation: WSBen can export both the Web service WSDL files and test requests into files in PDDL (McDermott, 1996) and STRIPS format, enabling concurrent comparison with state-of-the-art AI planners.

WSBen input framework: ***xTS***

WSBen input framework, *xTS* consists of five tuples, <$|J|$, G_r, η, G_p,$|W|$>. In detail:

1. $|J|$ is the total number of parameter clusters.
2. G_r denotes a graph model to specify the underlying topology of a parameter cluster network. G_r can be on of the following three models discussed in the Background section:
   - *Erdos-Renyi*($|J|$, *p*): This model has such a simple generation approach that it chooses each of the possible

     $$\frac{|J|(|J|-1)}{2}$$

     edges in the graph with $|J|$ nodes with probability *p*. The resulting graph becomes the same as the binomial graph.
   - *Newman-Watts-Strogatz*($|J|$, *k*, *p*): The initialization is a regular ring graph with *k* neighbors. During the generation process, new edges (shortcuts) are added randomly with probability *p* for each edge. Note that no edges are removed, differing from Watts-Strogatz model.
   - *Barabasi-Albert*($|J|$, *m*): This graph model is generated by adding new nodes with *m* edges that are preferentially attached to existing nodes with a high degree. The initialization is a graph with *m* nodes and no edges. Note that the current implementation of WSBen is limited because it can only generate the simplified version of the extended Barabai-Albert model, by setting $p = q = 0$ and $m_0 = m$, resulting in graphs with $\gamma = 2.0 \pm 0.1$, where γ is the exponent of a power function $P_w(v)$ defined over connectivity *v* range in the form of $P_w(v) \propto v^{-\gamma}$.

3. η denotes the parameter condense rate. With η, users can control the density of partial-matching cases in produced Web services.
4. M_p denotes the minimum number of parameters a cluster can contain. In other words, clusters may have a different number of parameters but all clusters must have at least M_p number of parameters.
5. $|W|$ denotes the total number of Web services of a test set.

With $|J|$ and G_r, the first two input elements of *xTS*, we can build G_{cl} with each empty cluster. Thus, we need a procedure to fill each empty cluster with parameters. For this purpose, WSBen uses the following procedure:

1. A parameter cluster network G_{cl} with empty clusters is built by specifying $|J|$ and G_r.
2. Co-occurrence probability of each cluster is measured by the following probability:

$$\Delta_j = \frac{k_j}{\max_{j \in V_{cl}} k_j}\eta \qquad (3)$$

where Δ_j is the co-occurrence probability of cluster j, and k_j is the edge degree of cluster j.

η is the parameter condense rate which is given by users.

3. $|Pa_j|$ is measured based on the following equation.

$$|Pa_j| = \frac{M_p}{\Delta_j} \qquad (4)$$

where Pa_j is the set of parameters contained in cluster j.

4. For each j cluster, atomic parameters are generated up to $|Pa_j|$, with duplicated parameters forbidden (i.e., $\forall i, j \in V_{cl}$, $Pa_i \cap Pa_j = \phi$).

Once a complete parameter cluster network, $G_{cl}(V_{cl}, E_{cl})$ is built, WSBen repeats the following procedure until $|W|$ number of Web services are generated:

1. A Web service template $<i, j>$ is chosen at random from E_{cl}.
2. WSBen generates a WSDL file, in which each input parameter is selected from i cluster with probability Δ_i, and each output parameter is selected from j cluster with probability Δ_j.

Figure 5 illustrates how WSBen builds G_{cl} and generates WSDL files based on the G_{cl}. Suppose that $xTS = <8, Barabasi-Albert\ (8, 2), 0.8, 1.5, 100>$ is given. Then, the generation steps are as follows:

1. WSBen generates a graph of *Barabasi-Albert*(8,2). The direction of each edge is determined at random.
2. Δ_j and $|Pa_j|$ are specified. For example, Cluster 5 has nine parameters as shown in Figure 5. That is, $|Pa_5| = 9$, as

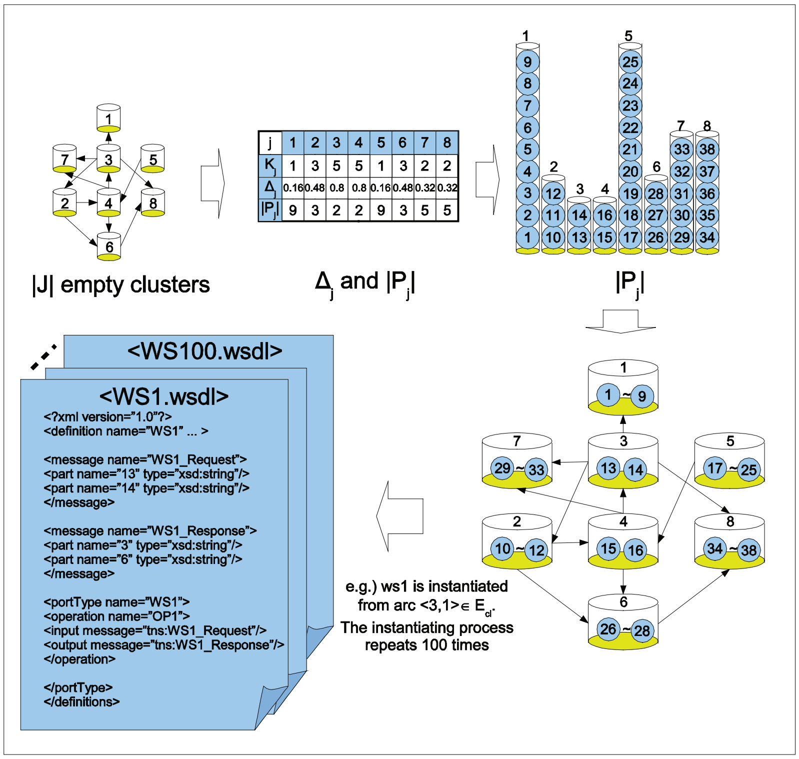
$$\Delta_5 = \frac{k_j}{\max_{j \in V_{cl}} k_j}\eta = \frac{1}{5} \times 0.8 = 0.16,$$

resulting in

$$|Pa_5| = \frac{M_p}{\Delta_5} \approx 9.$$

3. Pa_j is specified. For example Pa_5={17,18,19,20,21,22,23,24,25} as shown in Figure 5 because $|Pa_5| = 9$ and for $\forall i, j \in V_{cl}$, $Pa_i \cap Pa_j = \phi$. Note that the parameter names are automatically generated, and thus do not contain any semantics.
4. Finally, G_{cl} is built and WSBen generates $|W|$ Web services. For example, in Figure 5, `WS1` is instanced from a Web service template $<3, 1> \in E_{cl}$. Note that $\Delta_1 = 0.16$ and $\Delta_3 = 0.8$. $\Delta_1 = 0.16$ suggests that the occurrence probability of each parameter in Cluster 1 has 0.16. Due to the low probability, only "1" and "9" are selected from Cluster 1. Similarly, $\Delta_3 = 0.8$ means that the occurrence probability of each parameter in Cluster 3 has 0.8. Due to the high probabil-

Figure 5. Test set generation with <8, *Barabasi – Albert* (8, 2), 0.8, 1.5, 100>



ity, all parameters in Cluster 3 that are "13" and "14" are selected. In the case where no parameter is generated, dummy parameters "S" and "T" are filled in the input and output parameters, respectively.

The state, $s \in S$ is a collection of parameters in $|P|$. Therefore, r^i and r^o are states. The test request r is constructed such that r^o is farthest away from r^i in a parameter space in terms of $g_{r^i}(p)$ - the cost of achieving $p \in P$ from a state r^i. To obtain $g_{r^i}(p)$, we propose following *Forward Searching* algorithm.

Forward Searching: $g_{r^i}(p)$ can be characterized by the solution of a recursive equation as follows:

$$g_{r^i}(p) = \min_{w \in Ow(p)} [c(w) + \max_{p' \in w^i} g_{r^i}(p')] \qquad (5)$$

where $c(w)$ is an invocation cost of a Web service, $w \in W$ and is assumed to be 1. $Ow(p)$ is a set of Web services: $Ow(p) = \{w \in W \mid p \in w^o\}$. At first, $g_{r^i}(p)$ is initialized to 0 if $p \in r^i$, and to ∞ otherwise. Then, the current information state s is set to r^i (Line 1 in Algorithm 1). We denote $\Omega(s)$ by a set of Web services $w \in W$ such that $w^i \subseteq s$. That is, w can be invoked or applicable in the state s.

Every time for $\forall w \in \Omega(s)$, each parameter $p \in w^o$ is added to s, and $g_{r^i}(p)$ is updated until $\Omega(s)$

Algorithm 1. Forward searching algorithm of WSBen

Test Request Generation

Input : r^i
Output: $g_{r^i}(p)$ for all p reachable from r^i
1: $s = r^i$; $C = \phi$; $d = 1$;
2: **while** $(\delta \neq \phi)$ **do**
3: $\delta = \{w \mid w \in \Omega(s), w \notin C\}$;
4: **for** p in $w^o (w \in \delta)$ **do**
5: **if** $g_{r^i}(p) = \infty$ **then**
6: $g_{r^i}(p) = d$; $s = s \cup \{p\}$;
7: $C = C \cup \delta$; d++;

stops to increase, meaning that this process ends with finding $g_{r^i}(p)$ for all parameters reachable from r^i (Lines 2-6 in Algorithm 1).

We can use Equation (5) to drive the lower bound of the optimal cost of WSC solutions. Note that the invocation cost of a Web service is assumed to be 1. Thus, the optimal cost of a WSC problem coincides with the minimum number of Web services required to solve the WSC problem. For a set of parameters A, we can regard the following cost function:

$$g_{r^i}^{\max}(A) = \max_{p \in A} g_{r^i}(p) \tag{6}$$

The cost of achieving a set of parameters cannot be lower than the cost of achieving each of the parameters in the set. Thus, $g_{r^i}^{\max}(A)$ is the lower bound of the optimal cost of achieving r^o from r^i.

Based on the forward searching algorithm, WSBen create a test request r, as follows:

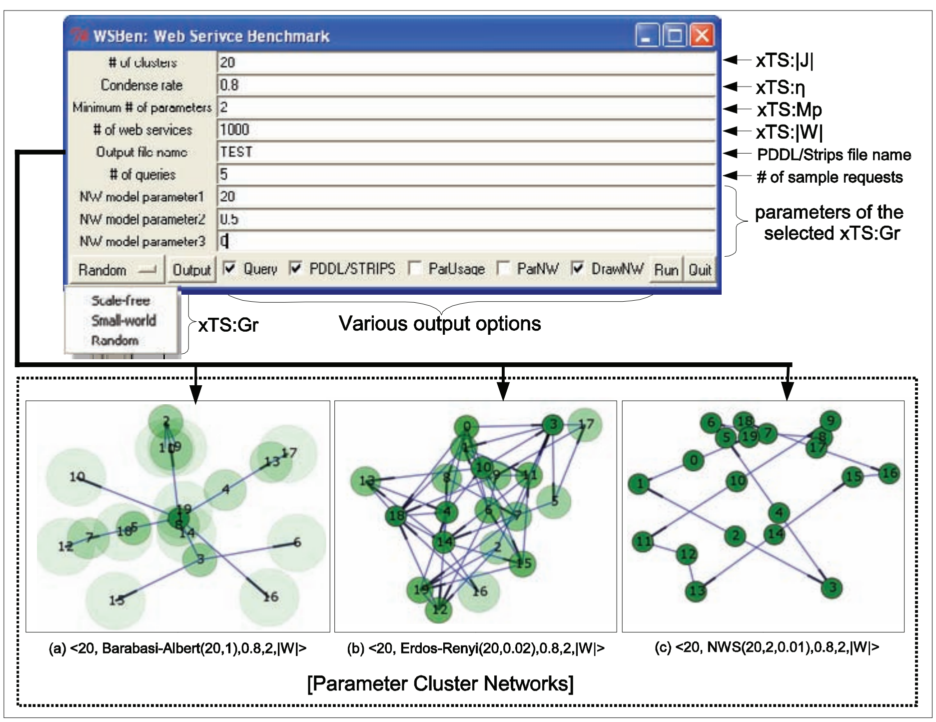
1. WSBen selects a Cluster $j \in G_{cl}$ at random.
2. WSBen copies all parameters in the Cluster j (i.e., Pa_j) into r^i, and then r^o is constructed so that it consists of the first five largest parameters in terms of $g_{r^i}(p)$. Consequently, parameters in r^o are farthest away from parameters in r^i in a parameter space.

As a default, WSBen repeats the above procedure five times, generating five request sets for each test set.

Implementation

As shown in Figure 6, WSBen provides user interfaces to specify *xTS* and several parameters, which are required to form a parameter cluster network and generate WSDL files. WSBen is implemented in Python, and run on Python 2.3 or later. It runs on Unix and Windows. For the creation, manipulation, and functions of complex networks, we used a Python package called NetworkX[8].

Figure 6. WSBen user interface



Current implementation of WSBen is limited as follows: (1) it supports only the exact matching without type compatibility check, and (2) each Web service contains only one operation so that a Web service can be viewed as equivalent to an operation. Therefore, w^i and w^o indicate the input and output parameter set of a Web service, *w*.

Figure 6 also shows three sample parameter cluster networks, where each circular node represents a cluster and edges with heads denote the Web service template, from which Web services are instanced. The size of node is proportional to the number of parameters in the node, while the transparency level of a node's color is inversely proportional to the degree of the node. For example, in the left cluster network, Cluster 18 can be considered a hub cluster in that it has the high degree. Therefore, it is presented by a small circle with denser color.

Following the mechanism of WSBen explained so far, we can build three illustrative test set frameworks by specifying *xTS* as follows:

1. *baTS* = <100, *Barabasi – Albert* (100, 6), 0.8, 5, |*W*|>
2. *nwsTS* = <100, *Newman – Watts – Strogatz*(100, 6, 0.1), 0.8, 5, |*W*|>
3. *erTS* = <100, *Erdos – Renyi*(100, 0.006), 0.8, 5, |*W*|>

Figures 7 and 8 show that there are distinctive differences between *baTS*, *nwsTS*, and *erTS* in terms of G_p and outgoing edge degree distribution.

ILLUSTRATIVE USE CASES OF WSBEN

In this section, we present three use cases to demonstrate the application of WSBen: (1) evaluating Web-service composition algorithms; (2) comparing performance of AI planners; and (3) estimating the size of giant component. These use cases are prepared to provide vigorous experiments and evaluation for assessing the usage of WSBen. For each use case, we will provide three Web-service test sets by varying *xTS* with three parameter cluster networks such as "random", "small-world", and "scale-free" types. Note that these three network models have the expression power enough to model many real-world networks sufficiently (Albert & Barabasi, 2002). This implies that our generated test cases can be appropriate for representing diverse real-world Web-service networks. Furthermore, these three Web-service test sets are significantly distinctive from each other in terms of their Web-service network topologies and degree distributions as we have shown in the previous section. This indicates that we have sufficient reason to analyze

Figure 7. G_p of baTS, erTS, and nwsTS when |W| = 1,000

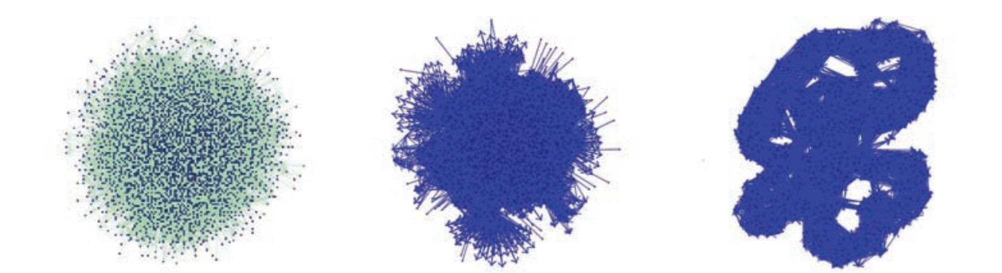


Figure 8. Outgoing edge degree of baTS, erTS, and nwsTS when |W| = 1,000

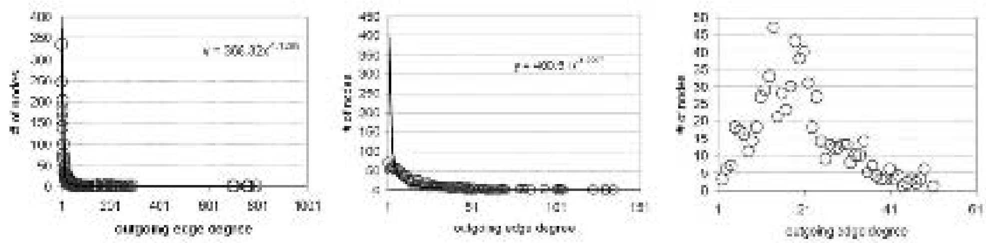
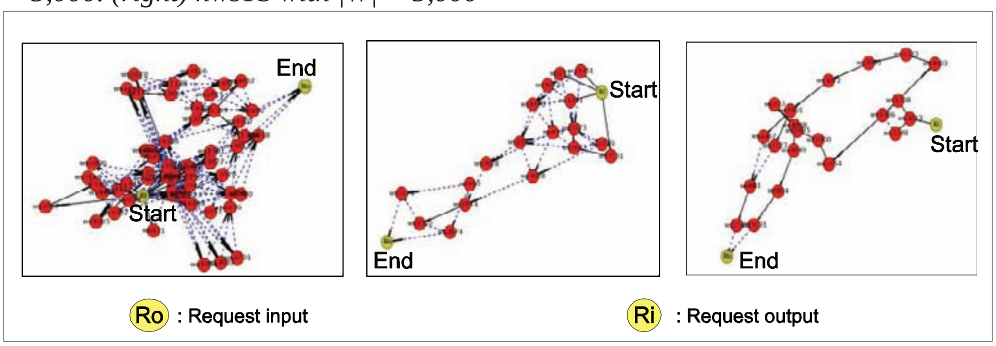


Figure 9. Composed services using WSPR for three test sets. (left) baTS with |W| = 5,000. (center) erTS with |W| = 5,000. (right) nwsTS with |W| = 5,000



how different network topologies can affect the performance of Web-service applications or environments.

Evaluating Web-Service Composition Algorithms

Recently, many WSC researches have been reported in the Web service community. As such, the EEE06 Web-service composition contest holds as many participants as 11. Among the 11 WSC algorithms, we choose a WSC algorithm named WSPR (Oh, Lee, & Kumara, 2007), which was proved effective and efficient in the contest, in order to demonstrate the application of WSBen.

In this case, we use three test sets generated by WSBen: (1) *baTS* with |*W*| = 5,000; (2) *erTS* with |*W*| = 5,000; and (3) *nwsTS* with |*W*| = 5,000. The resultant composed services generated by WSPR are shown in Figure 9, where WSPR addressed a request for each of the three test sets. Note that WSBen can automatically create sample requests for a given test set. In the graph, each composed solution has nodes such as "*Ri*" and "*Ro*", which represent the initial condition and goal state, respectively. Other nodes represent Web services. The directed arcs indicate the invocation flow, where a solid edge means full-matching invocation and a dotted edge represents partial-matching invocation. From the experiments based on diverse test sets such as *baTS*, *erTS*, and *nwsTS*, we can understand how different network models of G_{cl} influences the performance of WSC algorithms. In general, given the same number of clusters, the *Barabasi-Albert* model generates G_{cl} with a greater number of parameters, and a larger variance of the number of parameters between clusters than the *Newman-Watts-Strogatz* and *Erdos-Renyi* models do. Due to the greater number of parameters and larger variance, *baTS* needs more partial-matching Web services to fulfill the given requests than others. The increasing need for partial-matching Web services leads to increasing number of Web services in the composed service. This is the reason that the *baTS* case has more Web services to create a resultant composed service as shown in Figure 9 (left).

Comparing Performance of AI Planners

We demonstrate how WSBen can be used to compare the performance of AI planners. For this purpose, we choose three prominent AI planners – Graphplan (Blum & Furst, 1997), Blackbox (Kautz & Selman,1996), and IPP[9]. Blackbox and IPP are extended planning systems that originated from Graphplan. In particular, Blackbox is extended to be able to map a plan graph into a set of clauses for checking the satisfiability problem. Consequently, it can run even in large number of operators. For comparing the performance of three planners, we use two evaluation metrics as follows:

1. τ(Time): It measures how long an algorithm takes to find a right solution, in seconds. This is a measure of computational efficiency.
2. *#W*: The number of Web services in a right solution. This is a measure of effectiveness.

All AI planners run with their default options, except that the maximum number of nodes for Blackbox and Graphplan was set to 32,768 and 10,000, respectively. Commonly, the time to read operator and fact files is not included in τ. Blackbox and IPP accept only the PDDL format, while Graphplan accepts only the STRIPS format for their operator and fact files. Note that an operator file corresponds to a test set, and a fact file corresponds to a test request file. Also note that WSBen provides a function to convert test

Table 2. Results over baTS with |W| = 3,000

Requests	BlackBox		GraphPlan		IPP	
	#W	τ	*#W*	τ	*#W*	τ
r_1	61	478.69	-	-	-	-
r_2	-	-	-	-	-	-
r_3	5	5	5	0.09	5	26.22
r_4	9	27.78	9	0.11	9	28.56
r_5	4	1.4	4	0.04	4	23.97

Table 3. Results over erTS with |W| = 3,000

Requests	BlackBox		GraphPlan		IPP	
	#W	τ	*#W*	τ	*#W*	τ
r_1	75	38.09	-	-	-	-
r_2	50	16.02	-	-	-	-
r_3	22	18.68	-	-	22	24.78
r_4	23	4.38	-	-	23	21.06
r_5	38	4.01	-	-	38	21

sets and requests into PDDL and STRIPS files automatically. The experiments were performed on Linux with three Intel® Xeon™ CPU, running at 2.4GHz with 8GB RAM.

Tables 2, 3, and 4 shows the results of the five test requests for each of *baTS*, *erTS*, and *nwsTS* with $|W| = 3{,}000$. Graphplan ran out of memory in many cases. IPP also failed to solve the some requests. As a whole, Blackbox showed better performance than others, meaning that it can solve more requests than others. It is because Blackbox uses the local-search SAT solver, Walksat, for the satisfiability problem, so that Blackbox can run relatively well even with a large number of operators.

We can estimate the size of giant component in a service network using random graph theory. Often it is believed to be important to have a large and dense giant component in a service network. Otherwise, the isolated services will never have a chance to provide any services to clients. Newman, Strogatz, and Watts (2001) suggested the theoretical framework in order to estimate the giant component size in networks by using random graph theory. In order to see if their theoretical framework works, we generated the g_{op}^{f} with different network size for each of following cases:

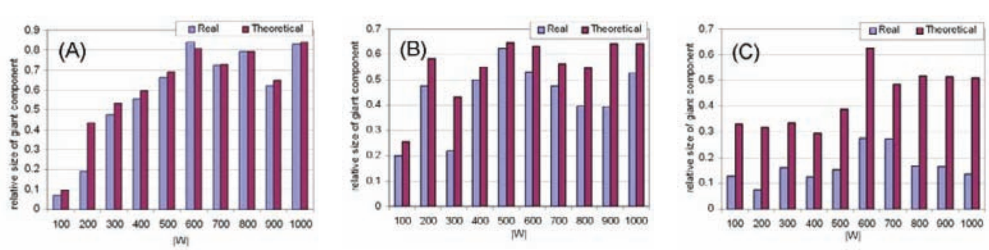
1. *Random model*: <50, *Erdos – Renyi*(100, 0.6), 0.8, 5, $|W|$>
2. *Scale-free model*: <50, *Barabasi – Albert*(100, 6), 0.8, 5, $|W|$>
3. *NWS model*: <50, *Newman – Watts – Strogatz*(100, 6, 0.1), 0.8, 5, $|W|$>

For each of these networks, we measured the real size of giant components. Then, we calculated the theoretical size of giant components according to the estimation model of Newman et al. (2001). The comparisons between real and theoretical sizes are summarized in Figure 10. For g_{op}^{f} based on the random parameter cluster network in Figure 10(A), the theoretical value of the giant component

Table 4. Results over nwsTS with $|W| = 3{,}000$

Requests	BlackBox		GraphPlan		IPP	
	$\#W$	τ	$\#W$	τ	$\#W$	τ
r_1	48	571.63	-	-	48	29.52
r_2	35	114.67	-	-	35	28.57
r_3	24	192.99	-	-	24	30.19
r_4	26	11.88	-	-	26	28.39
r_5	31	111.21	-	-	-	-

Figure 10. Comparison of real and theoretical size of giant components: (A) random, (B) scale-free, and (C) NWS models.



Estimating the size of giant component

size is very close to the measured one for each synthetic network. This implies that even a simple random model may be very helpful to estimate the inter-operable portion of such networks with random topology without even analyzing the available network beyond its degree distribution. However, Figure 10(B) shows that the estimation model is not a good model for Scale-free type. There is a considerable gap between theory and real value for many of the synthetic networks in this type. The deviation between theory and actual networks becomes more dramatic for the *NWS* (small world phenomenon and highly clustered property) type shown in Figure 10(C). The results show that the random network model might be good generative model for such Web services networks if these networks are entirely random, which is also in accordance with the basic assumption by Newman et al. (2001).

CONCLUSION

A novel Web-service benchmark toolkit, WSBen, is presented with three use cases in different application domains. The WSBen development is inspired by the study on real-world Web services, and is designed to provide diverse scenarios and configurations which users can fine-tune easily. As a result, using WSBen, users can conduct extensive experimental validation on their Web-service discovery and composition algorithms. It is our hope that WSBen will provide useful insights to the design and development of Web-services discovery and composition solutions and software. The latest version of WSBen is available at: http://pike.psu.edu/sw/wsben/. Further research is needed to extend WSBen to support approximate and semantic matching among Web services. Also, we plan to discover additional applications where the WSBen can be used.

ACKNOWLEDGMENT

Authors would like to thank Hyunyoung Kil and Ergin Elmacioglu for their help on the design and implementation of WSBen, and Professor Soundar R. T. Kumara at Penn State for his helpful comments on the earlier draft of this article.

REFERENCES

Albert, R. & Barabasi, A. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), pp. 47-95.

Albert, R. & Barabasi, A. (2000). Topology of evolving networks. *Phys. Rev.Lett., 85*, pp. 5234-5237.

Albert, R., Jeong, H. & Barabasi, A. (1999). The diameter of the world wide web. *Nature, 401*, pp. 130-131.

Bertsekas, D. (2000). *Dynamic programming and optimal control*, volume 1, 2nd ed. Athena Scientific, Boston.

Blum, A. & Furst, M. (1997). Fast planning through planning graph analysis. *Artificial Intelligence, 90*, pp. 281-300.

Brian, B., William, C., Michael, J., & Andreas, W. (2007). WSC-07: Evolving the web services challenge. *In Proc. of 9th IEEE International Conference on E-Commerce Technology and 4th IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC-EEE 2007)*, pp. 505-508, Tokyo, Japan.

Bylander, T. (1994). The computational complexity of propositional STRIPS planning. *Artificial Intelligence, 69*(1-2), pp. 165-204.

Constantinescu, I., Faltings, B. & Binder, W. (2004). Large scale testbed for type compatible service composition. In *Proc. of 14th International Conference on Automated Planning and Scheduling (ICAPS)*, pp. 23-28, Whistler, British Columbia, Canada.

Delgado, J. (2002). Emergence of social conventions in complex networks. *Artificial Intelligence*, *141*, pp. 171-185.

Denning, J. P. (2004). Network laws. Communications of the ACM, 47(11), pp. 15-20.

Erdos, P, Graham, R., & Nesetril, J. (1996). *The mathematics of Paul Erdos*. Springer-Verlag, Berlin, Germany.

Fan, J., & Kambhampati, S. (2005). A snapshot of public Web services. *SIGMOD Record*, *34*(1), pp. 24-32.

Kautz, H. & Selman, B. (1996). Unifying SAT-based and graph-based planning. In *Proc. of the 16th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 318-325, Stockholm, Sweden.

Kil, H., Oh, S., & Lee, D. (2006). On the Topological Landscape of Semantic Web services Matchmaking. In *Proc. of 1st International Workshop on Semantic Matchmaking and Resource Retreival (SMR2006)*, pp. 19-34, Seoul, Korea.

McDermott, D. (1996). A heuristic estimator for means-ends analysis in planning. In *Proc. of the third International Conference on Artificial Intelligence Planning Systems (AIPS)*, pp. 142-149, Edinburgh, Scotland.

McDermott, D. (2002). Estimated-regression planning for interactions with Web services. In *Proc. of the sixth International Conference on Artificial Intelligence Planning and Scheduling Systems (AIPS)*, pp. 67-73, Toulouse, France.

Newman, M. E. J., Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, *64*(026118).

Nilsson, J. (2001). *Artificial Intelligence: a new synthesis*. Morgan Kaufmann, San Francisco, CA.

Oh, S., Kil, H., Lee, D., & Kumara, S. (2006). WSBen: a Web-services discovery and composition benchmark. In *Proc. of the fourth International IEEE Conference on Web service (ICWS)*, pp. 239-246, Chicago.

Oh, S., Lee, D., & Kumara, S. (2007). WSPR: a heuristic algorithm for Web-service composition. *Int'l J. of Web services Research (IJWSR)*, *4*(1), pp. 1-22.

Sivashanmugam, K., Verma, K., Sheth, A., & Miller, J. (2003). Adding semantics to Web services standards. In *Proc. of the first IEEE International Conference on Web services (ICWS)*, pp. 395-401, Las Vegas, NV.

Watts, D. J. (1999). *The dynamics of networks between order and randomness*. Princeton Univ. Press, Princeton, NJ.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, *393*(4), pp. 440-442.

XMark. An XML Benchmark Project. Retrieve on September 17, 2006 from, http://monetdb.cwi.nl/xml/.

ENDNOTES

1. This article is a substantial extension from the short version that appeared in the proceedings of the 4th International IEEE Conference on Web Services (ICWS), held in Chicago, USA, 2006. The work of Seog-Chan Oh was done while the author was with the Pennsylvania State University.
2. As of August 2007, according to the estimation of Google Scholar, there are about 2,360 scholarly articles mentioning "Web Services Composition".
3. http://www.comp.hkbu.edu.hk/~ctr/wschallenge/
4. http://insel.flp.cs.tu-berlin.de/wsc06/

[5] The Semantic Web Services Challenge (2007): http://sws-challenge.org/wiki/index.php/Main_Page

[6] The Web Services Challenge (2007): http://www.wschallenge.org/wsc07/

This work was previously published in International Journal of Web Services Research, Vol. 6, Issue 1, edited by J.N.D. Gupta, S.K. Sharma, and J. Hsu, pp. 1-19, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 3.6
Architecture of the Organic.Edunet Web Portal

Nikos Manouselis
Greek Research & Technology Network (GRNET S.A.), Greece

Kostas Kastrantas
Greek Research & Technology Network (GRNET S.A.), Greece

Salvador Sanchez-Alonso
University of Alcalá, Spain

Jesús Cáceres
University of Alcalá, Spain

Hannes Ebner
Royal Institute of Technology (KTH), Sweden

Matthias Palmer
Royal Institute of Technology (KTH), Sweden

Ambjorn Naeve
Royal Institute of Technology (KTH), Sweden

ABSTRACT

The use of Semantic Web technologies in educational Web portals has been reported to facilitate users' search, access, and retrieval of learning resources. To achieve this, a number of different architectural components and services need to be harmonically combined and implemented. This article presents how this issue is dealt with in the context of a large-scale case study. More specifically, it describes the architecture behind the Organic.Edunet Web portal that aims to provide access to a federation of repositories with learning resources on agricultural topics. The various components of the architecture are presented and the supporting technologies are explained. In addition, the article focuses on how Semantic Web technologies are being adopted, specialized, and put in practice in order to facilitate ontology-aided sharing and reusing of learning resources.

INTRODUCTION

Following their introduction and commercial growth after 2000, Web portals have lately attracted increased research interest that focuses on a variety of aspects such as their business models, interface design, technical development, or their quality (Mahadevan, 2000; Tatnall, 2005a; Moraga et al., 2006; Tatnall, 2007). The term Web portal has been initially used to refer to well-known Internet search and navigation sites that provided a starting point for web visitors to explore and access information on the World Wide Web (Warner, 1999; Winkler, 2001). A Web portal can be generally viewed as a single, distilled view of information from various sources that integrates information, content, and other software services or applications (Averweg, 2007). Therefore, today Web portals can be simply defined as *gateways to information and services from multiple sources*, and their continuous development has been highlighted by relevant publications (Tatnall, 2005b).

A type of Web portals with particular interest are educational ones (Conceicao et al., 2003; Boff et al., 2006). Educational Web portals generally serve as gateways to information and services of some learning or teaching relevance and may cover a variety of types. They range from institutional Web portals that provide access to course listings and institutional information (such as Ethridge et al, 2000), to community portals that serve the needs of particular communities of learning and practice (such as DeSanctis et al., 2001; Luke et al., 2004). One category of educational portals that have recently received considerable interest (Neven & Duval, 2002; Richards et al., 2002; Hatala et al., 2004) is that of Web portals that provide access to some organized collection of learning resources. These portals usually facilitate users' access to the content in one or more learning repositories – that is, to database systems that facilitate the storage, location and retrieval of learning resources (Holden, 2003). Popular examples include both independent learning resources' portals such as MERLOT (http://www.merlot.org) and Teachers' Domain (http://www.teachersdomain.org/), as well as portals that list or aggregate learning resources from various other sources (e.g. other portals or repositories) such as OERCommons (http://www.oercommons.org).

Richards et al. (2002) stress that Web portals with learning resources may offer a wide variety of services based on what they seek to give to the user community behind them, although the more common are those aimed at facilitating users' search, access, and retrieval of the resources. For this purpose, they include services that will facilitate these processes, utilizing different types of user-related information (such as personal preferences) or resource-related information (such as the learning resource characteristics). One of the most recent trends in portal development is the use of Semantic Web technologies (Maedche et al., 2001). Semantic Web is an evolving extension of the World Wide Web (WWW) in which web content can be expressed not only in natural language, but also in a format that can be read and processed by software systems, thus permitting them to find, share and integrate information more easily (Berners-Lee, 1998). Numerous applications and case studies of Semantic Web technologies (e.g. ontologies for annotating information and expressing its semantics in a machine-processable manner) have been reported during the past few years. For instance, the World Wide Web Consortium (W3C) reports on several systems that have been put in production in existing organizations, as well as a number of commercial products (http://esw.w3.org/topic/CommercialProducts). Yet, the Semantic Web technologies have not so far reached the wide public. Some of the experts in the field claim that the reason is that large-scale applications, serving the needs of large user communities, have not been delivered yet (Shadbolt et al., 2006). To further illustrate their potential (and especially for Web portals), there is a need for implementing state-of-the-art Semantic Web technologies in large-scale applications. In the context of educational Web portals, this involves the semantic annotation of big collections of learning resources and their access and use from existing communities of users.

This article aims to contribute to this development by presenting such a large-scale implementation effort. More specifically, it discusses how semantic annotation and Semantic Web technologies are being adopted, specialized, and put in practice in order to set up a technical infrastructure that will facilitate sharing and reusing of learning resources for an educational Web portal. The case study is the Organic.Edunet Web portal, a portal that serves the needs of learning and teaching communities of the agricultural sector, by facilitating their access to a network (also called a federation) of learning repositories with learning resources on Organic Agriculture (OA) and Agroecology (AE) topics.

The article is structured as it follows. First, a short review of the way Semantic Web approaches are being implemented in similar applications is carried out. A description of the Organic.Edunet initiative is given, and the rationale for developing the Organic.Edunet Web portal is outlined. The main part of the article focuses on the description of the technical architecture of the Web portal, and on the way Semantic Web technologies are implemented in it. A discussion of perceived benefits and potential challenges is later carried out, to finally provide the main conclusions of this work.

BACKGROUND

Educational Semantic Web

From its initial conception around 1989 (Berners-Lee, 1998), the WWW (or simply, the Web) has been designed as an information space, with the goal that it should be useful not only for human-to-human communication, but also for machines that would be able to mediate and help. As Berners-Lee reports, one of the major obstacles to this has been the fact that most information on the Web is designed for human consumption, and even if it was derived from a very well specified database, the structure of the data is not evident to an automated software system browsing the Web. On the contrary, in his vision of the Semantic Web, data recovery for a particular context of use should be a routine, automated process. This is the reason why the empowering role of Semantic Web technologies has already been acknowledged in various contexts, such as education and training (Aroyo & Dimitrova, 2006). In this new paradigm, data would be specifically oriented to machine consumption by means of formal descriptions based on the existence and wide availability of ontologies, knowledge models of a given domain. Ontologies, which can be defined as collections of concepts representing domain-specific entities, the relationships between those concepts, and the range of admissible values for each concept (Daraselia et al., 2004) are in fact the key element of the Semantic Web. Ontologies serve as knowledge models for each particular domain of science, thus allowing to unambiguously represent, refer, and describe entities in that domain, and serving as the basis for interoperability and common understanding under formal and strict semantics.

Since the Web is becoming a popular educational medium at schools, universities and professional training institutions, a prominent new stream of research on the Educational Semantic Web has been established. Research studies already report semantic-based annotation and sharing of learning resources. For example, Forte et al. (1999) report on the principles underlying the semantic and pedagogic interoperability mechanisms built in the European Knowledge Pool System, a distributed repository of learning objects developed by the European research project ARIADNE (http://www.ariadne-eu.org).

In addition, Soto et al. (2007) designed an ontology schema capable to bring more flexibility to the description of the entities stored in semantic learning object repositories and, at the same time, to facilitate automated functions and task delegation to agents. Furthermore, Sicilia et al. (2005) describe the design of a learning object repository approach to what they called “semantic lifecycle” and illustrate thus through the concrete architecture of a semantic learning object repository prototype.

Moreover, semantic web applications are becoming more and more usual in education & training contexts. Sancho et al. (2005) for instance, applied these technologies to e-learning personalization by combining the information provided by ontologies, and the user profile, to create personalized units of learning. Santos et al. (2006) described an approach to promote interoperability among heterogeneous agents that are part of an educational portal. Their main contribution was to provide a means for social agents to communicate with agents outside its original scope through the use of semantic web technologies. Other implementations of Semantic Web technologies in educational portals also exist in the literature (Woukeu et al., 2003; Tane et al., 2004; Moreale & Vargas-Vera, 2004; Verdejo et al., 2004; Kotzinos et al., 2005).

To further illustrate the potential of the Educational Semantic Web, there is a need for implementing state-of-the-art Semantic Web technologies in large-scale applications that involve the semantic annotation of big collections of learning resources and their access and use from existing communities of users.

Organic.Edunet

To further promote the familiarization of consumers with the benefits of OA and AE - for their own health as well as for the benefits of the environment - the most dynamic consumer groups have to be properly educated. Young people at all stages of formal education have to be carefully approached through relevant educational programs in the curricula of all kinds of educational institutions, from elementary schools to relevant university departments. But apart from raising the awareness and education level of consumers, agricultural professionals must also be properly educated. By “agricultural professionals” we refer to the different types of future agricultural experts (e.g. natural production experts, veterinary experts, agricultural economists, extension officers, etc.), who study in agricultural universities around Europe, and who should be provided with a wide range of information related to OA and AE theories, methods, practices, and economic/environmental impacts.

Both groups (pupils and young agricultural students) constitute user groups of high importance. Children constitute tomorrow’s consumers, and they have to be properly approached and educated so that their nutritional, as well as their ecological and environmental awareness are developed. Students of agricultural universities constitute tomorrow’s agricultural professionals. They are expected to guide farmers through the adoption of OA and AE principles, or to serve themselves as the next generation of farmers/producers. Therefore, these two user groups have to be carefully approached through publicly available, quality, and multilingual educational content.

In this direction, the Organic.Edunet initiative (http://www.organic-edunet.eu), a European project that is funded by the *e*Content*plus* Programme and which involves 15 partners from 10 countries, aims to facilitate access, usage and exploitation of digital educational content related to OA and AE. Organic.Edunet will deploy a multilingual online federation of learning repositories, populated with quality content from various content producers. In addition, it will deploy a multilingual online environment (the Organic.Edunet Web portal) that will facilitate end-users’ search, retrieval, access and use of the content in the learning repositories. In this way, digital content resources that can be used to educate European youth about the benefits of OA and AE, will become easily accessible, usable and exploitable.

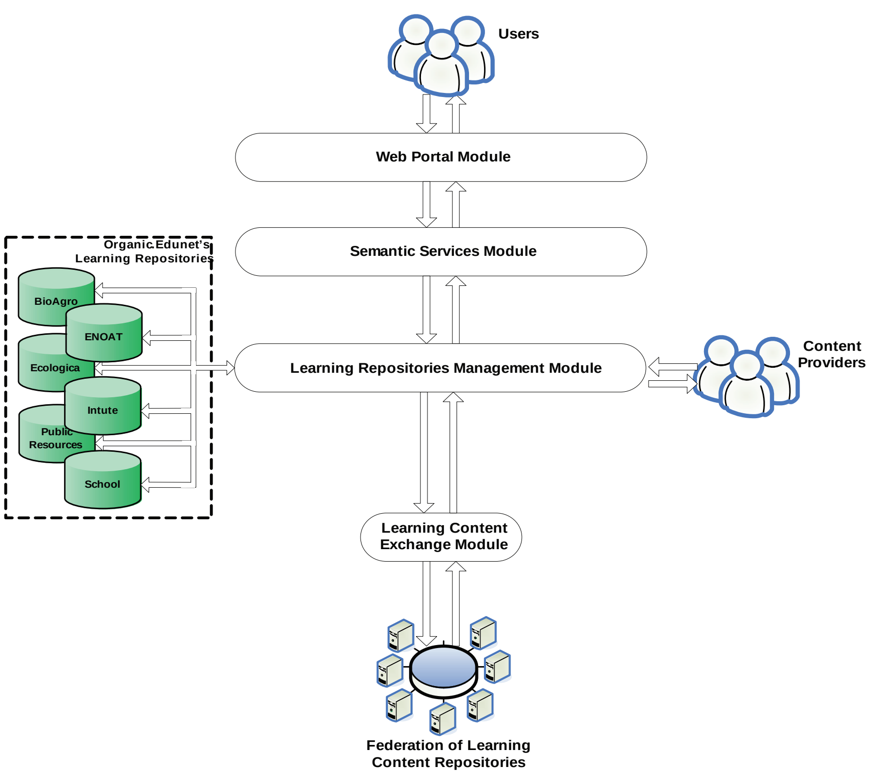
To achieve its aims, Organic.Edunet adopts state-of-art technologies that have been developed and tested in several research initiatives, but have yet to be proven in a real-life context. A characteristic example involves the implementation of Semantic Web technologies that have been previously developed in the context of the "LUISA: Learning Content Management System Using Innovative Semantic Web Services Architecture" EU project (http://www.luisa-project.eu/). The main characteristics of the LUISA architecture are its service-orientation and the built-in capabilities for semantic querying. For this purpose, semantic Web Services are involved, reusing the EU framework WSMO (http://www.wsmo.org/) for the brokering of multiple repositories. WSMO (which stands for Web Service Modeling Ontology) provides ontological specifications for the core elements of Semantic Web services. Taking the Web Service Modeling Framework (WSMF) as reference, WSMO defines four different elements for describing semantic Web Services: ontologies that provide the terminology used by other elements, goals that define the problems that should be solved by Web Services, Web Services descriptions that define various aspects of a Web Service, and finally mediators which bypass interpretability problems. In the following section, we describe the overall architecture of the Organic. Edunet infrastructure, and how technologies such as the ones adopted from LUISA are engaged.

ORGANIC.EDUNET INFRASTRUCTURE

Overall Architecture

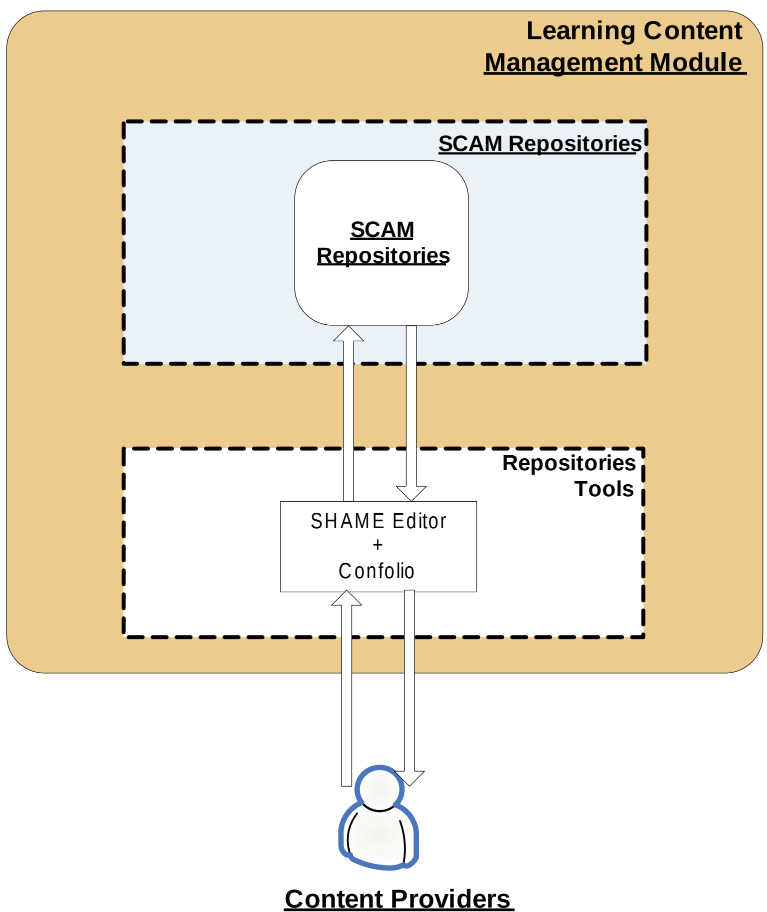
The overall architecture of Organic.Edunet is illustrated in Figure 1. The main elements of this architecture are the following:

Figure 1. Overall architecture of Organic.Edunet



- *Learning Repository Management Module*: includes the suite of tools that the Organic. Edunet content providers will use to create a digital collection of learning resources, to describe resources with appropriate metadata, and to publish resources in their own learning repository. Overall, six learning repositories are expected to be set up by the Organic.Edunet content providers (namely the Bio@gro, ENOAT, ECOLOGICA/COMPASS, Intute, School, and Public Resources ones).
- *Learning Resource Exchange Module*: concerns the connection of the Organic. Edunet federation with other federations of learning repositories, using open standards and specifications for the exchange of search queries and the harvesting of metadata. Organic.Edunet is expected to be connected with two external federations; the Learning Resource Exchange (LRE) of the European Schoolnet (http://lre.eun.org) and the ARIADNE Foundation (http://www.ariadne-eu.org/).
- *Semantic Services Module*: it is the core of the Semantic Web technologies' application in the architecture, and supports the semantically-enabled services that the Organic. Edunet Web portal will offer, by reasoning upon a number of integrated ontologies.
- *Web Portal Module*: refers to the end-user visible parts of the whole infrastructure, allowing users (including school teachers and pupils, university teachers and students,

Figure 2. Overview of the learning repository management module components



researchers etc.) to search, locate, retrieve and access learning resources on OA and AE throughout the whole Organic.Edunet federation.

Each module is further detailed in the paragraphs that follow.

Learning Repository Management Module

This module deals with the way content producers organize, annotate and publish learning resources and metadata in an Organic.Edunet repository. As illustrated in Figure 2, each of the Organic. Edunet content providers is expected to collect and annotate its learning resources, according to a multilingual application profile of the IEEE Learning Object Metadata (LOM) standard (LTSC, 2002). Two existing software tools are being adapted and integrated for this purpose:

- A configurable metadata editor built upon the code-library SHAME (available as Open Source at http://shame.sourceforge.net). With this code-library application programmers can develop flexible and easily extensible annotation tools for Semantic Web-based metadata. SHAME implements the Annotation Profile Model (Palmér et al., 2007a; Palmér et al., 2007b). This model is a configuration mechanism for the annotation of metadata and leaves the question of metadata standard compliance up to a metadata expert and not to the application developer.
- The electronic portfolio system Confolio (http://www.confolio.org) that allows the flexible management of folder-based repository interfaces.

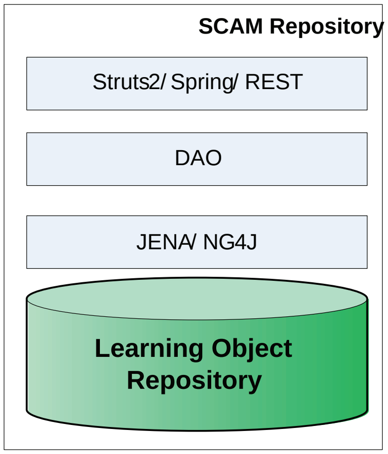
The content providers will use the integration of the SHAME editor and the Confolio tool in order to upload (if desired) their resources and the associated metadata into a learning repository that is called SCAM (Standardized Contextualized Access to Metadata), an Open Source Semantic Web repository solution for learning resources (Palmér et al., 2004). Figure 3 presents how a SCAM repository is accessed, using a combination of technologies.

The repository backend is resource-oriented and will store its metadata according to a Resource Description Framework (RDF, www.w3.org/RDF/) representation of the Organic.Edunet IEEE LOM application profile. The repository provides a range of connection interfaces, allowing the most appropriate to be chosen for each situation. An interface which exposes the repository closest to the internal representation is the REST (Representational State Transfer), a resource-based software architecture building fully on top of well established standards such as the HTTP protocol (Fielding, 2000). This makes it very easy to build interactive web applications on top of this interface.

The Confolio repository front-end builds on top of the REST-based web services exposed by the repository and an AJAX (Asynchronous JavaScript and XML) toolkit, which enables cross-browser compatibility and operating system independent application. The basic operations of Confolio can be separated in two groups: administrative (e.g. creation of new portfolios of learning resources) and end-user (e.g. creation of folders and description of resources using the SHAME metadata editor).

Using the Learning Repository Management Module, Organic.Edunet content providers may collect resources, annotate them using metadata conforming to the developed application profile, reviewing and approving resources, and then releasing resources for publication. Then, the metadata of the resources stored in a particular Organic.Edunet repository are (a) made available for harvesting from the Semantic Services Module and (b) made available for harvesting and/or search federation to external federations.

Figure 3. Overview of technology layers in a SCAM repository



Learning Resources Exchange Module

The Learning Resources Exchange Module allows for the communication of the Organic.Edunet repositories with external federations. Organic. Edunet will aim at the connection with the LRE and ARIADNE federations by adopting two widely used protocols and specifications:

- For communicating with ARIADNE: the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PHM, http://www.openarchives.org/OAI/ openarchivesprotocol.html) for making metadata available for harvesting from the ARIADNE services.
- For communicating with LRE: the Simple Query Interface (SQI, http://www.prolearn-project.org/lori) for serving/exchanging queries with the LRE services.

Metadata is transformed from its RDF representation into an XML representation, in order to be available for the external federations. Additional possibilities also exist for further interconnecting the Organic.Edunet repositories, due to their SCAM basis, e.g. the SPARQL Protocol and RDF Query Language - a W3C standardized query language and protocol for accessing RDF data (http://www.w3.org/TR/rdf-sparql-query/).

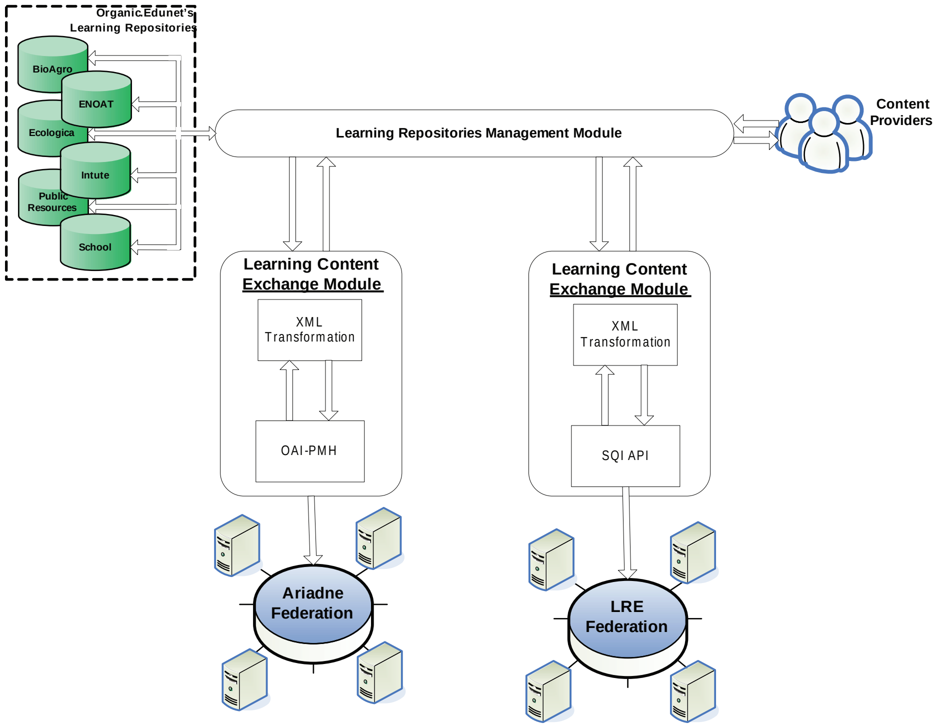
Semantic Services Module

The Semantic Services Module is the core engine behind the Organic.Edunet Web portal that allows offering users with semantic search capabilities. To support this, it is based on a semantic representation of the learning resources' metadata, as well as a number of ontologies that are engaged during search queries to provide reasoning capabilities. More specifically, metadata is transformed into an ontological representation inside a sub-module called LOMR.

The LOMR (standing for Learning Object Metadata Repository) is not itself a metadata repository but rather a framework which provides Web Service interfaces to any given, "real" learning object repository. LOMR instances allow developers to select the best repository implementation for a given application need, enabling specialized components, such as custom query resolvers and result composers, to benefit from the availability of different, heterogeneous LOMR instances. LOMR main features include the storage of learning object metadata in semantic format, the provision of a service-oriented interface and the import of metadata in non-semantic formats, among others.

In addition, LOMR offers semantic services to the Web Portal Module, following WSMO. It uses the Web Services Modeling Language (WSML, http://www.wsmo.org/wsml/) in order to provide formal syntax and semantics for WSMO, since it is richer in reasoning capabilities than the OWL Web ontology language (http://www.w3.org/TR/owl-features/) recommended by the W3C. Interoperability can be easily achieved through translating WSML to OWL through open source tools that are publicly available, such as the Web Service Modeling Toolkit (WSMT, http://wsmt.

Figure 4. Illustration of the way the Learning Resources Exchange Module operates

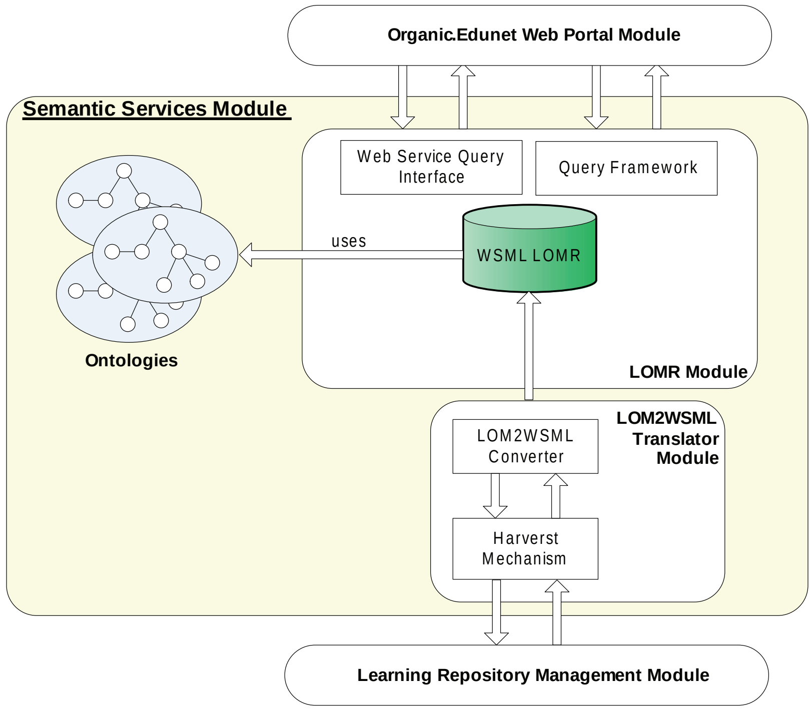


sourceforge.net) and WSMO Studio (WSMO4J, http://wsmo4j.sourceforge.net).

As a starting point, three ontologies are expected to be used by the Semantic Services Module. The first ontology will represent the domain area (OA and AE). It will serve all subject classification purposes, as well as allow for reasoning related to the semantics of the OA and AE concepts themselves. For example, searching for resources that have been classified using some concepts or terms related to the ones that the user has initially indicated. The popular AGROVOC (http://www.fao.org/aims/ag_intro.htm) ontology of FAO will be used as a basis for the construction of this ontology. The second ontology will be a geographical one. It will help reasoning related to the geographical origin and/or coverage of resources and their associated languages. For instance, it may allow users from a particular geographical region to search for resources in languages that have been indicated as related to the particular region, even if this has not been indicated in the initial search query. The third ontology will be representing IEEE LOM. It is expected to allow reasoning related to semantics of the LOM structure itself, such as searching for information in other elements than the ones that a user has initially indicated.

In LOMR, metadata will be harvested from the individual Organic.Edunet repositories using an appropriate harvesting mechanism. As Figure 5 shows, the RDF representations stored in the SCAM repositories will be converted to the WSML representation that LOMR requires. Once all the metadata information is stored in the LOMR repository in the formal, ontology-based format, LOMR will be able to expose various functionalities through semantic Web Services (described according to WSMO), allowing a wide variety of interactions with the Organic.Edunet Web portal.

Figure 5. Overview of the Semantic services module



Web Portal Module

The final module of the Organic.Edunet architecture is the Web Portal one. It actually comprises the online environment that will interact with the various user roles (school teachers & pupils, university teachers and students). For this purpose, it entails a role-filtering mechanism that will allow each user category to be presented with a user interface tailored to its specific needs. Apart from allowing users to semantically search and retrieve learning resources using the Semantic Services Module, the Web Portal Module will also provide the users with the option of evaluating/rating learning resources. Multi-dimensional numerical evaluations will be stored in appropriately defined evaluative metadata (Vuorikari et al., 2008). Then, they will give input to a collaborative filtering mechanism that will recommend users to look at resources that other users with similar preferences liked in the past (Manouselis & Costopoulou, 2007).

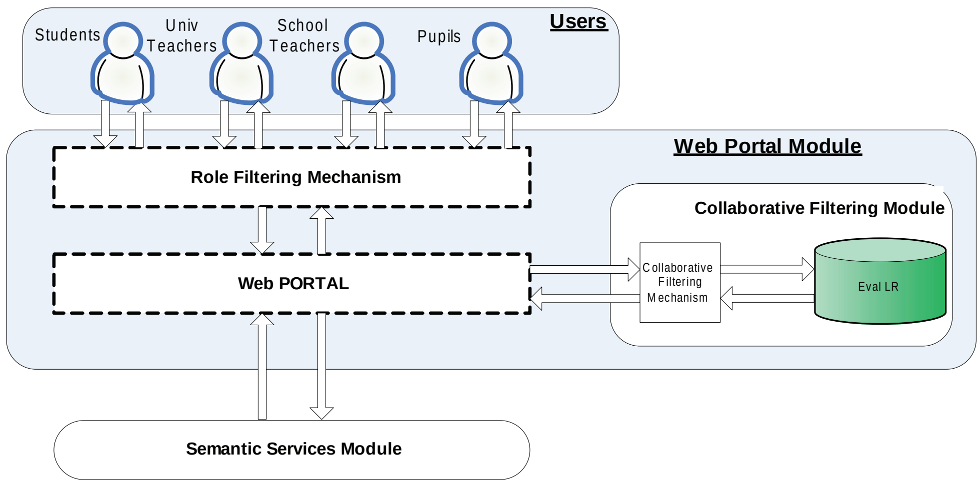
DISCUSSION

Benefits

As it has been described in the presentation of the Organic.Edunet portal architecture, there is a number of benefits expected from the adoption of Semantic Web technologies. The following paragraphs will go into more detail on these benefits.

The use of ontologies for the classification of learning resources will allow the refinement and expansion of queries. Users currently have to rely on keyword-based searches: for instance, a teacher looking for learning resources on the advantages of the use of organic fertilizers might try something like "advantages organic fertilizers" in Google (or in any other keyword-based search engine). A search on these keywords would return results containing either the terms "organic" or "fertilizer" or "advantage" or a combination of them, but many of those would be seen as non ap-

Figure 6. Elements of the Web portal module



propriate for most users. An example on the kind of (inadequate) resources that might be retrieved with this method –traditional keyword-based searches– would be the following:

- Commercial information on products by companies selling *organic fertilizers*.
- Resources on the *advantages* of non-*organic fertilizers* (matching the three keywords used for the search).
- Resources criticizing the use of *organic fertilizers* and discouraging users on its application due to their low efficiency and high prices.
- Resources on the elaboration of *organic* yoghurt (as they would match to at least one of the keywords provided).

Contrary to these examples, the use of ontologies for the description of materials would force the search engine to stick to strict-matching criteria to unambiguous definitions. It would also allow to search only educational-oriented materials explicitly annotated with the predicates such as "IsAbout" or "Provides BackgroundOn", which are related to e.g. organic fertilizers. This would even allow users to find just those learning resources explaining the advantages of the use of organic fertilizers and not criticizing them, and would allow users avoid suggestions for learning resources on other topics.

In addition, the use of ontologies will further enhance the search and browsing services offered to the users. More specifically, users will be able to browse through learning resources by selecting concepts of the ontologies used, together with an expression of their relationships. In a learning objects portal on organic farming, these technologies would help to easily access similar materials to a given one, as the relationships in the ontology would provide the ability to navigate from one instance to another. An example would be a search on learning objects about organic pest control, which would return e.g. a case study on the use of several types of insecticide-fungicide dust for use on fruit trees. Thus, portal users could navigate from the relationship from this object to fruit trees, and find e.g. learning objects on the commercialization of organic apples, or even lectures on organic fungicides applicable only to specific geographical regions.

Learning resources in current public repositories often have a high variability in their characterizations: from anything in digital format to well-defined educational oriented learning materials including metadata conformant to the IEEE LOM standard (McGreal, 2004). The description of all the knowledge about the domain of learning objects in the form of an ontology, and the use of this ontology as the basis for a learning object portal on organic farming, would provide the portal with the flexibility necessary to seamlessly accommodate different conceptualizations. It would also provide the ability to interact with external systems, even if each of these systems have a different understanding of what a learning object is, how their metadata should look like, etc. This model would eventually provide the users with a number of different functionalities, adapted to each particular concept of learning object, and not necessarily restricted by only one of these conceptualizations, applying technologies already in practice (Soto et al., 2007).

Challenges

Apart from the benefits, a number of challenges have also to be dealt with during the implementation of the Organic.Edunet architecture:

- The process of selecting, developing, and specifying the ontologies to be used (especially as far as domain-dependent ones such as the OA & AE ontology are concerned) is demanding and time-consuming, and needs the help of a number of experts from different disciplines. To make all the experts reach agreements is not always straight-forward (Sánchez-Alonso et al., 2008.).
- The process of engineering a new ontology often implies checking the new knowledge against the commonsense knowledge and general terms in an upper ontology. This process, which has to be carefully carried out, can be summarized in four iterative phases as described by Sánchez-Alonso & García (2006): (1) find one or several terms that subsume the category under consideration, (2) check if the mapping is consistent with the rest of the subsumers inside the upper ontology, (3) provide appropriate predicates to characterize the new category, and (4) edit it in an ontology editor to come up with the final formal version.
- Even though semantic web technologies are attractive and promise many benefits, they are not, unfortunately, ready for production use yet. Ontology-management systems hardly support the large ontologies needed for most production environments, and thus should be preferably used for research and experimentation purposes. A good example of the lack of maturity of these technologies are the APIs for ontology persistence in Java available today: Jena, Sesame and Protégé's persistence APIs find many difficulties in managing medium to large sized ontologies.
- Although the Organic.Edunet portal is based on a distributed architecture, the semantic module calls for a centralized repository that harvests the data from all repositories in the federation. Even though the harvesting tests carried out so far have shown good results, scalability, size and performance issues might arise as the project progresses and have an impact in the development.

CONCLUSION

To further illustrate the potential of Semantic Web technologies in Web portal applications, experiences from large-scale implementations are required. Especially in the case of Web portals that provide access to learning resources, the implementation of Web portals with services that are based on Semantic Web technologies that will be tested semantically annotating large collections of learning resources, and by being accessed and used from communities of users with numerous members. In this direction, this article presented a large-scale implementation effort that engages Semantic Web in order to set up a technical infrastructure that will facilitate sharing and reusing of learning resources for the agricultural domain. The next steps of this work concern reporting the results from the actual implementation, deployment, and initial testing of the technologies that are integrated in the Organic. Edunet Web portal.

ACKNOWLEDGMENT

The work presented in this article has been funded with support by the European Commission, and more specifically the project No ECP-2006-EDU-410012 "Organic.Edunet: A Multilingual Federation of Learning Repositories with Quality Content for the Awareness and Education of European Youth about Organic Agriculture and Agroecology" of the *e*Content*plus* Programme.

REFERENCES

Aroyo, L., Denaux, R., Dimitrova, V., & Pye, M. (2006). Interactive Ontology-Based User Knowledge Acquisition: A Case Study. *Lecture notes in computer science*, *4011*, 560.

Averweg U. (2007), Portal Technologies and Executive Information Systems Implementation, in A. Tatnall (Ed.), *The Encyclopedia of Portal Technology and Applications*, Hershey, PA: Idea Group Publishing.

Berners-Lee, T. (1998). Semantic Web Road Map. *W3C Design Issues. URL http://www. w3. org/ DesignIssues/Semantic. html, Oct.*

Boff E., Rizzon Santos E., Vicari R.M., "Social Agents to Improve Collaboration on an Educational Portal," Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06), IEEE Computer Press, 896-900, 2006.

Conceicao S., Sherry L., Gibson D., Amenta-Shin, G.,"Managing Digital Resources for an Urban Education Portal". in Proc. of World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA 2003), Honolulu, Hawaii, June 23-28, 2003.

Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., & Mazo, I. (2004). *Extracting human protein interactions from MEDLINE using a full-sentence parser*, 20, 604-611. Oxford Univ Press.

DeSanctis G., Wright M., Jiang L., "Building A Global Learning Community", Communications of the ACM, 44 (12), 80 – 82, December 2001.

EC (2004), European Action Plan for Organic Food and Farming, COM 415, Brussels 10 June 2004.

Ethridge R.R., Hadden C.M., Smith M.P., "Building a Personalized Education Portal: Get a Behind-the-Scenes Look at LSU's Award-Winning System", Educause Quarterly, 23(3), 12-19, 2000.

Fielding, R. T. (2000). Chapter 5: Representational State Transfer (REST), Architectural Styles and the Design of Network-based Software Architectures, Dissertation.

Forte, E., Haenni, F., Warkentyne, K., Duval, E., Cardinaels, K., Vervaet, E., et al. (1999). Semantic and pedagogic interoperability mechanisms in the ARIADNE educational repository. *SIGMOD Rec.*, *28*(1), 20-25.

Hatala M., Richards G., Eap T., Willms J., "The interoperability of learning object repositories and services: standards, implementations and lessons learned", in Proc. of the 13th International World Wide Web Conference, New York, NY, USA, 19-27, 2004.

Holden, C. (2003). From Local Challenges to a Global Community: Learning Repositories and the Global Learning Repositories Summit. Version 1.0, Academic ADL Co-Lab, November 11.

Kotzinos D., Pediaditaki S., Apostolidis A., Athanasis N., Christophides V., "Online curriculum on the semantic Web: the CSD-UoC portal for peer-to-peer e-learning", in Proc. of the 14th International Conference on World Wide Web, Chiba, Japan, 307 – 314, 2005.

LTSC (2002). IEEE Standard for Learning Object Metadata, 1484.12.1-2002. IEEE Learning Technology Standards Committee, 2002.

Luke R., Clement A., Terada R., Bortolussi D., Booth C., Brooks D., Christ D., "The promise and perils of a participatory approach to developing an open source community learning network", in Proc. of the 8th Conference on Participatory Design on "Artful integration: interweaving media, materials and practices", 11 – 19, Toronto, Ontario, Canada, 2004.

Maedche, A., Staab, S., Stojanovic, N., Studer, R., & Sure, Y. (2001). SEAL - A framework for developing SEmantic Web PortALs. *Lecture Notes in Computer Science*, 2097(1-22), 46.

Mahadevan B., "Business Models for Internet based E-Commerce: An Anatomy", California Management Review, 42(4), 55-69, 2000.

Manouselis N., Costopoulou C. (2007). Experimental Analysis of Design Choices in Multi-Attribute Utility Collaborative Filtering, International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), Special Issue on Personalization Techniques for Recommender Systems and Intelligent User Interfaces, 21(2), 311-331.

McGreal, R. (2004). Learning Objects: A Practical Definition. *International Journal of Instructional Technology and Distance Learning, 1*(9), 21-32.

Moraga A., Calero C., Piattini M., "Comparing different quality models for portals", Online Information Review, 30 (5), 555-568, 2006.

Moreale E., Vargas-Vera M., "Semantic Services in e-Learning: an Argumentation Case Study", Educational Technology & Society, 7 (4), 112-128, 2004.

Neven F., Duval E., "Reusable learning objects: a survey of LOM-based repositories", in Proc. of the 10th ACM International Conference on Multimedia, Juan-les-Pins, France, 291 – 294, 2002.

Palmér M, Enoksson F, Nilsson M, Naeve A (2007a). Annotation Profiles: Configuring forms to edit RDF. In: Proceedings of the Dublin Core Metadata Conference. DCMI Conference Papers, United States.

Palmér, M., Enoksson, F., Naeve, A (2007b). LUISA deliverable 3.2: Annotation Profile Specification,.

Palmér, M., Naeve, A., Paulsson, F., (2004), The SCAM-framework – helping applications to store and access metadata on the semantic web, Proceedings of the First European Semantic Web Symposium (ESWS 2004), Heraklion, Greece, May, 2004, Springer, ISBN 3-540-21999-4.

Richards, G., McGreal, R., Hatala, M., & Friesen, N. (2002). The Evolution of Learning Object Repository Technologies: Portals for On-line Objects for Learning. *Journal of Distance Education*, 17(3), 67-79.

Sánchez-Alonso S., Cáceres J., Holm A.S., Lieblein G., Breland T.A., Mills R.A., Manouselis N., "Engineering an ontology on organic agriculture and agroecology: the case of the Organic.Edunet project", in Proc. of the World Conference on Agricultural Information and IT (IAALD AFITA WCCA 2008), Tokyo, Japan, 24 - 27 August, 2008.

Sánchez-Alonso, S. and García, E. (2006) Making use of upper ontologies to foster interoperability between SKOS concept schemes. Online Information Review, 30(3), pp. 263-277.

Sancho, P., Martínez, I., & Fernández-Manjón, B. (2005). Semantic Web Technologies Applied to e-learning Personalization in< e-aula>. *Journal of Universal Computer Science*, *11*(9), 1470-1481.

Santos, E. R., Boff, E., & Vicari, R. M. (2006). Semantic web technologies applied to interoperability on an educational portal. In *Proc. of Intelligent Tutoring Systems, 8th International Conference*, 4053, 308-317.

Shadbolt N., Berners-Lee T., Hall W. (2006). The Semantic Web Revisited, *IEEE Intelligent Systems*, 21(3), 96-101, May/June.

Sicilia M. A, García E. Sánchez-Alonso S, and Soto J. (2005). A semantic lifecycle approach to learning object repositories. In *Proceedings of ELETE 2005 - eLearning on Telecommunications*, Lisbon, Portugal.

Soto, J., García, E. and Sánchez-Alonso, S. (2007). Semantic learning object repositories. *International Journal of Continuing Engineering Education and Life-Long Learning*, 17(6), pp. 432-446.

Stokes E., Edge A., West A. (2001). Environmental education in the educational systems of the European Union, Environment Directorate-General, European Commission, April.

Tane J., Schmitz C., Stumme G., “Semantic resource management for the web: an e-learning application”, in Proc. of the 13th International World Wide Web Conference, New York, NY, USA, 1-10, 2004.

Tatnall, A. (2005b), “Portals, Portals Everywhere...”, in Tatnall, A. (Ed.), Web Portals: the New Gateways to Internet Information and Services, Hershey, PA, Idea Group Publishing, 1-14.

Tatnall, A., Ed. (2005a). Web Portals: the New Gateways to Internet Information and Services. Hershey, PA, Idea Group Publishing.

Tatnall, A., Ed. (2007). Encyclopaedia of Portal Technology and Applications. Hershey, PA, Information Science Reference.

Verdejo M.F., Barros B., Mayorga J.I., Read T., “Designing a Semantic Portal for Collaborative Learning Communities”, in Selected Papers from the 10th Conference of the Spanish Association for Artificial Intelligence (CAEPIA03), Current Topics in Artificial Intelligence, LNCS 3040, Springer Berlin/Heidelberg, 251-259, 2004.

Vuorikari R., Manouselis N., Duval E. (2008). Using Metadata for Storing, Sharing, and Reusing Evaluations in Social Recommendation: the Case of Learning Resources, in Go D.H. & Foo S. (Eds.) *Social Information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectively*, Hershey, PA: Idea Group Publishing.

Warner, S. (1999). Internet portals, what are they and how to build a niche Internet portal to enhance the delivery of information services. In *Proceedings of 8th Asian-Pasific SHLL Conference*.

Winkler, R. (2001). Portals – The All-In-One Web Supersites: Features, Functions, Definition, Taxonomy. SAP Design Guild, Edition 3. Retrieved May 2, 2008, http://www.sapdesignguild.org/editions/edition3/portal_definition.asp.

Woukeu A., Wills G., Conole G., Carr L., Kampa S., Hall W., “Ontological Hypermedia in Education: A framework for building web-based educational portals”, in Proc. of World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA 2003), H Honolulu, Hawaii, June 23-28, 2003.

This work was previously published in International Journal of Web Portals, Vol. 1, Issue 1, edited by J. Polgar, pp. 71-91, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 3.7
Interactive Whiteboards in the Web 2.0 Classroom

David Miller
Keele University, UK

Derek Glover
Keele University, UK

ABSTRACT

This chapter summarizes the work underway to chart, critically evaluate, and systematize the introduction of interactive whiteboards (IWB) into modern foreign language classrooms in England. It is suggested that there is a developmental cycle whereby teachers take some time to understand the technology and become competent in its use. They then look to its advantages in presentation and the motivation of students before becoming aware of its pedagogical value and develop a changed classroom practice. This cycle is based upon enhanced teacher understanding of the nature of interactivity and the potential offered by the IWB in meeting a variety of learning needs. The relationship between IWB use and Web 2.0 arises from the potential of both to add impetus for teachers to structure lesson development and enhance activity. It is supported by teacher understanding of questioning techniques, and increasingly, by consideration of the use of gestures at the IWB. While IWBs are not a solution to all learning problems, it is suggested that they offers scope for greater student involvement and understanding in the learning process.

DOI: 10.4018/978-1-60566-190-2.ch027

INTRODUCTION

The interactive whiteboard (IWB) is part of the growing variety of equipment used in conjunction with a computer and data projector to incorporate software, Internet links and data equipment for whole class use. Increasingly schools are equipping each subject area, and in many cases every classroom, with an interactive whiteboard to supplement or replace traditional white or blackboards. This is happening in many parts of the world, for example in Mexico there has been a focus on IWB installation and use, wherever possible, to ensure that the full potential of the equipment and associated software can underpin quality lessons to be taught on the widest possible scale. This shows a fundamental belief that IWB technology and pedagogy can make a difference across a range of subjects (Hennessey, Wishart & Whitelock, 2007; Belli, 2005; McFarlane, 2005). Research shows that this may be true for certain young people and for a period of time but that fundamental changes promoting continued educational achievement are only possible where teachers recognize the significance of the word "interactive" and develop their approaches to teaching to promote this. Such approaches are concerned with driving student involvement and increasing understanding. They are based on the recognition of students' differing learning needs in order to ensure conceptual understanding and cognitive development (Armstrong et al., 2005; Hall & Higgins, 2005; Kent, 2006; Smith et al., 2005; Sturcke, 2004; Jones, 2004).

Glover and Miller (2003) have traced the pattern of increasing use in terms of the influence of "missioners, tentatives and luddites" within schools. More importantly they have demonstrated that teachers need to be helped through a three-stage development process so that they can move from traditional to increasingly more interactive approaches, specified as:

a. *Supported didactic*, where the teacher makes some use of the IWB but only as a visual support to the lesson and not as integral to conceptual development.
b. *Interactive*, where the teacher makes some use of the potential of the IWB to stimulate student responses from time to time in the lesson and to demonstrate some concepts.
c. *Enhanced interactivity*, where the teacher develops the materials so that the students focus upon the IWB as a means of prompting, explaining, developing and testing concepts for most of the lesson.

It is only at the third stage that the potential of the board as the focus of learning based upon a new understanding of the learning process, is recognized and realized by the teacher (Miller & Glover, 2004; Ziolkowski, 2004; Watson, 2006). The capacity to use the equipment in this way is dependent upon both technical fluency in the use of the equipment and associated software, and pedagogic understanding and flexibility to exploit the possibility of interactivity between teacher and student, and student and student. To achieve this has much in common with the educational development of all ICT and reflects a move, whether recognized or not, to the use of the Web 2.0 platform (Belshaw, 2007). Web 2.0 is here understood to be related to a focus on learning through concentration on multimedia use, age and ability linked group and individualized learning, and an awareness of variations in personal learning styles (Xhakli, 2008). This brings with it a change of emphasis from the teacher centered transmissive approach to learning to one characterized by interactivity, collaboration, user-generated content and immediacy of feedback. This is based on short attention switches from the teacher to the IWB as a mediating agency allowing access to other ICT technology within the classroom.

In a sense the IWB presents a new meta-language for classroom use. It certainly has developed its own vocabulary, which offers new technical terms. These become part of the basic language from initial training sessions with phrases such as "calibration," "drag and drop," and "hide and reveal" being early concepts for the user to understand. With the use of the interactive potential, phrases such as "virtual manipulatives" (Weiss, 2005) signify understanding of both process and pedagogic possibility, and as the integration of technology and pedagogy becomes better understood teachers and learners become aware of associated words from subject specific areas such as "the use of artifacts," which in both mathematics and modern languages has its own significance within the IWB focused classroom.

Language, however, is more than vocabulary, and IWB users become aware of the use of intonation, whereby the same word or phrase used in a different way signifies another meaning. This can be illustrated by considering the word "interactive" which is seen to operate at three levels — as indicating that there is a relationship between the technology and the user whereby a physical action leads to changes in the visual content on the boards; as an instruction to a user when using the board; or to one planning the sequence of conceptual developments and seeking a process by which movement on the board can lead to action in the brain of the recipient and subsequent action in the classroom.

This argument suggests that teaching and learning is limited by what occurs on the desk or the board but there is an intermediary in the process. This is the teacher, however defined; who acts as a mediator in the process of learning and who, we have noticed, develops a set of gestures as the non-verbal aspect of language. While not all users have the same hand and face gestures for similar aspects of mediation, research shows that users make consistent use of the same non-verbal expressions as lessons proceed (Miller & Glover, 2006).

This chapter concentrates on the outcomes of research that has been centered on the way in which IWB users, both teachers and learners, have developed their use of hardware and software to enhance teaching and learning in modern foreign language teaching. Our work was based on recent research and practice publications that highlighted the way in which IWBs could be a support in target language teaching. Research has also highlighted the role of ICT in language teaching and directed teachers to the use of the Internet, streaming videos and downloaded resources as a stimulant to interest in the classroom. Interactivity, however, is a feature of the Web 2.0 philosophy and this may extend beyond the classroom to include e-mail correspondence, blogging, and the use of *realia*. These are shown in the developing shareware from the Teacher Resource Exchange in England (tre.ngfl.gov.uk/server.php). As yet, though, there is very little modern language and IWB specific research. Glover et al., (2007) deal with the research outlined below in more detail and Gray et al., (2007) examine the integration of the IWB with teaching in the lower secondary school. Both are however, reporting on the need to move from didactic to more interactive approaches.

Our illustrations are taken from research into the learning of modern foreign languages within ten schools in England in 2004. These schools were all at an early stage of technology use and the experience of teachers in these schools accords with that of all new learners in that they have had to gain both competence and confidence in working with technology in enhancing pedagogy.

IWB AND MODERN LANGUAGE TEACHING

The selected schools were known to have previously good OfSTED (national inspection service) reports and were therefore likely to be showing good practice. Overall 13 lessons were video-recorded for subsequent analysis according to the following framework:

- The timeline and activity sequence in each lesson. This usually included a revision starter, and then moved through vocabulary use to sentence construction and grammatical understanding.
- Classroom management issues. These included the way in which the room was set out for the lesson, the nature of the environment to favor or inhibit IWB use for all students in the room, the integration of the IWB with traditional textbooks and other resources and the use of student groupings for learning activities.
- Enhancement from IWB use was sought within a framework of revision of past work, establishing new principles and data, sequencing of information and learning, as well as the demonstration of processes and reinforcement of learning through recall and the use of examples.

- The contribution of IWB use to cognitive development was assessed through the establishment of aims, the use of varied learning styles, stepped learning sequences with revision as needed, problem solving, and recall and discussion as a bridge to further learning.
- The contribution of IWB use to the conceptual development of discrete elements in the lesson through the identification of processes, manipulation of data, and review to ensure understanding and application as part of cognitive development.
- The nature of IWB techniques used within the lesson and the way in which these are perceived by students.
- An assessment of the teaching style used in the lesson.
- Identification of practical and pedagogical issues arising from the use of IWB technology in its contribution to effective learning.
- Measurement of the percentage of the lesson when the IWB was the focus of teaching and learning.

Structured interviews were also undertaken with ten teachers to probe aspects of their understanding of presentational, motivational and pedagogical issues inherent in technology use. The interviewers attempted to identify the reasons why, and how, teachers felt that the IWB made a difference to learning. Two groups of ten students each were interviewed in two schools to gain some triangulation with teacher opinion. There appears to be a run-in period of between eighteen months and two years while teachers develop competence in handling the technology, in developing fluency in its use and in establishing a battery of basic screens to support their teaching. Whilst teachers may have developed these skills their practice could still be grounded in older styles of teaching — or these styles may emerge in some lessons but not in others according to the needs of the topic and the class context.

PRESENTATION

During the lesson observation notes were made of the techniques used in the presentation of materials. In some lessons teachers used several techniques, in others they used just two or three but exploited them to the full as a further spur to learning. Overall, the frequency of use was as shown in Table 1.

Observation and interview evidence was also used to explore the processes by which IWB use promoted interactivity as understood by teachers in the lessons. In using techniques the four most common methods of securing interactivity were:

- Drag and drop, matching a response to a stimulant.
- Hide and reveal, opening a hidden response when the stimulant was understood.
- Matching equivalent terms, e.g. vocabulary in different languages.
- Movement, to demonstrate principles, e.g. sentence construction.

Students were also observed writing (and replacing) words, e.g. as they explained a story in a village mapped from the IWB, and shading e.g. to show rooms in a house where one would watch TV. In all of these the aim of the teachers was to: “have a number of children working at the board so that they could gain competence and confidence and to get others involved especially where we were using competitive approaches to keep them all involved” (Male teacher, Spanish lesson).

Teachers made use of superimposition by moving phrases or words and putting them alongside vocabulary or in sentences, and considerable use of matched verbal and visual representation of vocabulary. They made use of the coloring potential for parts of speech, and shading, to mark parts of a sentence as construction developed.

Nine of the teachers made some comment about higher standards of presentation as a result of the

Table 1. Use of techniques in IWB focused teaching

Techniques	Example	No of lessons (n=13)
Movement and animation	Cycle route on map	11
Drag and drop	Vocabulary	10
Overwriting of screen	Verb endings	10
Verbal and visual linkage	Sounds and objects	10
Superimposition	Labeling	10
Hide and reveal	Sentence construction	10
Shading	Comprehension	5
Imported sound	Clip	4
Gap infilling	Sentence construction	4
Internet access	Life in village	4
Highlighting	Parts of speech	4
Automatic responses	Vocabulary	4
Applet development	Describing actions	3
Tools	Connecting lines	2

use of IWB software and in each of the student groups there were three references to the way in which writing on the board had improved. According to teachers, it was "sort of professional looking" and "much easier to read than the writing we used to have."

At the same time both teachers and students spoke of the problems of "over-writing" where teachers made notes on diagrams on the board, and where "the writing looks odd, sort of angular." This is partly due to the level of fluency developed in the use of the pen on the IWB, but also related to the precision generated by the software.

Although increasing, at the time there were few commercial or professional programs specifically designed for teaching modern languages using the IWB. As a result practitioners speak of the need to develop their own materials often through electronically scanning textbooks, or from downloaded Internet material. In three of the lessons characterized as "supported didactic" a page of sketches had been scanned from a textbook and this lacked the movement, color and vitality of comparable material built up by the teacher from clipart collections, but given interest through attaching sounds. This incorporation of sound was a feature of half the lessons observed in modern languages. Four of the 13 lessons also made use of passages from the Internet as the basis of a comprehension activity and in two lessons students working with laptops were asked to pursue this at a higher level while the others in the class worked at the IWB.

Observation suggests that the use of the presentational aspects of the IWB varies as students get older. Year seven students (aged 11-12) showed enthusiasm and interest when filling in missing words in a competitive situation—the capacity for the IWB to have associated sounds for success and failure added to this. By year nine (aged 13-14), however, it appears that students are less willing to participate in either volunteering to write on the board unless all students are involved, or as a member of a small group at the board; completing "hide and reveal" type statements or hazarded answers, and demonstrating verbal relationships to the rest of the class. Indeed, there is some evidence that by this stage students will attempt to subvert some of the presentational advantages through spotting wrong results so that they incur the “noise of failure,” or give the wrong answer to “appear to be one of the gang.”

One skilled teacher pointed out, however, that this does not mean that students have outgrown the board. Rather they expect the teacher to be fluent in its use and to lead their learning in such a way that their consolidation takes place individually in their exercise books following teacher use of IWB materials. Discussion showed that even to age 16 students appreciate its value when the IWB is a source of further material for comprehension, or when it is used to demonstrate grammatical rules in action.

Consideration of the content and approach of the observed lessons indicates that the more didactic teaching was in lessons where there were fewer activities in the lesson period, where the pace was more limited and where there were longer periods of textbook or exercise work. In these lessons there were also fewer techniques used and teachers tended to make use of “drag and drop” or “hide and reveal” more than in lessons that used movement, automation (manipulatives) and color changes. In the lessons characterized by enhanced interactivity there was a tendency to use more activities with several techniques and a combination of commercially or professionally produced materials with those developed by the teacher. These lessons had greater pace and tended to use the IWB as the focus of all activity including board-based exercises and extension work. A year nine group learning German followed a three minute revision starter with three activities building vocabulary through highlighting, drag and drop and hide and reveal; building phrases through pair work drawing upon matching of vocabulary, gender and translation, to sentence construction based on an Internet activity. The lesson concluded with revisiting screens and the use of color highlighting to identify rules for case and gender agreement.

Teachers commented on, and used, color highlighting and arrows to indicate movement and positioning for parts of speech and to indicate verb endings. Over half the lessons observed made some use of associated sound, imported pictorial material and “real” newspaper or magazine extracts as a basis for comprehension work and the application of vocabulary. It was agreed that this was the greatest presentational advantage in that pre-prepared materials could be highlighted, expanded, developed and analyzed by over-written comment. In discussion respondents also considered the issues of “savability.” All except two participants had a battery of screens that they used as they prepared their lessons. The general view was that although it took time to prepare lessons for IWB use they could then be stored and used in three ways:

1. Catalogued by topic and then drawn out as each lesson was prepared.
2. Catalogued by lesson and then copied if the same screen was to be used in another lesson.
3. Catalogued by intended year group and then developed with further material if being used in a different context.

Teachers were less ready to regularly link their presentation to the printer so that materials could be made available for students. In 9 of the 13 observed lessons there was an element of copying from the IWB at some stage in the lesson. Table 2 shows the results of an analysis of the copying used in observed lessons:

Some copied activity characterizes all the teaching described as “modified didacticism,” but also occurs in the other styles of IWB use. It seems that teachers are less willing to explore or use the copying facility than is claimed by the promoters of IWB technology. The more positive view emerges from a linguist who commented on the time saved by being able to print off materials for those needing extra help.

Table 2. Analysis of copying activity during observed lessons (multiple activities possible)

Nature of copying activity	Number of language lessons (n=9)
Examples for exercises	6
Rules of grammar or process	4
Copied screen as a record	2
Aims of the lesson	1
Homework material	3

MOTIVATION

In all the discussions with teachers it has been difficult to sort out the motivational factors from the presentational or pedagogic in the successful use of the IWB. Seven of the teachers made reference to the intuitive use of the technology as a feature in the everyday lives of students and felt that the schools should be offering a high level of presentation and attractiveness so that "what happens in school should not be seen as a poor relation to what they see on TV and computer screens."

Our evidence suggests that the major features that encourage student motivation are as follows:

- The intrinsic stimulation provided by the combination of the visual, kinesthetic and auditory paths to learning.
- Those aspects of classroom management that lead to a focus on the IWB with linked desk activities throughout the lesson.
- The stepped learning that characterizes much IWB teaching offering constant challenges with frequent assessment of achievement as a stimulant to further involvement.
- The particular advantages for slower learning students or those who need reinforcement through the presentation of data or processes with more than one learning style (i.e. the ability of the board to allow material to be presented or represented in a variety of ways.

The observed lessons show, however, that older and more able students gain from the IWB because they appreciate the visualization of structures more readily than through verbally dominated approaches. A German lesson for 15-16 year olds exploited the IWB to build up and then analyze sentences in terms of constituent vocabulary, constructional frameworks and comprehension alongside continuous and enthusiastic encouragement from the member of staff who constantly referred them back to earlier screens. It was not simply the IWB, but also the way in which it had become integrated into the teaching method in a highly personal way combining visualization and encouragement of all students, that enhanced learning.

Another factor in the motivation of students stemmed from the way in which teachers exploited a "different type of contact with the lesson in the student's hands." Good practice obviously builds upon knowledge of particular groups and of individuals within the groups and a realistic assessment was that "the IWB still doesn't mean that we shall have a lesson where all the students are paying attention all the time." Boys, for example, are generally more ready to demonstrate or complete work at the IWB than girls of the same age. Older boys were more ready to demonstrate in part because it provides an opportunity for them to show their superiority in technological fields when teachers comment upon inadequacies of programs or available tools, while girls were more concerned about "being right" before they would commit themselves to the board. Evidence from the two student groups showed that they thought that "lessons had less wasted time" and that "they moved with more pace so that they didn't want them to come to an end." If there is one single motivational factor during lessons it appears to be that the immediacy of response ensures maintained interest. Seven of the teachers refer to the enhanced engagement in lessons and four referred to the ways in which the use of the IWB encouraged participation.

Although there was general agreement that teachers needed to consider aspects of lighting, student seating arrangements, sight lines, and the area of the board in use by students considering their physical characteristics, the observed lessons highlighted continuing issues. In four of the 13 lessons tables were organized in such a way that students were in rows at right angles to the board, or at grouped tables where half of the students naturally had their backs to the board. This problem is not subject specific but is related to the size of the room, access problems and the need for teachers to move around while desk work is in progress. In three classrooms light infiltration rendered vision difficult for those seated at the near front of the sides of the rooms. Amelioration was achieved in one school by using laptops with the same screen program so that vision was achieved and in another by breaking the lesson up in such a way that board activity was distinct from grouped activity. The latter was dependent upon group work using laptops and linked audio material while one group worked with the IWB and then groups moved to different activities in a subsequent lesson.

When the student groups were asked to identify why lessons were of greater interest than in traditional teaching they identified:

- The inherent interest of color, shading, dynamics, hide and reveal and demonstration.
- The sequential development of ideas and exemplars resulting from pre-prepared and commercial software.
- The availability of games that support learning, require responses that can be immediately assessed and then linked to a scoring system with team races or noughts and crosses.
- The "fun" arising from the use of pictorial matter and the immediacy of any processing built into the programs.
- The opportunity to revisit earlier concepts and examples in underpinning understanding.

Where lessons have such a dynamism and attraction it is likely that they will offer interest and challenge. This supports both revision of earlier work and enhanced understanding of new work. Above all as one teacher commented this offers "credible media for a new age." Teachers were conscious, however, of the time demands for preparation even when using commercial materials, and four referred to the problems of technology that could inhibit slick use of the IWB.

These data show that those lessons characterized by enhanced interactivity focused on the board for a greater proportion of the lesson, while those where the board was a support for more didactic approaches used the board for a significantly more limited period. For linguists more of the lesson may have to take place away from the board, e.g. in practicing vocabulary use, constructing sentences, and repeating words and phrases. The most interactive lessons were those where these activities were linked to the board. In four of the 13 lessons this led to a combination of choral reading, repetition of phrases and word completion using sentences from the board. Overall time on task is greater when the IWB is the focus of teaching and learning.

There was considerable concern that there could be a novelty value in the use of the new technology, "but we have to remember that students are used to this at home" and "that they think advanced technology now." One teacher commented "there is now danger that if we don't use the technology we will be seen as lacking in some way." All the respondents accept this but it is clear that teachers have developed strategies to ensure that there should be a continuing upward progression in learning and attainment. In a year seven French lesson the teacher used an introductory activity based upon naming colors, then moved to five vocabulary development exercises and finished with a learning check linked to boys versus girls scoring to ensure that momentum was maintained, that all the students were taking part and that visual stimulation was used to the full with a total of ten screens during the 35 minute lesson. That said, the dynamism of the teacher was important in supporting continuing learning — even broken with a two minute march to the French alphabet to stimulate renewed activity.

While it would be easy to claim great advantages for the IWB in motivating students at all ages it is evident that it is the quality of the teaching that ensures progress. Comparison was made of two lessons of vocabulary development with year seven groups. In one there were seven screens used in the course of the lesson but these were interspersed with pair work, a brief exercise and a discussion about rooms in the house. The students were animated throughout. In a comparable lesson, again with seven screens used, the teaching approach was much more didactic, there was little variation in activity from stage to stage in the lesson and the inter-relationship between teacher and learners was authoritarian and defensive. In such circumstances the lesson could not have the vigor, and "fun" element shown with a different teacher.

But there is another subtle influence noted by four of the respondents. This is because the constant progression in an interactive situation absorbs those who might otherwise become fidgety in a traditional classroom situation. They, in turn are less "nagged" during the lesson, enjoyment increases and motivation is supported: "It enhances collaborative work. This may just take the form of kids shouting out, correcting each other, say in a multiple-choice selection. This is very noticeable. As the teacher you too are working in a community, where you are visible. It does give a sense of competition, of expectation, the idea of can you beat it?" (male teacher, Spanish lesson).

PEDAGOGY

It was clear that teachers were using the learning of concepts as a basis for cognitive understanding. As a result in all but two of the 13 lessons there were discernible cognitive aims and a series of activities to explore, develop, explain and reinforce subsequent understanding. This was summed up one teacher as follows: "Sustained learner interest works in a number of different levels. It is not just a gimmick ... the interaction is important, like kids coming out to the board, having choices, e.g. they can decide on the verb ending, find the stem and match up the right pronoun. It makes concrete in their minds how the language works" (male teacher, German lesson).

There was a high level of understanding that students learn in different ways. This was seen where a pattern of viewing pictures, learning associated vocabulary, repeating its use in sentence construction, and then undertaking written or spoken group work ensured that: "we both enjoy teaching and learning more ... you can give clearer examples which are more interesting because of access to color and clip art. It's more aesthetically pleasing and is good for visual and kinesthetic learners and it's useful in that you can jumble up sentences and get them involved in reconstruction" (male teacher, French lesson).

Although there was some use of commercially developed activities, such as a short color recognition program — "we have developed our own materials from a number of sources, including download from the net, magazine and picture scanning and my own extensive library of clip art images" — this was seen to have advantages in that what was developed was meeting specific needs. Two teachers, however, expressed reservations — one about the time taken to produce good professional looking materials, and the other about "the danger of getting too structured and then unable to work flexibly if a problem occurs in the learning process for a particular topic" (female teacher, French lesson).

Teachers were all conscious of the need to maximize interactivity between themselves, the students and the learning materials. This is achieved through developing the opportunity to use "visual manipulation" so that concepts can be illustrated and worked upon by the students; the growth of shared evaluation of resources and the use of shared materials developed within subject areas, and exploitation of immediacy of feedback either through programmed software or through the use of presentational tools as with the colors program in French, or with right and wrong answer symbols. These programs are most effective as starters or for work with the least able when rapid responses and moving on enhance word manipulation.

There was also much debate about the place of traditional textbooks, exercise books, homework and other data sources in teaching. Over-writing was seen to offer scope for assisting cognitive development by "showing the same thing in different ways." Much of the Internet use was to download games and activities that did just this by underpinning learning of vocabulary and phrase development, or even with some audio links to check pronunciation. Most importantly, however, were the ways in which the IWB was being used to underpin lesson structure and to enhance cognitive development. Teachers variously appear to use a structure of:

- Setting objectives with or without revisiting earlier IWB slides.
- Using a bright and lively starter including "drag and drop," "hide and reveal" and multiple answers to stimulate interest, to offer a chance for brainstorming as a bridge to the main part of the lesson, and to revise necessary associated learning.
- Proceeding to the main part of the lesson where the IWB is the focus of much activity being used for illustration, explanation, sequenced ideas and the development of main principles. The progression was through the use of vocabulary and its application in sentences reinforced by practice and comprehension. During this section of the work the approach was distinguished by challenged responses with the emphasis on understanding and then using language correctly — with practice in the completion of sentences on the IWB reinforced by group activities. In this way, as one teacher commented, "you move the students with you." Interview respondents identified a tension between those who thought that time taken in managing the students' use of the IWB while others were watching could be seen as a loss to active learning but in eight language lessons students were given tasks alongside the work being illustrated on the board so that all the students were active.
- Concluding with a plenary session involving the use of recall, examples and previously worked material to ensure understanding and to act as the basis for extension work. This section of the lesson was more usually concerned with revisiting vocabulary and structures and then looking at an associated screen requiring comprehension or conversation as a consolidation for the lesson.

Awareness of the need for cognitive development and the place of concepts within this was shown in the frequent reference to sequencing of ideas, the availability of a range of pre-prepared examples appropriate "to age and ability," adaptability of materials to allow for "alternative approaches and the use of different ways of learning." This was through vocabulary understanding and pronunciation, and through phrase and sentence construction to use in verbal and aural comprehension. Three linguists outlined the use of supportive materials from the net or other sources, and three referred to the need to help students understand the technology e.g. in the use of pens and programs, so that they could become fluent in the interactivity required if whole class participation was to be assured.

There were comments that dependence on sequenced slides in some pre-prepared materials in PowerPoint and Excel, as well as in some of the commercial materials, could inhibit flexibility in revisiting ideas and in offering alternative explanations appropriate to "whether they can learn verbally or not." This was not seen to be a problem in the observed lessons because of the technological fluency shown in accessing screens. There was a general view amongst those interviewed that when the staff have the time to develop their materials and access to appropriate technological support it was possible to use the IWB to generate faster and more effective learning, with tighter planning and the implementation of lesson plans according to the need to cover the prepared material.

There was frequent reference in the interviews to the need to match materials to the needs of the students and that some differentiation of task, activity or outcome required teachers to be flexible, adaptable, and "aware of the ways in which consolidation can occur without going back to old fashioned practices such as copying." This was illustrated in a comparison between two groups learning and applying clothing vocabulary showing that the more able group moved on to determine the difference between summer and winter clothing while using similar screens of information.

In pedagogic review the teachers also drew attention to the clear match of objectives to activities and the understanding of these by students so that they could use the board to help in their evaluation of progress. They showed an awareness of what the IWB could offer and in the two most stimulating lessons Web 2.0 approaches were integrated into the teaching. In one lesson there were five groups working at their own level in differing learning situations. These included the use of an interactive software program at the IWB, access to the net by a group using a laptop, randomized questioning in pairs with an interactive program on a desktop, and the preparation of a presentation by a group working with PowerPoint. It is possible that all these approaches can exist individually without being specifically labeled as Web 2.0 but they are now being used to shift the emphasis from teacher to student, from lecture to learning.

THE DEVELOPING AGENDA

Arising from the agenda it appears that there are two pedagogic areas for further investigation. The first is the relationship between the teacher, the student and the materials involved. For enhanced interactivity to occur this has to be understood as a chain reaction where the IWB is a means of mediation between learners and learning. There are four elements in this process:

a. *Teaching approach*. Ernest (1994) suggests a simple scale for the approach used by teachers. At the lowest level the teacher is an "instructor" concerned with the presentation of concepts as rules followed by practice. At the higher level the teacher is "facilitator" offering approaches that enhance understanding, and at the highest level the teacher is a "mediator" bridging between student understanding and development. In their use of the interactive whiteboards the instructor is concerned with elements of presentation. Conversely the mediator deals with issues arising from questions and thereby regards the interactive whiteboard as a vehicle for interaction with students.

b. *The use of the interactive whiteboard.* In both the approaches discussed above, it is evident that the interactive whiteboard enhances the role of the teacher regardless of where s/he is on the spectrum. The teacher-as-instructor will be working with prepared material, to be presented in a logical sequence, and often with a PowerPoint sequence as the basis of the teaching. The material is likely to be focused on statements of facts and definitions, headings etc. but there will also be examples to be copied and exercises to be completed. Such material is likely to be organized, clear and monotone. On the other hand, the teacher-as-mediator will be concerned with how the IWB can support the features of mediation such as modeling and coaching in relation to the topic under consideration. In collaborative classrooms, modeling serves to share with students not only what one is thinking about the content to be learned, but also the process of communication and collaborative learning. Modelling may involve thinking aloud (sharing thoughts about something) or demonstrating (showing students how to do something in a step-by-step fashion). Coaching involves giving hints or cues, providing feedback, redirecting students' efforts, and helping them use a strategy. A major principle of coaching is to provide the right amount of help when students need it — neither too much nor too little so that students retain as much responsibility as possible for their own learning (Tinzmann et al., 1990). This can be seen in the selection of appropriate adjectives or in the search for word meanings. Miller, Glover and Averis (2005) have suggested that as competence improves teachers become more ready to develop and use manipulatives as the basis of interaction. This is seen to particularly good effect in consideration of the accommodation available at differing costs within a French holiday town where the input of so many Euros into a slot machine then produced a range of menus for description and selection. It is our contention that the use of particular manipulations might be used effectively to support the role of teacher-as-mediator (Miller, Glover, Averis & Door, 2005).

c. *Questioning*. Experienced and effective teachers use questioning intuitively. They probably think little about the nature or level of the question but proceed as they think fit. Inexperienced and poor teachers appear not to have such skills. Much has been written about the nature of questions and the art of questioning. Mason (2000), in his commentary on the work of many in this field, clearly demonstrates the complexities of the process and relates questioning to both conceptual and cognitive development. Analysis of the video recorded lessons suggests that open and closed questions and those focusing on product or process are frequently used but are only partially helpful in developing higher order learning.

d. *Learning Models*. The fourth element in developing interactivity stems from the learning model espoused by the teacher. Observations have been made on the way in which teachers use the constructivist and social-constructivist views of learning as defined by Piaget (Piaget & Inhelder, 1974) and Vygotsky (1978). Students construct concepts and meaning, as a solo activity, based on their own experience. Associated with this model is the notion of "cognitive conflict" whereby children are exposed to something that is different from (conflicts with) their currently perceived models. From Vygotsky, the focus is on the social-constructivist view of language and the extent to which it is linked with the formation of knowledge. Furthermore, all knowledge is a social construction and based on shared views and images. In language teaching the social context of much learning offers scope for constructivist learning to be enhanced. The opportunity to call on a vast range of Internet resources helps when technological fluency allows access.

GESTURE

In the introduction to this chapter we spoke of the impact of intonation on language understanding and we return to this in considering the way in which teachers, and indeed board-using students, gesture while mediating between board and class.

There is an increasing awareness that teaching is a multi-modal activity drawing upon a range of communicatory activity including verbal, visual and interpersonal communication, as well as associated technology. Jewett (2004) has shown that knowledge of multi-modal perception and pedagogy can support both teachers and taught. Abrahamson (2003) outlines the role of artifacts or bridging tools, including gesture in that learning process. Watson and De Geest (2005) outline the need for consideration of all aspects of communication in teaching and learning, and Rasmussen et al (2004) explore the use of consistent gesture as part of these multi-modal approaches. Goldin-Meadow and Wagner (2005) take these patterns of gesture further and consider the impact of these on both learners and their learning environment through reflection of the state of knowledge and subsequent change through cognitive understanding.

There is considerable evidence of the way in which the teachers using enhanced interactive approaches were constantly using recognizable gesture patterns. One female teacher used all-embracing movements to secure attention at the start of most lessons almost sweeping the students along with her as she summed up her aims and then moved towards the IWB. During starter periods her hands were used in a quick to and fro movement linking students to the IWB but ensuring that the pace of the lesson was maintained. In the main section of the lesson her movements were slower, often indicating building or process stages, and then opened in an invitational way as explanation was returned to the students for consolidation. There was then a return to quicker, pointing and sequencing gestures as stages were revisited in the plenary section of the lesson. When asked about the pattern of interaction the teacher referred to "the need to keep them on their toes, but to feel that we were learning together."

Ferscha et al., (2005) attempt to extend the gesture typology with three families of gestures — hand gestures, gestures of an artifact held permanently (e.g. an IWB pen) and gestures that are detached from the hand and manipulated occasionally (e.g. change of software). All of these convey messages by the way in which they are used. While such a system is of potential value for user interface computer technology development it does not offer the sort of vocabulary of gestures that match the instinctive activity by teacher and student in the classroom. However, it is the basis of gesture sensing devices and could well offer an insight into a typology because it may be that students read more into body language, as shown when recall of an IWB screen fails and frustration is indicated, or when invitations are issued for students to work at the IWB and they respond with acceptance gestures.

In our analysis of video-recorded lessons it was possible to ascertain the reliance on gesture by both teacher and students and the combination of gesture as explanation, indication and invitation. The IWB both encourages and reinforces learning through the use of visual as well as more transitory gestures that offer shapes in the air. During this lesson gestures were used and emulated in an often involuntary manner, in all three areas of gesture; hand, software and artifact. The hand movements that mediated technology and learning through movements were:

- *Invitational*, with the use of movement linking students to the IWB, offering the pen for use, showing a step and offering an opportunity for participation – often encouraged with IWB software.
- *Displaying*, with hand gestures pointing to material on the IWB and then using movement, highlighting or overwriting to indicate content or process.
- *Blocking*, with hand gestures putting a barrier between the students and the IWB as a result of mistakes or the need to re-think a process and then followed by an invitational reinforcement of process and use of drag and drop and over-writing to support this.
- *Sequencing*, with the gestures to indicate progression and using gestures to pose a question and then to work through sequences of example questions.

It would seem that students learn not only because of the difference in presentation but also because the IWB offers additional modes of gesturing that support verbal and visual explanation. It may be that this kinaesthetic quality will meet the needs of those who cannot readily learn with didactic approaches. Our observations suggest that where teachers are using enhanced interactivity with the IWB they are employing considerable gesturing to great advantage

ACHIEVING INTERACTIVITY

The starting point for the effective use of IWB technology has to be in teacher training. Nevertheless, the move from traditional didactic approaches to changed pedagogy is complex. It has been recognized that although UK student teachers are required to have a basic knowledge of computer use as a requirement for certification many already have a high degree of computer literacy and technological understanding. Whether this can be harnessed to enhance teaching appears to be related to other factors including the nature of curriculum development programs, school technology resource levels, and individual teachers' planning and reflection. Kennewell (2001) suggests that effective evaluation of ICT use will prompt more awareness of, and adaptation to, the complexity of influences in the classroom. More pessimistically, Robertson (2003) argues that despite the potential impact of ICT on teaching and learning it remains a marginal influence on student attainment. He argues that other significant changes have been more willingly achieved in education and that the slow pace of change in ICT may be related to social, anthropological and cultural aspects of the human and computer interaction. Kirschner and Selinger (2003) point to the disparate technological competence of teachers and the children they teach and argue that if ICT is to be a core technology then teachers need to recognize not only how to use the different technologies but also follow through a five stage development from pre-novice through novice, apprentice, and practitioner to expert user. The elements of this stage are the ability to reflect, evaluate and adapt both content and approach to address student needs. If this is to be achieved, then the work of teacher educators takes on a major role extending beyond the "how to" to the "why" of ICT and the use of interactive whiteboards (Sturkle, 2004).

This requires understanding of the potential of Web 2.0 tools in association with IWB use to change the way in which teachers encourage learning. Interactivity may be a matter of question and answer but Web 2.0 approaches in modern language teaching may open the way for the use of interactive software, as for example in vocabulary extension work; for the use of the net in developing comprehension; for the use of search engines in preparing presentations, and for enhanced understanding of the cultural context. In this way presentation spurs motivation and this, in turn, promotes higher attainment. This is especially so where collaborative group work has been developed to meet differing learning styles. Web 2.0 tools provide the means of both conceptual and cognitive development.

However at the time of the investigation it was not possible to podcast and share videos. These technologies offer considerable opportunities (and threats) for teachers and pupils. The possibilities will undoubtedly be constrained by the technical, pedagogical and attitudinal backgrounds of the teachers. Further limiting factors will be the way in which uses of some Web 2.0 technologies are “censored” and restricted by school firewalls.

In language teaching students may find considerable benefits in using (and creating) products that may help them with their study. Generally the technological skills will be within the grasp of learners — but the option to demonstrate and use these skills may be overlooked.

Even at the most basic technological level this may require fundamental changes in aspects of initial teacher education. In simple terms, the assertion that mentoring teachers should be at least competent in ICT use was found wanting by Cuckle and Clarke (2003) who comment on the considerable variation in student support between schools. When that competence occurs for Knezek and Christensen (2002) the focus of subsequent change is determined by evidence that:

as teachers progress from lower level, simple applications toward full integration of technology in the classroom in support of higher cognitive functions, attitudes progress in predictable patterns along with changes in their needs. (p. 375)

Once established as teachers and in continuing professional development there is some evidence that successful one-to-one coaching can be achieved where the technologically adept students are paired with teachers having a much wider pedagogic experience to mutual benefit (Matthew et al., 2002). Mooij (2004) argues that teachers have to be aware not only of the technical aspects of newer technologies but also of the curricular and instructional gains that can be made, and more importantly of the way in which technology and pedagogy can be integrated to achieve flexible and individually sensitive learning situations. Triggs and John (2004) have demonstrated the need for working groups at departmental, whole-school and educational service levels, interconnecting for professional growth through the sharing of technical and pedagogic experience.

The recurrent theme is one of a discrete way of teaching and learning using ICT and Taylor (2004) suggests that this requires a three stage development from personalization to achieve fluency in using the technology, through pedagogic sensitivity to its potential, to the development of contingent thinking to allow responsive and reflective use of materials. In the context of continuing professional development, this requires strong support within teacher training institutions and the schools with whom they work in partnership. This will then help teachers who have been inappropriately, or inadequately, trained in the pedagogy and do not realize the need to develop interactivity through the use of a variety of teaching and learning styles, artefacts and gesture — in short, coping with the affordances of the technology (Conole & Dyke, 2004). Failure to make a significant pedagogic change will, we suggest, lead to wasted opportunities and the danger that equipment with the potential to change understanding, application and the conceptual development of learning will be at worst, unused, and at best a presentational aid.

For this to occur there has to be further consideration of the professional development provided for users. Glazer and Hannafin (2006) building on Vygotsky’s social constructivist approaches suggest that this exploration of what happens in the classroom is best undertaken as a social enterprise where peers rely on the expertise and support of one another to adopt innovative practices: "Reciprocal interactions in a community of practice, where teachers take responsibility for each other's learning and development, may provide an effective means of supporting situated professional learning" (p. 179). Contextual work by Schrum et al. (2005) points to the need for departments to continually refine, reassess and redevelop their teaching approaches. Eekelen et al. (2005) have shown that this process needs to be regulated rather than self-regulated and unstructured — with implications for those responsible for professional development, and Tearle (2003) shows that this is particularly true of learning in technology based contexts where the learning culture is fundamental to teacher involvement and shared experience

CONCLUSION

There appears to be a learning curve for both teachers and students. The former need time to develop their technological fluency, apply pedagogic principles to the available materials or to the development of materials, and then to incorporate the IWB seamlessly into their teaching. Few teachers base all their lesson on the IWB all the time, and over half those interviewed stressed that the IWB has to be seen as part of the equipment available but that there was still a need for the use of texts, exercises and other media. Teachers then appear to become more aware of the nature of interactivity and its stimulation as the basis for conceptual development and cognitive understanding. Students also need to have a range of manipulative skills if they are to take part in lessons without loss of self-esteem as technologically incompetent. Even so good practitioners ensure that all students have access to the board, and are given help if there are signs of unhappiness with the medium.

It is only when basic technological fluency and pedagogical understanding has been achieved that teachers can then overcome the novelty factor. Our evidence suggests that there is an initial period where interest is stimulated by the cleverness of the technology, but after a period students are more aware of three great gains:

1. Brighter and clearer presentation of material
2. Stepped learning and the ability to recall earlier material
3. Rapid responses to interactive examples so that learning is reinforced or revisited

Where students have reached this stage, they accept the IWB as part of the battery of learning resources offered to them and progress beyond novelty to enhanced learning. At this stage any possible behavioral problems are usually overcome because students are caught up in the sequence and pace of learning and appear to "take off" in their understanding, achievement and consequent self-esteem.

There is evidence that language teaching is being transformed by competent and confident teachers but this is not to suggest that the IWB is a panacea for all ills. As yet, there is only a limited shift in classroom practice and student learning and transformation will require markedly changed teacher understanding. Our evidence suggests that there is a teacher progression from supported didactic to enhanced interactivity in their classroom and pedagogical management. Where there is still reliance on the copying of material, textbook exercises and minimal conceptualization of learning so that it can be interactive, the gains are minimized. Effective learning is inhibited where the IWB is given a novelty value by the teacher so that it becomes something different, where the physical surroundings are not conducive to IWB use and where the lesson lacks pace. It is not sufficient to argue that the use of the IWB will, of itself, bring the classroom into the Twenty First Century and the visually stimulated environment. Effective teaching requires that the technology and the pedagogy are directed towards enhanced and structured understanding. “I love my board because it gives so much to the kids,” as one teacher said, may be the clue that enthusiasm can be regenerated not just in the students but in the staff also.

REFERENCES

Abrahamson, D. (2003). Embodied spatial articulation: a gesture perspective on student negotiation between kinaesthetic schemas and epistemic forms in learning mathematics. In McDougall, D. E., & Ross, J. A. (Eds.). *Proceedings of the 26th Annual Meeting of the International Group for the Psychology of Mathematics Education* 2, 791-797.

Armstrong, V., Barnes, S., Sutherland, R., Curran, S., Mills, S., & Thompson, I. (2005). Collaborative research methodology for investigating teaching and learning: The use of interactive whiteboard technology. *Educational Review*, *57*(4), 457–469. doi:10.1080/00131910500279551

Belli, M. (2005). Technology is not white and black. *Times Educational Supplement*. Retrieved May 1, 2008, from: http://www.tes.co.uk/search/story/?story_id=2063467

Belshaw, D. (2007). Flattening the world: How to harness Web 2.0 tools to engage learners inside and outside the classroom. *In-service development day Kings Heath Leading Edge Partnership*. Retrieved November 27, 2007, from: http://www.slideshare.net/edte.ch

Conole, G., & Dyke, M. (2004). What are the affordances of information and communication technologies? *ALT-J*, *12*(2), 113–124. doi:10.1080/0968776042000216183

Cuckle, P., & Clarke, S. (2003). Secondary school teachers, mentors and student teachers' views in the value of ICT in teaching. *Technology, Pedagogy and Education*, *12*(3), 377–391. doi:10.1080/14759390300200168

Eekelen, I., Boshuizen, H., & Vermunt, J. (2005). Self-regulation in higher education teacher learning. *Higher Education*, *50*(3), 447–447. doi:10.1007/s10734-004-6362-0

Ernest, P. (1994). The impact of beliefs on the teaching of mathematics. In Bloomfield, A. & Harries, T. (Eds.). *Teaching and Learning Mathematics*. Derby: Association of Teachers of Mathematics.

Fersha, A., Resmerita, S., Holzmann, C., & Reichor, M. (2005). Orientation sensing for gesture-based interaction with smart artefacts. *Computer Communications*, *28*(13), 1552–1563. doi:10.1016/j.comcom.2004.12.046

Glazer, E. M., & Hannafin, M. J. (2006). The collaborative apprenticeship model: Situated professional development within school settings. *Teaching and Teacher Education*, *22*(2), 179–193. doi:10.1016/j.tate.2005.09.004

Glover, D., Miller, D., Averis, D., & Door, V. (2007). The evolution of an effective pedagogy for teachers using the interactive whiteboard in mathematics and modern languages: An empirical analysis from the secondary sector. *Learning, Media and Technology*, *32*(1), 5–20. doi:10.1080/17439880601141146

Glover, D., & Miller, D. J. (2003). Players in the Management of Change: introducing interactive whiteboards into schools. *Management in Education*, *17*(1), 20–23. doi:10.1177/08920206030170010701

Goldin-Meadow, S., & Wagner, S. M. (2005). How our hands help us learn. *Trends in Cognitive Sciences*, *9*, 234–241. doi:10.1016/j.tics.2005.03.006

Gray, C., Pilkington, R., Hagger-Vaughan, L., & Tomkins, S. A. (2007). Integrating ICT into classroom practice in modern foreign language teaching in England: Making room for teachers' voices. *European Journal of Teacher Education*, *30*(4), 407–429. doi:10.1080/02619760701664193

Hall, I., & Higgins, S. (2005). Primary school students' perceptions of interactive whiteboards. *Journal of Computer Assisted Learning*, *21*(2), 102–117. doi:10.1111/j.1365-2729-.2005.00118.x

Hennessy, S., Wishart, J., & Whitelock, D. (2007). Pedagogical approaches for technology-integrated science teaching. *Computers & Education*, *48*(1), 137–152. doi:10.1016/j.compedu.2006.02.004

Jewitt, C. (2004). *Knowledge, Literacy and Learning: Multimodality and New Technologies.* London: Routledge Falmer.

Jones, K. (2004). Using Interactive Whiteboards in the Teaching and Learning of Mathematics: a research bibliography. *Micro Math*, *20*(2), 5–6.

Kennewell, S. (2001). Interactive whiteboards — Yet another solution looking for a problem. *Information Technology in Teacher Education*, *39*, 3–6.

Kent, M. (2006). Our journey into whiteboard hell. *Times Educational Supplement*. Retrieved May 1, 2008, from: http://www.tes.co.uk/search/story/?story_id=2287557

Kirschner, P., & Selinger, M. (2003). The state of affairs of teacher education with respect to information and communications technology . *Technology, Pedagogy and Education*, *12*(1), 5–17. doi:10.1080/14759390300200143

Knezek, G., & Christensen, R. (2002). Impact of new information technologies on teachers and students. *Education and Information Technologies*, *7*(4), 369–376. doi:10.1023/A:1020921807131

Mason, J. (2000). Asking mathematics questions mathematically. *International Journal of Mathematical Education in Science and Technology*, *31*(1), 97–111. doi:10.1080/002073900287426

Matthew, K., Callaway, R., Letendre, C., Kimbell-Lopez, K., & Stephens, E. (2002). Adoption of ICT by teacher educators: one-on-one coaching. *Technology, Pedagogy and Education*, *11*(1), 45–62. doi:10.1080/14759390200200122

McFarlane, A. (2005). The taming of the whiteboard. *Times Educational Supplement*. Retrieved May 1, 2008, from: http://www.tes.co.uk/search/story/?story_id=2153815

Miller, D., & Glover, D. (2004). *Enhancing mathematics teaching through new technology: The use of the interactive whiteboard*. Retrieved May 1, 2008, from: http://www.keele.ac.uk/depts/ed/iaw/nuffield.htm.

Miller, D. J., Glover, D., & Averis, D. (2005). Developing Pedagogic Skills for the Use of the Interactive Whiteboard in Mathematics. *British Educational Research Association, Glamorgan*. Retrieved September 25, 2007, from: http://www.keele.ac.uk/depts/ed/iaw/docs/BERA%20Paper%20Sep%202005.pdf.

Miller, D. J., & Glover, D. C. (2006) *Enhanced secondary mathematics teaching: gesture and the interactive whiteboard*. Paper presented at the British Educational Research Association Annual Conference, University of Warwick, September 6-9.

Miller, D. J., Glover, D. C., Averis, D., & Door, V. (2005). The Interactive Whiteboard — A literature survey. *Technology, Pedagogy and Education*, *14*(2), 155–170. doi:10.1080/14759390500200199

Mooij, T. (2004). Optimising ICT effectiveness in instruction and learning. *Computers & Education*, *42*(1), 25–44. doi:10.1016/S0360-1315-(03)00063-0

Piaget, J., & Inhelder, B. (1974). *The child's construction of quantities*. London: Routledge & Keegan Paul.

Rasmussen, C., Stephan, M., & Allen, K. (2004). Classroom mathematical practices and gesturing. *The Journal of Mathematical Behavior*, *23*, 301–323. doi:10.1016/j.jmathb.2004.06.003

Robertson, J. W. (2003). Stepping out of the box: rethinking the failure of ICT to transform schools. *Journal of Educational Change*, *4*(94), 323–344. doi:10.1023/B:JEDU.0000006047.67433.c5

Schrum, L., Burbank, M. D., Engle, J., Chambers, J. A., & Glassett, K. F. (2005). Post-secondary educators' professional development: Investigation of an online approach to enhancing teaching and learning. *The Internet and Higher Education*, *8*(4), 279–289. doi:10.1016/j.iheduc.2005.08.001

Smith, J., Higgins, S., Wall, K., & Miller, J. (2005). Interactive whiteboards: boon or bandwagon? A critical review of the literature. *Journal of Computer Assisted Learning*, *21*(2), 91–101. doi:10.1111/j.1365-2729.2005.00117.x

Sturcke, J. (2004). ICT strategy? 'Just keep taking the tablets'. *Times Educational Supplement*. Retrieved May 1, 2008, from: http://www.tes.co.uk/search/story/?story_id=2055679

Taylor, L. (2004). How student teachers develop their understanding of teaching using ICT. *Journal of Education for Teaching*, *30*(1), 43–56. doi:10.1080/0260747032000162307

Tearle, P. (2003). Enabling teachers to use information and communications technology for teaching and learning through professional development: influential factors. *Teacher Development*, *7*(3), 457–472. doi:10.1080/13664530300200222

Tinzmann, M. B., Jones, B. F., Fennimore, T. F., Bakker, J., Fine, C., & Pierce, J. (1990). *What Is the Collaborative Classroom?* NCREL: Oak Brook.

Triggs, P., & John, P. (2004). From transaction to transformation: information and communication technology, professional development and the formation of communities of practice. *Journal of Computer Assisted Learning*, *20*(6), 426–439. doi:10.1111/j.1365-2729.2004.00101.x

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Watson, A., & De Geest, E. (2005). Principled teaching for deep progress: improving mathematical learning beyond methods and materials. *Educational Studies in Mathematics*, *58*, 209–234. doi:10.1007/s10649-005-2756-x

Watson, D. (2006). Understanding the relationship between ICT and education means exploring innovation and change. *Education and Information Technologies*, *11*(3/4), 199. doi:10.1007/s10639-006-9016-2

Weiss, D. F. (2005). Keeping it real: The rationale for using manipulatives in the middle grades. *Mathematics Teaching in the Middle School*, *11*(5), 238–242.

Xhakli, K. (2008). *Going online: How innovative projects can transform education and ICT opportunities*. Kosovo: IPKO Institute.

Ziolkowski, R. (2004). Interactive Whiteboards: Impacting Teaching and Learning. *Media & Methods*, *40*(4), 44.

KEY TERMS

Artifact (BE Artefact): Artifact is an object or item. However it can also be the on screen representation of an object or an item.

Gesture: This is a term encompassing human actions here associated with the use of the interactive whiteboard e.g. hand and body movements and facial expressions. There is evidence that users develop consistent hand and facial gestures e.g. in seeking responses, rejecting wrong responses and that learners assimilate these as part of the teaching package offered by individual teachers.

Interactive Whiteboard (IWB): An interactive whiteboard consists of a computer linked to a data projector and to a touch sensitive large electronic screen usually fixed to a wall. Images from the computer are then displayed onto the whiteboard by means of the data projector. These images can be manipulated at the electronic screen usually by means of a special pen or a finger (this depends on the properties of the electronic screen). The term interactive whiteboard often refers only to the electronic screen.

Interactivity: Interactivity is an approach to learning in which teacher and learner interact to ensure understanding, enhance conceptual development and stimulate debate. Learning is stimulated through participation rather than through rote or passive learning which characterises didactic approaches.

Motivation: In this context, is an outcome of presentation because of the greater interest offered to learners and the reinforcing of concepts through learner engagement.

Presentation: Presentation is the use of the software potential of the interactive whiteboard to enhance the way in which words, concepts, ideas and relationships are displayed. Design, color, movement and more complex virtual manipulatives offer a superior way of showing data on an interactive whiteboard with the intention of prompting learner participation. The use of a variety of means of display may meet the needs of learners with differing learning styles.

Social Constructivist Approaches: These are based upon the complex interaction between teacher and learner, or between learners, and relate to the way in which we learn from each other with greater facility once the social network of the context is known and when the culture of the learning group has been developed.

Virtual Manipulatives: A virtual manipulative is a computer program that represents a piece of equipment on a computer screen. Examples include a cannon that can fire cannon balls, a protractor for measuring angles and a geoboard where you can place and manoeuvre “elastic bands” on a grid on “nails.” Virtual manipulatives are most commonly written in Flash and JavaScript.

This work was previously published in Handbook of Research on Web 2.0 and Second Language Learning, edited by M. Thomas, pp. 506-525, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 3.8

Web 2.0 Technologies:
Social Software Applied to Higher Education and Adult Learning

Teresa Torres-Coronas
Universitat Rovira i Virgili, Spain

Ricard Monclús-Guitart
Universitat Rovira i Virgili, Spain

Araceli Rodríguez-Merayo
Universitat Rovira i Virgili, Spain

M. Arántzazu Vidal-Blasco
Universitat Rovira i Virgili, Spain

M. José Simón-Olmos
Universitat Rovira i Virgili, Spain

ABSTRACT

Web 2.0 technologies are playing an important role in building social capital through increasing flows of information, and building on knowledge and human capacity of learning. The purpose of this chapter is to show the role that social software, a component of Web 2.0 technologies, can play in higher education and adult learning. This chapter focuses on the role of Web 2.0 technologies in promoting learning. New learning paradigms and pedagogical applications are also discussed.

DOI: 10.4018/978-1-60566-739-3.ch059

INTRODUCTION

Education has traditionally been conducted face-to-face, with professors performing outstanding magisterial classes in front of the learners. During the centuries, students and professors have shared the same time and same space frame. Nowadays, things are quite different. Information technology (IT) is a reality affecting the whole education system from primary school to higher education and adult learning. IT is having a considerable impact on the learning providers, on the learning process itself and, of course, on any agent involved in the process.

History has demonstrated that technology affects education profoundly. Considering the definition of technology broadly, one may say that prehistoric people used primitive technologies to teach skills to their young (Frick, 1991). Whenever a new medium entered the picture, a new wave of educational delivery arrived. Radio, television, and now computers have all impacted the field of distance education. Though some studies (see Russell, 1999) report no significant differences in performance between face-to-face instruction and technology supported environments.

Nowadays, campuses are networked, faculty post their notes on Web pages, students access the library from their rooms, and entire classes can have discussions via chat software (Rice-Lively, 2000). This development has recently come to be labeled under the by now commonly accepted term e-learning (Hudson, 2003).

The European e-Learning Action Plan 2001 (European Commission, 2001) defines e-learning as the use of new multimedia technologies and the Internet to improve the quality of learning by facilitating access to resources and services as well as remote exchanges and collaboration. This requires new e-interaction and e-communication competencies and a reorganization of e-learning structures. Components can include content delivery in multiple formats, management of the learning, and a networked community of learners (Gunasekaran, McNeil, & Shaul, 2002). Internet/World Wide Web have meant that opportunities have been identified for developing distance learning activity into a more advanced online environment. It is known as Virtual Learning Environment (VLE), which eliminate geographical barriers while providing increased convenience, flexibility, individualized learning, and feedback over traditional classroom (Kiser, 1999). Higher education institutions devote substantial resources to providing students with access to internet-based information, VLEs and other forms of e-learning. These efforts are predicated upon an assumption that "university students are inherently inclined towards using the internet as a source of information within their day-to-day lives and, it follows, disposed towards academic use of the internet" (Selwyn, 2008, p. 12).

But, today, the traditional approach to e-learning is currently changing from the use of Virtual Learning Environment (VLE) to e-learning 2.0, an approach that combines the use of complementary tools and Web services -such as blogs, wikis, trackback, podcasting, videoblogs, and other social networking tools- to support the creation of ad-hoc learning communities. In this context, most of the current research tends to be concerned with the potential of the worldwide Web and other internet applications to accelerate university students' learning and knowledge-building, and support interactivity, interaction and collaboration (Selwyn, 2008).

This proposal aims to provide an introductory perspective on the learning impacts of new media and Web 2.0 information and communication technologies on the e-learning environment. Web 2.0 technologies are playing a crucial role in building of social capital through increasing flows of information, and building on knowledge and human capacity for learning. Social software has emerged as a major component of the Web 2.0 technology movement. But, how can social software play a role in higher education and adult learning? To answer this question, this proposal will focus on the role of Web 2.0 technologies in promoting learning. Pedagogical applications, which stem from their affordance of collaborative knowledge discovery, will be discussed. At the same time the chapter will also explore the pedagogical methodology involved considering that e-learning Web 2.0 leads us from constructivism to navigationism. Finally, some suggestions are made for future research in this field.

BACKGROUND

Social Software and Web 2.0 Technologies as a Must for a Digital Life

The term social software is generally attributed to Tim O'Reilly. Social software includes a large number of tools used for online communication, e.g. instant messaging, text chat, internet fora, Weblogs (or blogs for short), wikis, social network services, social guides, social bookmarking, social citations, social libraries and virtual worlds.

O'Reilly (2005) presented Web 2.0 as a second stage in the development of the Web. He describes Web 2.0 as an "architecture of participation'' where collective intelligence generates a ''network effect'' leading to Websites that become more valuable as more people participate (O'Reilly, 2003). For McGee and Begg (2008) "Web 2.0 represents a group of Web technologies with a user-centric focus that actively change and evolve with user participation" (p. 164). Web 2.0 is referred to as a technology (Franklin & Van Harmelen, 2007) and at the same time as a community-driven online platform or an attitude rather than technology (Downes, 2005).

Web 2.0 technologies are already having a significant impact on the way in which we communicate in both our personal and professional lives. Mejias (2005) wrote down a list of non-definitive kinds of social software applications, arranging technologies according to their social function (learning, selling, classifying, defining communities, and so on). Mejias (2005) stated that most social software products incorporate functions from more than one category and, also, most of them pose challenges to pedagogical approaches. And, these challenges are today instructors' challenges.

Organizational structures in the 21st century are also increasingly networked and with virtual teams becoming the norm. Virtual team working requires tools that enable the exchange of documents and information and collaborative creation. Wikis and blogs have taken relatively little time to become part of the suite of tools used for collaborative virtual projects. In this new organizational landscape, enterprise social or "collaborative software is probably the most visible current

Table 1. Different types of social software and its applications

Social software	Applications
Multiplayer gaming environments	Multi-User Dungeons (MUDs), Massively-Multiplayer Online Games (MMOGs).
Discourse facilitation systems	Synchronous: instant messaging (IM) and chat (e.g. Windows® Live Messenger, AOL Instant Messenger, Yahoo® Instant Messenger, Google™ Chat, Skype™); chat. Asynchronous: e-mail, bulletin boards, discussion boards, moderated commenting systems (e.g. Slashdot)
Content management systems	Blogs, wikis, document management (e.g. Plone™) and, Web annotation utilities.
Product development systems	Especially for Open Source software (e.g. Sourceforge.net®, Libresource)
Peer-to-peer (P2P) file sharing systems	Napster®, Gnutella, BitTorrent™, eMule, iMesh
Selling/purchasing management systems	eBay™
Learning management systems (LMSs):	Blackboard, WebCT, Moodle
Relationship management systems	Friendster®, Orkut
Syndication systems	list-servs, RSS aggregators
Distributed classification systems	Flickr®, del.icio.us.

challenge. Interpersonal communication has become an integral part of the process of content creation, hence the value placed on communities and networks" (Abell, Chapman, Phillips, Stewart & Ward, 2006, pp. 244-245).

Enterprise Social Software is a term describing social software in "enterprise" (business) contexts -definition provided by Wikipedia- [http://en.wikipedia.org/wiki/enterprise_social_software/]. It includes social and networked modifications to company intranets and other classic software platforms used by large companies to organize their communication'. Enterprise 2.0 is a paradigm shift. Organizations are increasingly focusing on leveraging internal information and on connecting people to people and people to content. Web 2.0 tools and techniques focus on collaboration and information/knowledge sharing. Business Information Survey explores the penetration of Web 2.0 tools. Results show that there is great strategic interest in social technology and Web 2.0 tools and techniques, but not much serious deployment yet (Foster, 2008). But, in our digital world, *digital natives* (Prensky, 2001) eagerly embrace social software developing the skills necessary to engage with social and technical change, and to continue learning throughout the rest of their lives..

As workers live Web 2.0 digital lives, organizations also will need to update their e-learning corporate practices. In that sense, Trondsen (2006) predicts strong uptake of virtual worlds in corporate learning and notes a number of pilot projects underway in company learning contexts. As students live Web 2.0 digital lives, instructors need to begin to deeply explore and develop new learning paradigms with these technologies and practices. And, finally, as the students of today grow into the leaders of tomorrow, they will bring these technologies into their organizations, making their use an essential part of the future of world of work and life-long learning.

Learning Paradigm Shifts

Since many years ago, different theories have been developed to explain how we learn. Behaviorism, cognitivism, and constructivism are the three broad learning theories most often utilized in the creation of learning environments. Neither of these views can be regarded as exclusively right or wrong. It is, however, necessary to know that constructivism is presently accepted as the most relevant of the three. In the pedagogical arena it is a must to analyze how these models allow instructors to create the circumstances best suited to facilitate student learning.

The first one, behaviorism, is a worldview that assumes a learner is essentially passive, responding to environmental stimuli. It stems from the work of Pavlov –the father of classical conditioning- and Skinner –the father of operant conditioning. Behavior theorists define learning as nothing more than the acquisition of new behavior. Learning is "any more o less permanent change in behaviour which is result of experience" (Borger & Seaborne, 1966, p.16). The behaviourist definition of learning focuses on the behavioural outcomes of learning, rather than on knowledge, attitudes and values.

After the behavioural theories came cognitive ones. The most influential theorists were Piaget and Vygotsky. Cognitivism theories seek to explain how the brain processes and stores new information. People are rational beings that require active participation in order to learn, and whose actions are a consequence of thinking. The learner is viewed as an information processor.

Constructivism as a paradigm posits that learning is an active, constructive process. According to a constructivist view, learning is seen as the individualized construction of meanings by the learner. The learner is an information constructor. Constructivist learning theories posit that knowledge is built by the learner, not supplied by the teacher (Piaget, 1967). People, by reflecting on their experiences, actively construct their own subjective representations of objective reality. Each of us generates our own mental models, which we use to make sense of our experiences. Learning, therefore, is simply the process of adjusting our mental models to accommodate new experiences.

In the present landscape of technological change, important transformations are underway in terms of how we teach and learn. There is a growing shift on the need to support the acquisition of knowledge and competencies to continue learning throughout life. "With respect to ICT, we are witnessing the rapid expansion and proliferation of technologies that are less about "narrowcasting", and more focussed on creating communities in which people come together to collaborate, learn and build knowledge" (McLoughlin & Lee, 2007, p. 664). So, constructivist approaches have grown to include social constructivism, which refers "to learning as the result of active participation in a community" where new meanings are co-constructed" (Brown, 2006, p. 111). Different learning strategies have been designed based on a community supported constructionist approach in which constructionism strategy –a strategy connected with experiential learning and based upon constructivist theories of learning- is situated in a supportive community context (Bruckman, 1998). This approach emphasizes the importance of social aspect of learning environment. The construction of new knowledge is the aim of these learning theories.

But beyond constructivism and social constructivism new paradigms are emerging. Brown (2006) focus on navigationism as the last learning paradigm shift. In this new learning paradigm the emphasis will be on knowledge navigation. Learning activities will be focused on exploring, connecting, evaluating, manipulating, integrating and navigating. Learning will take place when learners solve contextual real life problems through active engagement in problem-solving activities, and networking and collaboration. Siemens' principles of connectivism (Siemens, 2004) provides a summary of the connectivist learning skills required within a navigationist learning paradigm:

- Learning is a process of connecting specialized nodes or information sources.
- Capacity to know more is more critical than what is currently known.
- Nurturing and maintaining connections is needed to facilitate continual learning.
- Ability to see connections between fields, ideas, and concepts is a core skill.
- Currency (accurate, up-to-date knowledge) is the intent of all connectivist learning activities.
- Decision making is itself a learning process.

Connectivist learning skills are required to learn within a navigationist learning paradigm. And this is why Brown (2006) states that "connectivism is part and parcel of navigationism," (p. 117) a learning paradigm that needs further development. The main practical implication of Brown's work is that teachers and trainers should become coaches and mentors within the knowledge and digital era and learners should acquire navigating skills for a navigationist learning paradigm. To enhance e-learning Web 2.0 over time, it is vital for instructors to ground their designs on established learning theories and report how related learning experiences are integrated with Web 2.0 tools so instructors can determine what Web 2.0 tools have the greatest effect on learner motivation and performance. E-learning Web 2.0 is the key solution to equipping people with the evolving knowledge and skills that will be needed to adapt to the continuously changing nature of the information society. At the same time, the major aim in education is to produce autonomous learners. For Franklin and van Harmelen (2007, p. 21) "the growing Personal Learning Environment (PLE) movement has a significant Web 2.0 following which claims that PLEs are social software

Table 2. Comparison of characteristics of Web 1.0 versus Web 2.0 educational Websites. Source: McGee and Begg (2008, p. 167)

Web 1.0	Web 2.0
Course Websites using content management systems. An expert (course director) produces a syllabus which resides on a curriculum Website. Single Website, which displays the same content and design for all users. Posting problem based learning cases to a curriculum Website.	Faculty blogs, student discussion groups. Podcasts. Students in a course contribute to syllabus content with questions and answers to supplement expert materials. Personal Websites, with customized data sources and layout for individual users Small groups have their own Website to which they add learning objectives and educational content related to their coursework

tools that help or enable learners to take control of their own education" and learning processes throughout their lives.

POSSIBILITIES AND APPLICATIONS OF WEB 2.O TOOLS

As Owen, Grant, Sayers, and Facer (2006) state "Web 2.0 will lead to e-Learning 2.0, to a rethinking of the relationship between technology and learning, to the development of educational practices that place the learner at their heart through the creation of collaborative, community-based learning experiences. To explore this further we touch now on the key theme of the potential shift in thinking from 'e-learning' to 'c-learning' (p. 10). Virtual communities of learning also offer the promise of bridging the worlds of work and education.

Some of the key attributes of social software in relation to education are that it (Owen et al., 2006): "Delivers communication between groups, enables communication between many people, provides gathering and sharing resources, delivers collaborative collecting and indexing of information, allows syndication and assists personalization of priorities, has new tools for knowledge aggregation and creation of new knowledge and, delivers to many platforms as is appropriate to the creator, recipient and context".

To help apply Web 2.0 to education McGee and Begg (2008, p. 167) summarize briefly the key differences between Web 1.0 and Web 2.0 (see Table 2). The new user-centered paradigm in which users are both producers and consumers of content and services has evolved from previous Web developments. The Web before the dot.com crash is usually referred to as Web 1.0. O'Reilly (2005) cites a number of examples of how Web 2.0 can be distinguished from Web 1.0, such as Web 1.0 was mainly a platform for information, but Web 2.0 is also a platform for participation. Web 1.0 tools can be used for the delivery of the course materials and for communication but Web 2.0 tools (such as blogs) can be integrated in a e-learning environment to a shift from a "knowledge transfer model" to a "knowledge construction model" as presented by Virkus (2008).

If one were to apply Web 2.0 concepts, "the lecture notes could become wikis (Wikipedia), the slides would become an image sharing collection (akin to Flick®), and students would subscribe to audio and video recordings (on a site like iTunes™), ideally all within an integrated ''virtual learning environment.'' This online environment would allow students to create their own views of their learning material and combine, with their own notes and external information resources. In Web 2.0 parlance this is a ''mash up,'' where content from different sources is combined by a user to create something new" (McGee & Begg, 2008, p. 167). Web 2.0 is suitable for educational and lifelong learning, because our knowledge society is built on digital environments of work and social communication, and educational practices must

Table 3. Educational applications of Web 2.0. Source: Franklin & Van Harmelen (2007, pp. 5-7)

Web 2.0 tool	Description	Educational application
Blogs	A system that allows an author to publicly display time-ordered articles.	A blogger can build up a corpus of interrelated knowledge. Teachers can use a blog for course announcements, news and feedback. Blogs can be used with syndication technologies to enable groups of learners and teachers to easily keep track of new posts.
Wikis	A system that allows one or more people to build up a corpus of knowledge in a set of interlinked Web pages.	Wikis can be used for the creation of annotated reading lists by one or more teachers Wikis are suited to the incremental accretion of knowledge by a group, or production of collaboratively edited material.
Social bookmarking	It provides users the ability to record (bookmark) Web pages, and tag those records with significant words (tags) that describe the pages being recorded.	To build up collections of resources. Groups of users with a common interest can team together to use the same bookmarking service to bookmark items of common interest.
Media-sharing services	Sstore user-contributed media that allows users to search for and display content. Compelling examples include YouTube™ (movies), iTunes® (podcasts and vidcasts), Flickr® (photos), Slideshare (presentations), DeviantArt (art work) and Scribd (documents).	Podcasts can be used to record lectures Podcasts can be used to supply audio tutorial material Instructional videos and seminar records can be hosted on video sharing systems.
Social networking and social presence systems	Systems that allow people to network together for various purposes, such as Facebook© and MySpace® (for social networking / socialising), LinkedIn® (for professional networking), Second Life™ (virtual world) and Elgg (for knowledge accretion and learning).	LinkedIn® acts, at a professional level, as a model of educational use in the way in which it can be used to disseminate questions across the community for users seeking particular information. There are a wide variety of educational experiments being carried out in Second Life.
Collaborative editing tools	These allow users in different locations to collaboratively edit the same document at the same time, such as Google™ Docs & Spreadsheets.	For collaborative work over the Web.
Syndication and notification technologies	A world of newly added and updated shared content. A feed reader (or aggregator) is used to centralize all the recent changes in the sources of interest, and a user can easily use the reader/aggregator to view recent additions and changes. This relies on protocols called RSS (Really Simple Syndication) and Atom to list changes (these lists of changes are called feeds, giving rise to the name feed reader).	Feed Readers enable students and teachers to become aware of new blog posts in educational blogging scenarios, to track the use of tags in social bookmarking systems, to keep track of new shared media, and to be aware of current news

foster a creative and collaborative engagement of learners with this digital environment in the learning process (Guntram, 2007, p. 17).

Table 3 summarizes some educational applications of Web 2.0 tools included in Franklin and Van Harmelen's (2007) work.

To embed Web 2.0 tools and processes within mainstream higher education practice the following need to be in place (Collis & Moonen, 2008, p. 100):

- Both instructors and students must value an educational approach where learner participation and contribution are balanced with acquisition.
- A pedagogical approach must be used that reflects contribution-oriented activities where students create some of their own learning resources.
- The approach must be scaffolded in practice by interlinked support resources for

both instructors and students. Uncertainty must be reduced as much as possible for the students in terms of what is expected of them, and to what standard.

- The processes as well as the products produced by the students must be assessed as part of overall course assessment practices.

In higher education and adult learning educational applications of Web 2.0 tools add extra value to the learning experience and have an unlimited potential. So far, we have briefly summarized the increasingly varied ways in which these new tools can be used to construct the navigationist learning paradigm. This new learning paradigm 2.0 represents an opportunity to revolutionize the way human beings learn, interact, innovate and develop.

FUTURE TRENDS

Different subjects need to be explored in detail to step up research —educational, socio-economic and technological — in the field of e-learning 2.0 and in the use of Web 2.0 tools in higher education and adult learning.

- Special attention need to be devoted to using emerging technologies (GRID, Web 3.0) for the development of innovative applications for education and training. In this new technological environment, the question of how to motivate and socialize the student as an active learner needs also to be raised. As Hvid and Godsk (2006) state "e-learning platforms needs an aesthetic perspective instead of mainly addressing usability and function". (p. 210)
- In the near future, portable and personal technologies will offer new opportunities to connect people and to create new e-learning 2.0 environments. We are only beginning to understand the opportunities that mobiles technologies provide for learning. As Wilson (2006) points out "Web platforms that allow moblogging (blogging from mobile phones), vlogging (video blogging) and other forms of 3G-enabled participation are increasingly popular and show clearly the potential for user-generated 3G content to be integrated in an architecture of participation" (p. 239). Mobile technology will play a key role in the new e-learning 2.0 paradigm.
- e-Learning 2.0 indicators need to be further developed in order to monitor progress in the use of Web 2.0 in formal and informal education.
- Education methods, learning communities organization are essential aspects in this context.
- Research also needs to provide a holistic view of students' actual use of the social software in higher education and adult learning.
- Another key issue for any future research is to explore what forms of knowledge students obtain from social software and, most importantly, how students use such knowledge. In-depth qualitative research should be carried out to understand how is built through Web 2.0 technologies.
- The concept of virtual campus and virtual networks for cooperation and collaboration needs to be revisited.
- In the virtual world, social networking functions can enable learners to aggregate into communities of interest and evolve into communities of learning or practice. We need to understand the formation of these communities and ways to facilitate the contribution of cybersocial networking to the learning and engagement of students and teachers (Computing Research Association, 2005).

- E-learning Web 2.0 may be able to reach learners who are disadvantaged by the digital divide. If it also important to define a research agenda which takes into account individual differences in learning, and special needs education to exploit the potential of Web 2.0 technology to provide remedial measures in the case of disability, exclusion, difficulty in gaining access to learning, or where conventional education does not work.
- Finally, special attention needs to be given to the promotion of gender equality in building e-learning 2.0 communities and social capital.

Bearing in mind all these agendas, e-learning 2.0 are likely to be a fertile research field.

CONCLUSION

Each new wave of technological innovation promises to revolutionize education, as we know it. The emergence of e-learning Web 2.0 is currently affecting most colleges, universities, and corporations. Now it is time to step back and question the pedagogical principles that inform our learning paradigms because Web 2.0 technologies have to be implemented taking into account pedagogical perspectives. The use of Web 2.0 technologies in higher education and adult learning is still a new technological phenomenon which will only "become valuable in education if learners and teachers can do something useful with it" (OECD, 2001, pp. 24-25).

REFERENCES

Abell, A., Chapman, D., Phillips, P., Stewart, H., & Ward, S. (2006). Roles in the e-landscape: Who is managing information? *Business Information Review*, *23*(4), 241–251. doi:10.1177/0266382106072249

Borger, R., & Seaborne, A. (1966). *The psychology of learning*. Harmondsworth: Penguin Books.

Brown, T. H. (2006). Beyond constructivism: Navigationism in the knowledge era. *Horizon*, *14*(3), 108–120. doi:10.1108/10748120610690681

Bruckman, A. (1998). Community support for constructionist learning. *Computer Supported Cooperative Work: The Journal of Collaborative Computing*, *7*, 47–86. doi:10.1023/A:1008684120893

Collis, B., & Moonen, J. (2008). Web 2.0 tools and processes in higher education: Quality perspectives. *Educational Media International*, *45*(2), 93–106. doi:10.1080/09523980802107179

Computing Research Association. (2005). *Cyber infrastructure for education and learning for the future: A vision and research agenda*. Retrieved September 15, 2008, from http://www.cra.org/reports/cyberinfrastructure.pdf

Downes, S. (2005). e-Learning 2.0. *eLearn Magazine: Education and Technology in Perspective*. Retrieved September 15, 2008, from http://www.elearnmag.org/subpage.cfm?section=articles&article=29-1

European Commission. (2001). *The elearning action plan. Designing tomorrow's education*. Retrieved September 15, 2008, from http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2001:0172:FIN:EN:PDF

Foster, A. (2008). Business information survey. *Business Information Review*, *25*(1), 13–31. doi:10.1177/0266382107088221

Franklin, T., & Van Harmelen, M. (2007). *Web 2.0 for content for learning and teaching in higher education*, JISC, Bristol. Retrieved September 15, 2008, from http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/Web2-content-learning-and-teaching.pdf

Frick, T. (1991). *Restructuring education through technology*. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Gunasekaran, A., McNeil, R. D., & Shaul, D. (2002). E-learning: Research and applications. *Industrial and Commercial Training*, *34*(2), 44–53. doi:10.1108/00197850210417528

Guntram, G. (Ed.). (2007). *Open educational practices and resources. OLCOS Roadmap 2012*. Salzburg Research EduMedia Group, Salzburg. Retrieved September 15, 2008, from, available at: http://www.olcos.org/cms/upload/docs/olcos_roadmap.pdf

Hudson, B. (2003). Promoting collaboration in and international online learning community. *Industrial and Commercial Training*, *35*(3), 88–93. doi:10.1108/00197850310470294

Hvid, M., & Godsk, M. (2006, June). *The pleasure of e-learning - Towards aesthetic e-learning platforms*. In Proceedings of the 12th International Conference of European University Information Systems, University of Tartu & EUNIS, Tartu, Estonia, (pp. 210-212).

Kiser, K. (1999). 10 things we know so far about online training. *Training (New York, N.Y.)*, *36*(11), 66–74.

McGee, J. B., & Begg, M. (2008). What medical educators need to know about "Web 2.0". *Medical Teacher*, *30*(2), 164–169. doi:10.1080/01421590701881673

McLoughlin, C., & Lee, M. J. W. (2007, December). S*ocial software and participatory learning: Pedagogical choices with technology affordances in the Web 2.0 era*. In ICT: Providing choices for learners and learning. Proceedings ASCILITE (pp. 664-675), Singapore. Retrieved September 15, 2008, from http://www.ascilite.org.au/conferences/singapore07/procs/mcloughlin.pdf

Mejias, U. (2005). *A nomad's guide to learning and social software.* Retrieved September 15, 2008, from http://knowledgetree.flexiblelearning.net.au/edition07/download/la_mejias.pdf

O'Reilly, T. (2003). *Architecture of participation*. Retrieved September 15, 2008, from http://www.oreillynet.com/pub/wlg/3017

O'Reilly, T. (2005). *What is Web 2.0?* Retrieved September 15, 2008, from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-Web-20.html?page=1

OECD. (2001). *OECD E-learning: The partnership challenge*. Paris: OECD.

Owen, M., Grant, L., Sayers, S., & Facer, K. (2006). *Social software and learning.* FutureLabs. Retrieved September 15, 2008, from http://www.futurelab.org.uk/research/opening_education.htm

Piaget, J. (1967). *Six psychological studies*. New York, NY: Vintage Books.

Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon*, *9*(5). Retrieved September 15, 2008, from http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf

Rice-Lively, M. L. (2000). Bordeless education at UT-Austin GSLIS. *Texas Library Journal*, *76*(2), 58–60.

Russell, T. L. (1999). *The no significance difference phenomenon*. Raleigh, NC: North Carolina State University Press.

Selwyn, N. (2008). An investigation of differences in undergraduates' academic use of Internet. *Active Learning in Higher Education*, *9*(1), 11–22. doi:10.1177/1469787407086744

Siemens, G. (2004). *Connectivism: A learning theory for the digital age*. Retrieved September 15, 2008, from http://www.elearnspace.org/Articles/connectivism.htm

Trondsen, E. (2006). *Virtual worlds for learning and training*. Menlo Park, CA: SRI Consulting Business Intelligence.

Virkus (2008). Use of Web 2.0 technologies in LIS education: experiences at Tallinn University, Estonia. *Program: Electronic Library and Information Systems, 42*(3), 262-274.

Wilson, J. (2006). 3G to Web 2.0? Can mobile telephony become an architecture of participation? *Convergence: The International Journal of Research into New Media Technologies*, *12*(2), 229–242. doi:10.1177/1354856506066122

KEY TERMS

Collaborative Learning: An educational approach based the idea that learning is a naturally social act. The learner actively constructs knowledge by formulating ideas into words, and these ideas are built upon through reactions and responses of others. In other words, collaborative learning is not only active but also interactive. It is a student-centered approach in which social software tools are currently used for building and sharing knowledge.

Connectivism: A learning theory for the digital era. It is based upon the idea that knowledge is networked and so the act of learning takes place inside virtual networks and communities through social interaction. It is a networked model of learning.

E-Learning (electronic learning): Technology-supported learning and delivery of content via all electronic media. These may include Internet, intranets, computer-based technology, or interactive television. They may also include the use of e-technology to support traditional methods of learning, for example using electronic whiteboards or video conferencing. This terms covers a wide set of applications and processes, such as Web-based learning, computer-based learning, virtual classrooms, and digital collaboration.

Personal Learning Environments (PLE): A learning environment in which learners manage their own learning by selecting, integrating and using various software tools and services. It takes advantages of Web 2.0 affordances such as collaborative information and knowledge sharing.

Social Capital: A cross-disciplinary concept referring to the benefits of social networks and connections. Social capital is constructed and maintained in the interaction between individuals or groups. Social networks promote different types of social capital: bonding –referring to horizontal ties between individuals-, bridging – referring to ties that cut across different communities- or linking –referring to vertical ties.

Social Software: Software that allows the creation of communities and resources in which individuals come together to learn, collaborate and build knowledge. It is also known as Web 2.0 and it supports social interaction and collaborative learning. Current typical examples include Flickr® and YouTube™ –as audiovisual social software.

Virtual Learning Environments (VLE): A set of teaching and learning tools designed to enhance a student's learning experience by including computers and the Internet in the learning process. The principal components of a VLE package include curriculum mapping, student tracking, online support for both teacher and student, electronic communication, and Internet links to outside curriculum resources. There are a number of commercial VLE software packages available, including Blackboard, WebCT, Lotus® LearningSpace, and COSE.

This work was previously published in Handbook of Research on E-Learning Applications for Career and Technical Education: Technologies for Vocational Training, edited by V. C.X. Wang, pp. 779-790, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 3.9
SWELS: A Semantic Web System Supporting E-Learning

Gianluca Elia
University of Salento, Italy

Giustina Secundo
University of Salento, Italy

Cesare Taurino
University of Salento, Italy

ABSTRACT

This chapter presents a prototypal e-learning system based on the Semantic Web paradigm, called SWELS (Semantic Web E-Learning System). The chapter starts by introducing e-learning as an efficient and just-in-time tool supporting the learning processes. Then a brief description of the evolution of distance learning technologies will be provided, starting from first generation e-learning systems through the current Virtual Learning Environments and Managed Learning Environments, by underling the main differences between them and the need to introduce standards for e-learning with which to manage and overcome problems related to learning content personalization and updating. Furthermore, some limits of the traditional approaches and technologies for e-learning will be provided, by proposing the Semantic Web as an efficient and effective tool for implementing new generation e-Learning systems. In the last section of the chapter, the SWELS system is proposed by describing the methodology adopted for organizing and modeling its knowledge base, by illustrating its main functionalities, and by providing the design of the tool followed by the implementation choices. Finally, future developments of SWELS will be presented, together with some remarks regarding the benefits for the final user in using such system.

DOI: 10.4018/978-1-60566-034-9.ch006

INTRODUCTION

In a context of rapid environmental and technological change, characterized by an increasing obsolescence of knowledge, organizations need to accelerate the renewal and to increase the effectiveness of their managerial competences. Such continuous change is a determinant of continuous learning processes that calls for the capacity to organize at all levels of the organization new working processes that have to be more knowledge intensive, multidisciplinary, and collaborative.

This requires a profound rethinking of the processes supporting the design, development, and delivery of learning (McCrea et al., 2000) in a way that the learning process becomes more effective, just-in-time, and customized.

As a consequence, learning should not be a passive activity which is only done when people are in the educational institutions without knowing how the knowledge is used in the real world. It should be a continuous and active process performed under a specified goal and situation where the knowledge is really needed. Moreover, as huge amount of knowledge becomes available through Internet in the information society, it becomes possible for people to access the knowledge they need when necessary. In such a circumstance, the most important thing about having a lot of knowledge is to know how to find the knowledge, to be ready to understand and master the new knowledge, and to create knowledge for future use to close the loop of knowledge production and consumption. For these reasons, the goal of education and learning should be augmented to include training of learning capability and creativity of the learners (Mizoguchi, 2000).

Such considerations are some prominent drivers of the e-learning. Since e-learning applications are accessible from anywhere at any time, ICT-based learning environment has been gaining increasing attention from the research community.

In this context, the recently emerged (VLE) Virtual Learning Environments revealed themselves very effective from the pedagogical point of view, especially if they are compared with the previous (CBT) Computer Based Training and (WBT) Web Based Training systems.

However, VLE did not completely solve the problems related to the organization and navigation of the learning materials. Indeed, most of the current Web-based learning solutions show some limits in accessing the right knowledge, as well as in the learning pattern navigation process (given that they do not allow a complete and multi-dimensional vision of the knowledge available, therefore users are obliged to follow the learning modules according to a linear path designed for a generic learner). In addition, there is the need to optimize the processes related to learning resource organization and aggregation, and the subsequent access and reuse of such resources with respect to a not scheduled learner profile.

Our focus here is on the creation of a Web-based learning environment that enables fast, just-in-time and relevant learning. Indeed, current Web-based solutions do not meet the above mentioned requirements, and some pitfalls are for example information overload, lack of accurate information, and content that is not machine-understandable.

These limits suggest the application of Semantic Web technologies (Barnes-Lee, 2000) to e-learning as means for implementing new generation e-learning systems. The Semantic Web technologies support the innovation process in a learning environment, exploiting the opportunity to create and manage data that are *machine understandable* and not *only machine readable* (Secundo et al., 2004).

An effective way to apply the Semantic Web approach to e-learning could be the use of the ontology backbone, which allows the *ontology-based* description of the learning materials (knowledge base), adding small semantic annotations to each learning object created (Nejdl, 2001). By using an ontology-based approach, learning resources can be easily organized into customized learning patterns and delivered on demand to the learner, according to her/his profile and knowledge needs.

Moreover, such an approach allows to virtuously combine the content description process with the content navigation one: content description to easily identify the learning resources required to achieve the desired learning goals; content navigation to minimize the required time for accessing the learning resources by adopting the right approach of exploring the learning space.

Therefore, according to this approach and to the alignment between e-learning and Knowledge Management we present an application of the KIWI approach called SWELS (Semantic Web E-Learning System). This tool is a prototypal e-learning system based on the Semantic Web paradigm which main functionalities are:

- The creation of an ontology-based view;
- The semantic representation and visualization of learning modules (knowledge base);
- Learning modules (knowledge base) accessing;
- The visualization of the structure of the ontology.

Moreover, SWELS provides an innovative functionality to learners--the opportunity to navigate a domain ontology explicitly. By explicitly navigating the domain ontology, learners not only have the direct access to the knowledge they need inside the knowledge base, but also they are empowered in reaching the following goals (Secundo et al., 2004):

1. The complete exploration of the knowledge base, keeping the awareness and the visibility of the learning path performed to reach the extracted knowledge;
2. The gradual, but deep, understanding of the semantic structure of the knowledge domain they are exploring, through the comprehension of the meanings of concepts and relations of the ontology.

The chapter starts by introducing e-learning as an efficient and just-in-time tool supporting the learning processes, arisen from the learning requirements of the new, dynamically changing knowledge society. Then a brief description of the evolution of distance learning technologies will be provided, starting from first generation of e-learning systems (CBT) through the current Virtual Learning Environments and MLEs (Managed Learning Environments), by underling the main differences between them and the need to introduce standards for e-learning with which to manage and overcome problems related to learning content personalization and updating. Furthermore, some limits of the traditional approaches and technologies for e-learning will be provided, especially referring to the knowledge organization and access as well as to the learning content navigation. At this point, the Semantic Web will be proposed as an efficient and effective tool for implementing new generation e-learning systems, since that the application of such technology to e-learning provides an ontology-based description and organization of learning materials around small pieces of semantically annotated learning objects. In the last section of the chapter the SWELS e-learning system is proposed. The description of such an innovative solution starts with some insights on the methodology adopted for organizing and modeling the Knowledge Base of SWELS, then the main functionalities of the e-learning system will be illustrated; following, the design of the tool together with the implementation choices will be provided; finally, future developments and some remarks will be presented regarding the benefits for the final user (the learner) in using such system.

E-LEARNING: A TECHNOLOGY FACILITATING THE LEARNING PROCESSES

Learning is a critical support mechanism for organizations to compete, not only from the point of view of education, but also from the point of view of the New Economy (Drucker, 2000). The incredible velocity and volatility of today's markets require just-in-time methods for supporting the need-to-know of employees, partners, and

Table 1. Summary of problems and needs in education (Adapted from: Koper, 2004)

Dimension	*Problems/Needs*
I. Changes in Societal Demands	1. Current higher education infrastructure cannot accommodate the growing college-aged population and life-long learning enrolments, making more distance education programs necessary. 2. Knowledge and information are growing exponentially and Lifelong learning is becoming a competitive necessity. 3. Education is becoming more seamless between high school, college, and further studies.
II. Changes in Learning Teaching process	4. Instruction is becoming more learner-centred, non-linear, and self-directed. 5. There is an increasing need for new learning and teaching strategies that a) is grounded in new instructional design research and b) exploit the capabilities of technology. 6. Learning is most effective when learners are engaged in solving real-world problems; learning environments need to be designed to support this problem-centred approach. 7. Students demand more flexibility; they are shopping for courses that meet their schedules and circumstances. 8. Higher-education learner profiles, including online, information-age, and adult learners, are changing. 9. Academic emphasis is shifting from course-completion to competency. 10. The need for faculty development, support, and training is growing. 11. Instructors of distance courses can feel isolated.
III. Changes in Organization of Educational Institutions	12. There is a shift in organizational structure toward decentralization. 13. Higher education outsourcing and partnerships are increasing. 14. Retention rates and length of time taken to completion concern administrators, faculty members, students and tax payers. 15. The distinction between distance and local education is disappearing. 16. Faculty members demand reduced workload and increased compensation for distance courses. 17. Traditional faculty roles are shifting or unbundling.

distribution paths.

Time, or the lack of it, is the reason given by most businesses for failing to invest in learning. Therefore, learning processes need to be fast and just-in-time. Speed requires not only a suitable content of the learning material (highly specified, not too general), but also a powerful mechanism for organizing such material. Also, learning must be a customized on-line service, initiated by user profiles and business demands. In addition, it must be integrated into day-to-day work patterns and needs to represent a clear competitive edge for the business. In a few words, learning needs to be relevant to the (semantic) context of the business of people and organizations (Adelsberger et al., 2001).

In this scenario, Web-based learning environments have been gaining increasing attention from the research community, since e-learning applications can represent real facilitator of the learning processes both in business and in academic contexts. The following table (Table 1) underlines some problems and needs that can be effectively overcome with e-learning; such problems and needs are summarized and grouped on several e-learning domain dimensions (Koper, 2004):

But, what does e-learning mean? E-learning is the use of Internet technologies to create and deliver a rich learning environment that includes a broad array of instruction and information resources and solutions, the goal of which is to enhance individual and organizational performance (Rosenberg, 2006). E-learning is not just concerned with providing easy access to learning resources, anytime, anywhere, via a repository of learning resources, but is also concerned with supporting such features as the personal definition of learning goals, and the synchronous and asynchronous communication, and collaboration, between learners and between learners and instructors (Kolovski et al., 2003). It aims at replacing old-fashioned time/place/content predetermined learning with a just-in-time/at work-place/customized/on-demand process of learning (Stojanovic et al., 2001).

Traditional learning process could be characterised by centralised authority (content is selected by the educator), strong push delivery (instructors push knowledge to students), lack of a personalisation (content must satisfy the needs of many), and the linear/static learning process (unchanged content). The consequences of such organisation on the learning are expensive, slow and too unfocused (problem-independent) learning process.

But dynamically changed business environment puts completely different challenges on learning process--fast, just-in-time (cheap) and relevant (problem-dependent) learning, as mentioned above. This can be solved with the distributed, student-oriented, personalised, nonlinear/dynamic learning process--e-learning. The principle behind e-learning is that the tools and knowledge needed to perform work are moved to the knowledge workers--wherever and whoever they are.

In the recent years, new breeds of IS (Information System) known as LMS (Learning Management Systems) and LCMS (Learning Content Management Systems) are evolving to enable learning in organisations (Brennan et al., 2001). In essence, LMS replace isolated and fragmented learning programmes with a systematic means of assessing and raising competency and performance levels throughout the organisation, by offering a strategic IS solution for planning, delivering, and managing all learning events, including both online and classroom-based learning (Greenberg, 2002). LMS are often coupled with LCMS which facilitate the management and administration of the learning content for the online learning prgrammes in the form of learning objects (Brennan et al., 2001).

E-LEARNING SYSTEMS: FROM VIRTUAL LEARNING ENVIRONMENTS TO MANAGED LEARNING ENVIRONMENTS

In the 90s the primary impact of the Internet technologies on distance learning was mainly on the possibility of having different ways for aggregating and delivering learning content. Indeed, the application of such technologies to the learning processes has introduced a set of opportunities and advantages: the possibility to generate and transport on the Web multimedia audio/video flows at low costs (therefore promoting the diffusion of synchronous learning environments on asynchronous ones); the use of standard technologies for information exchange that allow to dynamically and effectively structure and navigate learning content; the possibility for learners to acquire knowledge and to continuously revise and update it by adapting the learning environment to their needs. In other words, Web-based learning environments shifted from stand-alone technologies towards highly integrated e-learning and knowledge management infrastructures and tools enabling the creation of learning communities and supporting the collaboration between members and organizations. However, in the last years Internet technologies failed in the process of creating and managing learning contents in a way that they could be easily and dynamically reused and updated. This because of the inability of trainers and learning managers to create learning materials that could be fast and easily adapted to the learning needs of learners as well as to the new ways of content delivery. Such considerations have been driven the shift from first generation e-learning systems, based on the delivery of Web-based learning content and on the basic Internet standards, towards second generation ones, based on "ad-hoc" e-learning standards (Damiani et al., 2002).

Second generation e-learning systems are based on the creation of VLEs that have risen from the integration between e-learning and knowledge management solutions. The primary aim of VLEs is to allow people to share knowledge, interests and experiences, thereby encouraging the creation of Virtual Learning Communities based on blended learning solutions in which face-to-face and virtual classroom meetings are combined with Web-based learning patterns to provide to learners a complete, interactive and effective learning experience. Nevertheless, if from one side VLEs revealed themselves very effective from the pedagogical point of view (especially if considered in relation to the e-learning platform of first generation), from the other side they showed some limits as regards to the problem of learning content aggregation and organization, and the subsequent access and reuse by learners with a non scheduled user profile. The classification and management of contents are the strength points of the so-called MLE. MLEs privilege the content design, creation, and management in respect to the content delivery infrastructure, considered as an element with which the content has to interoperate. The main goal of MLEs is to manage in an integrated way a complete system for analyzing, developing, and evaluating competences, for scheduling and organizing learning patterns, for managing roles and virtual classrooms, for tracking the learners and for final evaluation of the competences reached (Secundo et al., 2004). The complete separation proposed by MLEs between the management infrastructure and the final output of the learning material is enabling the development and the diffusion of standards for e-learning in several applicative contexts (Lockwood et al., 2001):

- DRM (Digital Right Management) and privacy management;
- Low level formats for learning content;
- Metadata for content description
- Personalization of content according to the learner profile and to the linguistic/social/cultural environment;
- XML-based models and languages for structuring and describing dynamic learning patterns (i.e., Educational Modeling Language);
- Technologies and methodologies for interoperability with Internet/Intranet delivery infrastructures.

Nowadays there are several key international players (including IEEE, IMS, ARIADNE, ADL and AICC) that are focusing their efforts on the issues of interoperability and reuse, resulting in a multitude of standards that can be used for building interoperable learning systems. These attempts at building learning platforms for interoperability are mainly targeted to ease the need of LMSs for adaptation to standards, but as a consequence, learners can be expected to gain more freedom. For example, the goal of SCORM (Sharable Content Object Reference Model) is to provide a reference model for content that is durable (survives system changes), interoperable (between systems), accessible (indexed and searchable) and reusable (able to be modified in different contexts and by different tools). This will hopefully allow students to move more freely between LMSs and even to combine several services from different LMSs.

SOME LIMITS OF TRADITIONAL APPROACHES AND TECHNOLOGIES FOR E-LEARNING

Current approaches, models and technologies for e-learning introduce, on the other hand, several problems. First, most content providers have large monolithic systems where adaptation to standards will not significantly change the underlying teacher-learner model. Students will be presented with material in a context often leading up to some form of (standardized) test. New and more interesting methods for learning--such as techniques for collaboration, annotation, conceptual modeling, and so forth.--will not profit from such adaptation. Second, even though monolithic, closed or proprietary systems will be able to exchange learning resources, course-like structures and keep track of students with the help of those standards, they will need to go through yet another process of adaptation to the next big batch of agreements on learning technologies, such as e.g. profiling and tracking of student performance. Third, the current perspective on metadata is too limited. Anyone who has something to say about a learning resource should be able to do so. This includes learners, teachers and content contributors such as authors and providers. Communicating this metadata is equally important as it can help, direct or encourage others to actively participate and learn. Proposed solutions, such as the adoption of SCORM, will result in learning resources (and their metadata) that will reappear in different versions and formats rather than dynamically evolve and improve (Naeve et al., 2001).

In a few words, today many of the e-learning systems available on market lack in specific functionalities for the creation and delivery of dynamic, modular learning paths that match the knowledge needs in a contextualized (according to learner's current activities) and individualized (according to learner's experiences, competences profiles, learning history and personal preferences) way.

This suggests a strong integration among e-learning and knowledge management functionalities to define a rich learning environment with wealth and variety of resources available just-in-time to learners, both through structured and unstructured knowledge objects through interaction with other people (Elia et al., 2006). The key to success is therefore the ability to reduce the cycle time for learning and to adapt "content, size and style" of learning to the learner and to the business. Therefore, to overcome such problems, a new learning framework is required, based on the key points mentioned above and opened to a multitude of new services. In order to be effective, it needs a powerful language for expressing facts about resources and schemas that will allow machines as well as humans to understand how these facts are related without relying on heuristics. Moreover, there is a need for expressing facts about remote (identifiable) objects without accessing remote data stores.

EMERGING PERSPECTIVES OF E-LEARNING IN THE SEMANTIC AGE

In an e-learning environment, the learning content should be oriented around small modules (the so called learning objects) coupled with associated semantics (the metadata) so that learners are able to find what they need in a specific moment and context. Furthermore, these modules should be related by a "dependency network" or "conceptual Web" to allow individualised learning. Such a dependency network permits, for example, the learning objects to be presented to the learner in an orderly manner, with prerequisite material being presented first. Additionally, in an e-learning environment, the learner should be able to add extra material and links (i.e., annotate) to the learning objects for their own benefit or for that of later students. This framework lends itself to an implementation based on the Semantic Web, incorporating cooperating software agents, which additionally make use of appropriate Web services to provide the functionality. The facilities the applications based on these technologies can provide, include allowing e-learning content to be created, annotated, shared, and discussed, together with supplying resources such lecture notes, student portfolios, group projects, information pages, discussion forums, and question-and-answer bulletin boards. Moreover, such applications allow students to benefit from more interaction with their peers (for example, sharing resources found on the Web), as well as with the instructors and tutors, by also providing an easy way for sharing and archiving information, whether of general interest or specific to a group project they are involved in (Kolovski et al., 2003).

T first generation WWW was a powerful tool for research and education, but its utility is hampered by the inability of the users to navigate easily the huge amount of sources for the information they require. The Semantic Web is a vision to solve this problem. It is proposed that a new WWW architecture will support not only Web content, but also associated formal semantics (Barnes-Lee, 1998). The idea is that the Web content and the related semantics (or metadata) will be accessed by Web agents, allowing these agents to reason about the content and produce intelligent answers to user queries. The Semantic Web, in practice, comprises a layered framework: an XML layer for expressing the Web content (the structure of data); a RDF (Resource Description Framework) layer for representing the semantics of the content (the meaning of data); an ontology layer for describing the vocabulary of the domain; and a logic layer to enable intelligent reasoning with meaningful data (Stojanovic et al., 2001).

Within an e-learning framework, the Semantic Web provides the technology that allows a learning object to be (Naeve et al., 2001):

- **Described with metadata.** Since a resource can have uses outside the domains foreseen by the provider, any given description (metadata instance) is bound to be incomplete. Because of the distributed structure of RDF, a description can be expanded or new descriptions following new formats (schemas) can be added. This allows for creative uses of content in new and unforeseen ways. Hence, one of the most important features of the current Web - the fact that anyone can link anything to anything--has been carried over into RDF.
- **Annotated.** Every resource identifiable by an URI can be annotated, with personal notes and links by anyone.
- **Extended.** In terms of content (structured, by means of XML descriptors), permitting multiple versions to exist. Indeed, successive editing of the content can be done via special RDF-schemas allowing private, group consensus or author-specific versions of a common base document. The versioning history will be a tree with known and unknown branches, which can be traversed with the help of appropriate versioning tools.
- **Shared by, and communicated to, anyone** who has expressed an interest in such content. RDF is application independent. Since the metadata is expressed in a standard format, which is independent of the underlying schemas, even simplistic applications can understand parts of complex RDF graphs. If the learner's favourite tool does not support the corresponding schemas, it can at least present them in a rough graph, table or whatever standard form it has for describing resources and their properties. If more advanced processing software is available (such as logic engines), more advanced treatment of the RDF descriptions is possible.
- **Certified**. There is no reason why only big organizations should certify learning resources. Individuals, such as teachers, may want to certify certain content as a quality learning resource that is well suited for specific learning tasks.

Apart from these uses, it is possible to invent new schemas describing structures, personalization, results from monitoring and tracking, processes and interactions that can enrich the learning environment in various ways.

The key property of the Semantic Web architecture (common-shared-meaning, machine-processable metadata), enabled by a set of suitable agents, establishes a powerful approach to satisfy the e-learning requirements: efficient, just-in-time, and task relevant learning. Learning material is semantically annotated and for a new learning demand it may be easily combined in a new learning course. According to his/her preferences, a learner can find useful learning material very easily. The process is based on semantic querying and navigation through learning materials, enabled by the ontological background. So, the Semantic Web can be exploited as a very suitable platform for implementing an e-learning system, because it provides all means for the ontology development, the ontology-based annotation of learning materials, as well as their composition in learning modules and proactive delivery of the learning materials through e-Learning portals.

In the following table (Table 2), the most important characteristics (or pitfalls) of the traditional learning and improvements achieved using the e-learning environment are showed; furthermore, a summary view of the possibility to use the Semantic Web for realizing the e-learning requirements is presented (Drucker, 2000; Stojanovic et al., 2001).

An important aspect related to the use of Semantic Web in educational contexts is how to represent a course in a formal, semantic way so that it can be interpreted and manipulated by computers as well as humans (i.e., the creation and management of data that are machine understandable and not only machine readable). This process is known in the literature as "educational modeling." A semantic model is developed using a variety of methods: literature research, expert group discussions, validation sessions, and so forth, and the result is described with a formal modeling language, like UML. The UML class diagrams can be translated to RDF-Schema and/or OWL Web Ontology Language, depending on the richness of the model. XML-Schema's (XSD) and other semantic bindings like Topic Maps can also be generated from the UML models (Koper, 2004). A semantic representation of learning content provides efficient solutions to the following problems (Koper, 2004):

- The development of Web-based courses that are flexible, problem-based, nonlinear, incorporate multimedia and are adaptive to learner characteristics, is expensive and extremely time-consuming. A semantic framework can help the course developers in the structuring and integration of the development work. In addition, authoring and design support agents and tools could be created to help the developers to do their jobs more effectively and efficiently.
- An explicit notation of learning content can preserve and share knowledge about effective learning designs. It gives the possibility to build and share catalogues of effective learning patterns that can be communicated very precisely and can be adapted to other contexts, problems, and content.
- Instantiation of an e-learning course in current LMSs (Learning Management Systems) can be a time-consuming job that has to be repeated for every new run of the course. One has to assign users, create groups, but also has to set-up the communication and collaboration services (e.g., discussion forums, workspaces, etc.) mostly by hand. A representation of a course that includes a specification of the set-up of the services enables the automation of this instantiation process.
- When the representation of the learning material includes a semantic, higher level description of the interactive processes that occur during the learning process, software agents can interpret these to support learners and staff in managing the workflow of activities in learning. These agents can also support the filtering of the

Table 2. Differences between training and e-learning and main benefits of applying Semantic Web technologies to e-learning (Adapted from Drucker, 2000; Stojanovic et al., 2001)

Dimension	Training	e-Learning	Semantic Web
Delivery	Push--Instructor determines agenda	Pull--Learner determines agenda	Knowledge items (learning materials) are distributed on the Web, but they are linked to commonly agreed ontologies. This enables the creation of user-specific learning patterns, by semantic querying for topics of interest.
Responsiveness	Anticipatory--Assumes to know the problem	Reactionary--Responds to problem at hand	Software agents on the Semantic Web may use commonly agreed service language, which enables co-ordination between agents and proactive delivery of learning materials in the context of actual problems. The vision is that each user has his/her own personalised agent that communicates with other agents.
Access	Linear-- Pre-defined progression of knowledge	Non-linear --Allows direct access to knowledge in whatever sequence makes sense to the situation at hand	User can describe situation at hand (goal of learning, previous knowledge) and perform semantic querying for the suitable learning material. The user profile is also accounted for. Access to knowledge can be expanded by semantically defined navigation.
Symmetry	Asymmetric --Training occurs as a separate activity	Symmetric --Learning occurs as an integrated activity	The Semantic Web offers the potential to become an integration platform for all business processes in an organisation, including learning activities.
Modality	Discrete-- Training takes place in dedicated chunks with defined starts and stop	Continuous --Learning runs in the parallel loops and never stops	Active delivery of information (based on personalised agents) creates a dynamic virtual learning environment.
Authority	Centralized --Content is selected from a library of materials developed by the educator	Distributed --Content comes from the interaction of the participants and the educators	The Semantic Web will be as decentralised as possible. This enables an effective co-operative content management.
Personalization	Mass produced-- Content must satisfy the needs of many	Personalized--Content is determined by the individual user's needs and aims to satisfy the needs of every user	A user (using personalised agent) searches for learning material customised for her/his needs. The ontology is the link between user profile and needs, and characteristics of the learning material.
Adaptiveness	Static-- Content and organization/taxonomy remains in their original authored form without regard to environmental changes	Dynamic-- Content changes constantly through user input, experiences, new practices, business rules and heuristics	The Semantic Web enables the use of knowledge provided in various forms, by semantic annotation of content. Distributed nature of the Semantic Web enables continuous improvement of learning materials.

appropriate resources to be used during the performance of an activity.

- Adaptation to individual learner characteristics (i.e., his/her learner profile) is highly desirable, since learners have not the same learning pre-requisites, skills, aptitudes or motivations. However, such adaptation can only be done realistically when the adaptation is wholly or at least partially automated (therefore, including descriptions of the conditions for adaptation). Otherwise, it becomes a very demanding work for the learner and/or his/her learning manager.

- A semantic annotation of learning content enables and facilitates sharing and re-use of learning objects (that is one of the major objectives in the field of e-learning). This sharing and re-use is needed to make the content development process more efficient. On the contrary, if learning objects are not semantically represented, it might be hard to find them on local or remote repositories, hard to integrate them into new contexts and--relating to the problem of interoperability and learning object exchange among different LMSs--hard to interpret and structure them in the correct way.
- An explicit semantic representation can serve as a means to create more advanced and complex, but consistent learning designs than is possible without such a representation. This is a characteristic of any language with semantic that enables one to write, read, rewrite and share meaning (e.g., natural language).

SWELS: AN E-LEARNING SYSTEM BASED ON SEMANTIC WEB TECHNOLOGIES

According to the alignment between e-learning and knowledge management approaches, a prototypal e-learning system based on the Semantic Web paradigm has been implemented called SWELS. Such system has been designed and developed at the eBMS (e-Business Management Section) () of the Scuola Superiore ISUFI, University of Lecce (Italy) and it is the result of a research activity under the KIWI project.

This paper represents an extended version of a previous publication that the authors G. Secundo, A. Corallo, G. Elia G. Passiante (2004) published in the proceedings of the International Conference on Information Technology Based Higher Education and Training, May 29th--June 2th, 2004 Istanbul, Turkey.

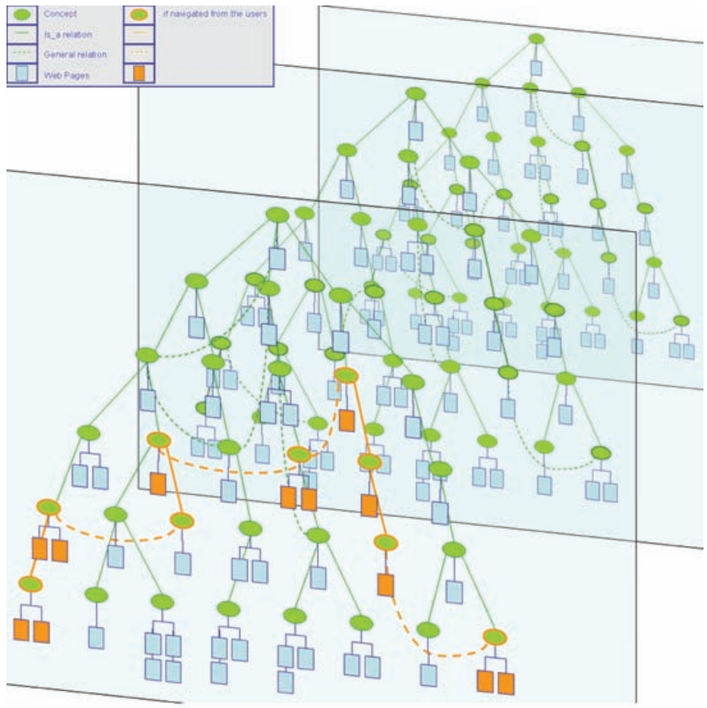
The SWELS system is intended to be an innovative tool for knowledge acquisition and competence development of learners and knowledge workers that exploits Semantic Web technologies in order to provide an effective and useful support to online learning processes. The system, indeed, is conceived as a tool with which to potentially overcome the limits of the current e-learning applications in terms of learning content creation and delivery, that is, the inability of existing tools to create dynamics and adaptive learning paths that match the learning profile of learners as well as their knowledge needs. SWELS points out a proactive behaviour based on a matching process among the profile of the user, his/her interests as well as his/her just-in-time choices during the learning activities, and the learning content available in the knowledge base; as a consequence learning resources can be easily organized into customized learning patterns and delivered on demand to the user.

Learning materials which SWELS refers to are focused on "Change Management and Leadership" knowledge domain, that has been modeled through a domain ontology. Such an ontology contains the list of concepts and semantic relations with which to provide a semantic description of the learning objects (text files, images, graphs, as well as multimedia audio-video files) of the domain.

KNOWLEDGE BASE ORGANIZATION AND MODELING

Learning materials (i.e., the knowledge base) are described by means of a domain ontology that provides a semantic representation of con-

Figure 1. A representation of knowledge base flexibility



tent, adding small semantic annotations to each learning resource. In particular, the knowledge base modeling process can be organized in two main steps:

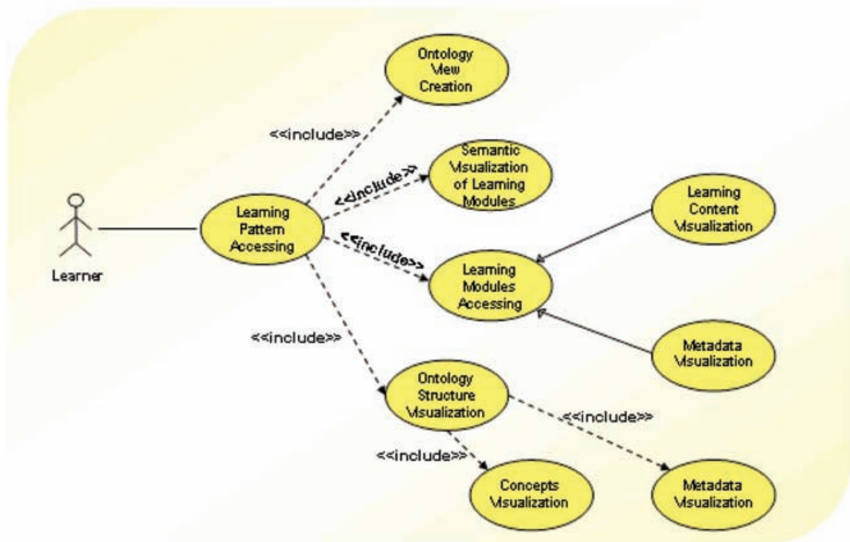
1. Definition of the knowledge base ontology. The ontology definition consists in identifying the learning module structure and defines the abstract notions and vocabulary that will be available for the learner to conceptualize the learning modules.
2. Description of the knowledge resources. Knowledge items are tagged with one concept belonging to the ontology. In this way, learner can identify each resource and, using the ontological relationships, he/she can explore new resources tagged by different domain concepts.

Such a description of learning content allows an effective organization of them in the knowledge base, therefore providing to learners the possibility to have an explicit navigation of the domain ontology.

The two main advantages for the final users are, from one hand a complete exploration of the modeled knowledge base, which allows them to have a total awareness of the available content, as well as the visibility of the performed learning path to reach the required knowledge. So, learners are conscious both of the total amount of knowledge present in the knowledge base, and of the knowledge extracted till then and of knowledge heritage to explore in the future. From the other hand, learners can understand, step by step, the semantic structure of the knowledge domain they're exploring by surfing the ontology, by gradually being aware of the meanings of ontology's concepts and relations.

This approach to the knowledge base organization and modeling provides more flexibility for learners as regard to the learning content

Figure 2. Use case diagram



access, since that they can explicitly browse the knowledge base and dynamically configure their learning patterns (Figure 1).

FUNCTIONAL DESCRIPTION OF SWELS

The interaction between the learner and the system can be represented through a use case diagram that shows the main functionalities of the tool (Figure 2).

As the use cases show, in order to have access to the dynamically created learning patterns, learners have to perform a set of different steps.

The following state chart diagram describes the overall behaviour of the system by underling the logic sequence of the states and the list of the state transitions related to the user events according to interaction between learner and system described before (see Figure 3).

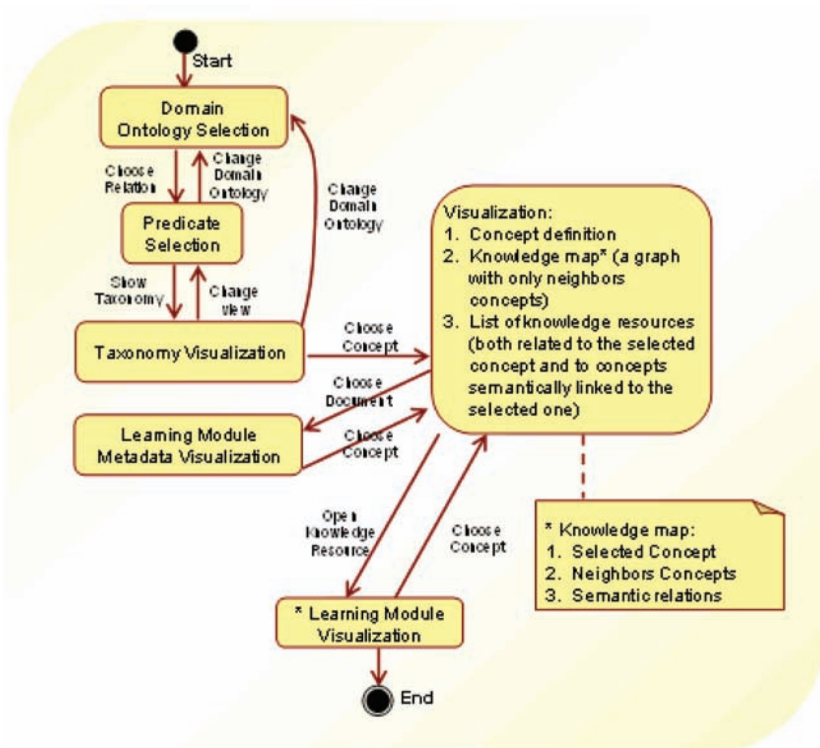
CREATION OF THE ONTOLOGY-BASED VIEW

When a learner accesses the SWELS, he/she has to select the domain ontology and the relation by which creating the ontology view. When predicate is chosen, the tool generates the taxonomic representation of the ontology, through a tree-structure (Figure 4).

SEMANTIC REPRESENTATION AND VISUALIZATION OF LEARNING MODULES

After the ontology view creation, the learner can generate his/her own personalized learning pattern by browsing the concepts of the ontology. By clicking on each concept, a list of elements will be shown:

Figure 3. State chart diagram



- Concept definition (top of the page);
- List of knowledge resources indexed on the selected concept (body of the page--with blank relation);
- List of knowledge resources indexed on concepts linked to the selected concept through one of the ontology relationships (body of the page – with specified relation).

Such information is organized in the tab "Documenti" (Knowledge Resources) as follows (Figure 5).

Ontology Structure Visualization

When a learner selects the specific concept which he/she is interested to, together with the semantic visualization of the learning modules, SWELS generates also a knowledge map containing both the selected concept and the neighbour concepts. Such a graph is organized in the tab "Grafico" (Graph) of the application, and represents the semantic boundary of the concept (specifying the neighbour concepts, the semantic connections and the direction of these connections). The semantic boundary is illustrated through a radial layout (neighbourhood view) --as TGViz one (a Plug-in of Protégé) and Visualizer one (a Plug-in of OntoEdit)--to give to the learner an explicit and immediate representation of the ontology structure (Figure 6). It is important to note that, referring to each triple (subject, predicate, object), the direction of the arrows connecting two concepts goes from the subject to the object of the triple; this allows the learner to have a unique interpretation of the semantic map.

In this way, two different and complementary representations of the domain ontology are available: the tree-structure (on the left of the page)

Figure 4. Ontology view creation

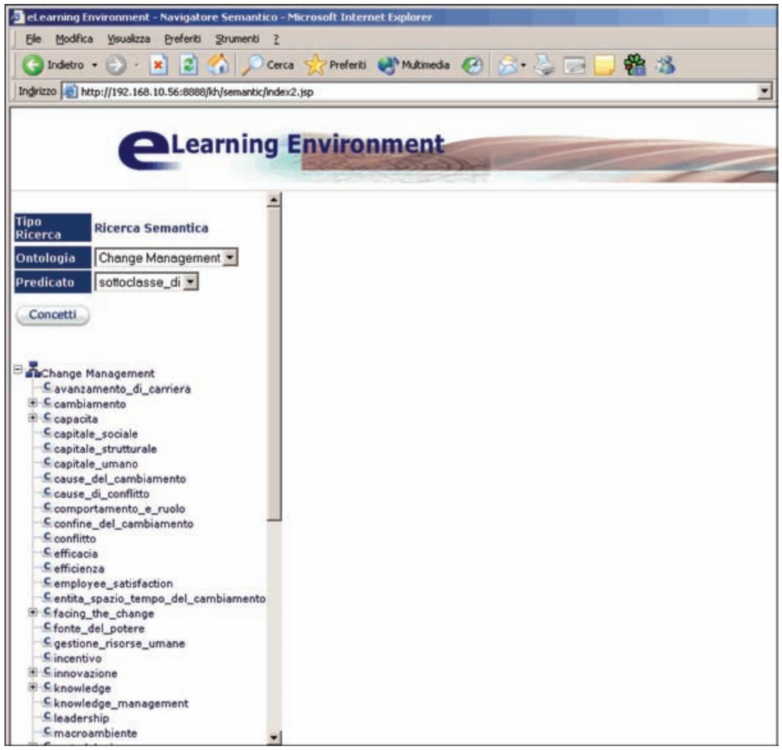


Figure 5. Semantic representation and visualization of learning modules

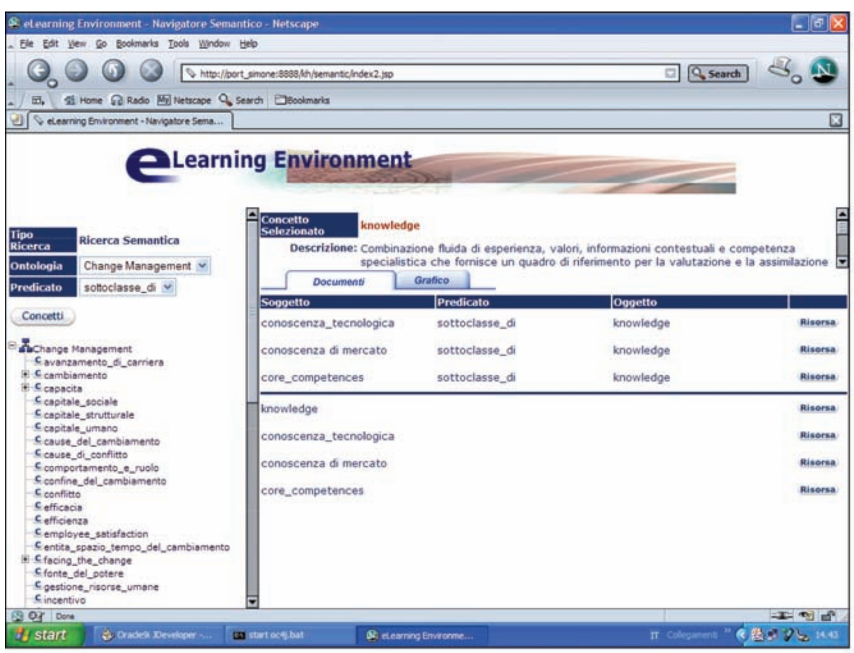
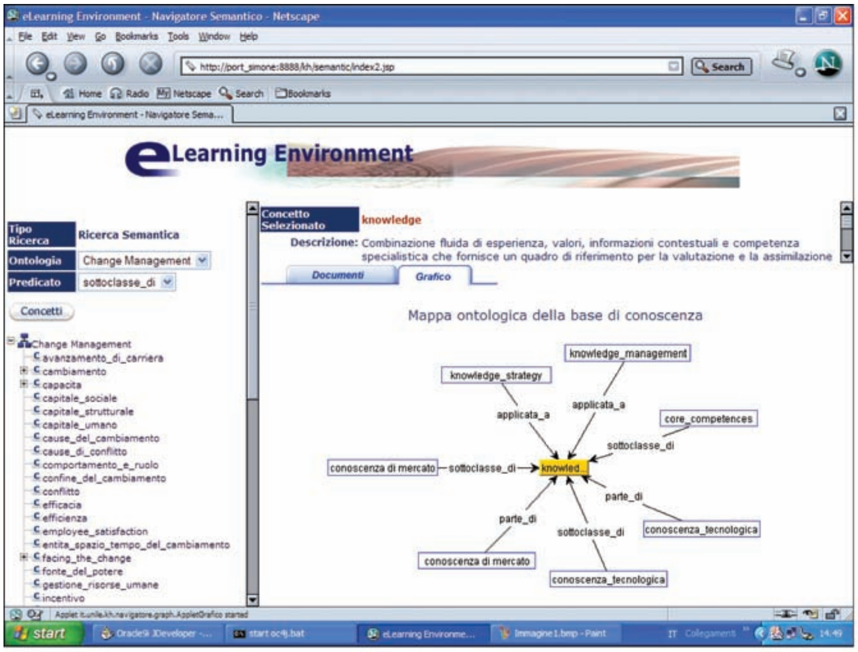


Figure 6. Ontology structure visualization



and the graph-structure (on the right). This choice allows a better understanding of the knowledge domain (since that it provides two different ways for representing the knowledge available in the domain) and gives the learner the opportunity to select and extract the right learning resources according to his/her "learning profile."

Learning Module Accessing

By clicking on the button "Risorsa" (Resource), in the list of the knowledge resources indexed in the tab "Documenti" (Figure 5), the learner has the direct access to the chosen learning module; in this way, the selected learning module will be launched in a new browser window and he/she can attend it autonomously (Figure 7).

Moreover, the learner can also access metadata describing knowledge resources, by clicking on resource name. In this way, Dublin Core metadata (Dublin Core Metadata Initiative, 2006) will be shown (Figure 8).

Design of The Tool

Concerning the design of SWELS, we decided to adopt the MVC (Model-View-Controller) design pattern (Figure 9) since that it allows enterprise applications to support multiple types of users with multiple types of interfaces. By representing the logic architecture of the system with such a "Three Tier" model, it is possible to keep separated core business model functionalities from the presentation and the control logic that uses those functionalities. Such separation permits multiple views to share the same enterprise data model, which makes supporting multiple clients easier to implement, test, and maintain (Sun Microsystems, Inc., 2002).

According to the Three Tier model adopted for the SWELS design, the first diagram proposed is the package diagram that shows developed class packages and the dependencies among them (Fowler et al., 1999) (Figure 10).

Figure 7. Learning module accessing

Going on in the description of the SWELS design, following are shown the class diagrams describing the types of the objects in the system and the various kinds of static relationships that exist among them. In particular, we propose the class diagram related to the ontology view creation (Figure 11), the class diagram related to the semantic visualization of the learning modules (Figure 12), and the class diagram related to the ontology structure visualization (Figure 13).

Technological Issues

With regard to implementation choices, SWELS is a J2EE Web-based application, developed according to the MVC (Model-View-Control) pattern (which implies together the use of Servlets as well as JSPs technologies), by using two suitable frameworks:

- *Jakarta Struts*, an open source framework for creating Java Web applications that utilize a MVC architecture. The framework gives three key components: a "request" handler provided by the application developer that is mapped to a standard URI, a "response" handler that transfers control to another resource which completes the response, a tag library that helps developers create interactive form-based applications with server pages (The Apache Software Foundation, 2006).
- *Oracle9iAS Toplink*, an ORM (Object Relational Mapping) framework for implementing the 'Model' layer that is free only for non-commercial applications.

Furthermore, the ontology is codified in RDFS and is stored in a relational database. The DBMS

Figure 8. E-learning metadata

is Oracle 9i; the relational database schema for the application is the following (Figure 15):

Finally, the standard for e-learning metadata is Dublin Core; the implementation of SCORM 1.2 is a work-in-progress.

These implementation choices give the tool a high level of flexibility and scalability; indeed, it can be used on several knowledge bases by developing a specified domain ontology and by exploiting the potentialities of ORM framework.

Empirical Evidence

During the exciting experience of conceptualization, design and implementation phases of SWELS, some attempts to validate the effectiveness of the whole approach were made. Specifically, we refer to a process realized on empirical evidence basis to acquire some insights for improving the overall system. This process is articulated in two test phases--the ‘alpha test’ phase and the ‘beta test’ phase.

The “Alpha test” phase was performed by the team involved in the implementation of the system itself. From one side, software developers tested

Figure 9. The MVC (Model-View-Controller) design pattern

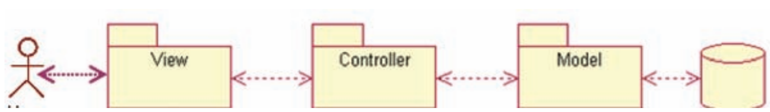


Figure 10. Package diagram

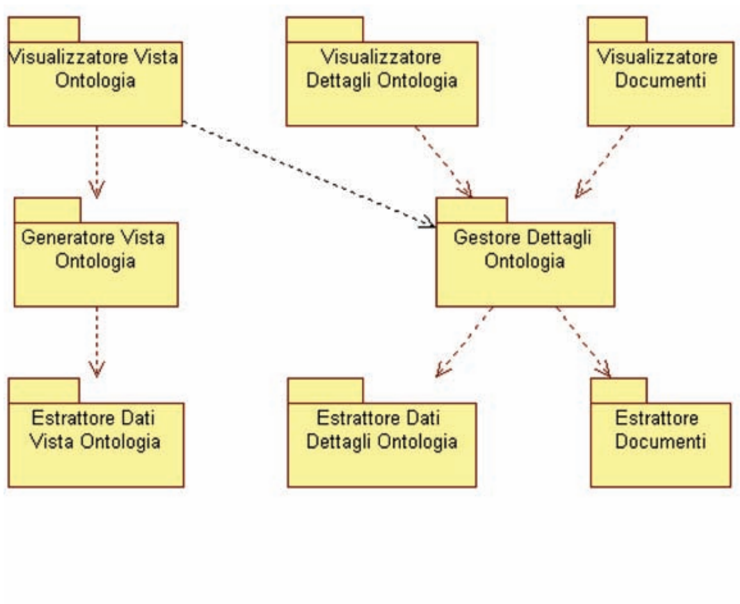


Figure 11. Class diagram for the ontology view creation

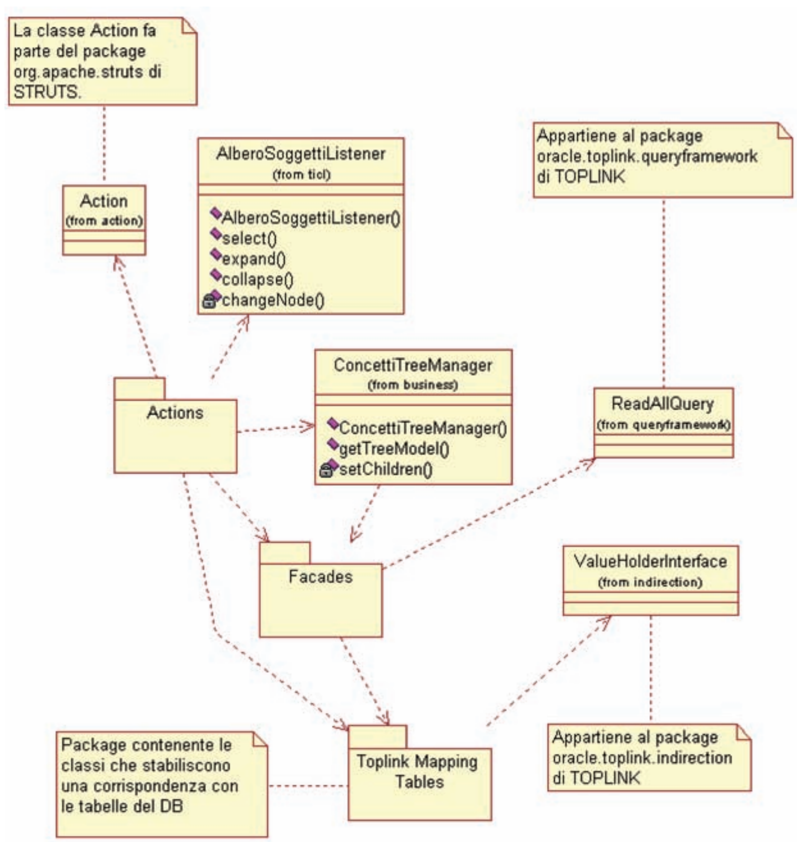
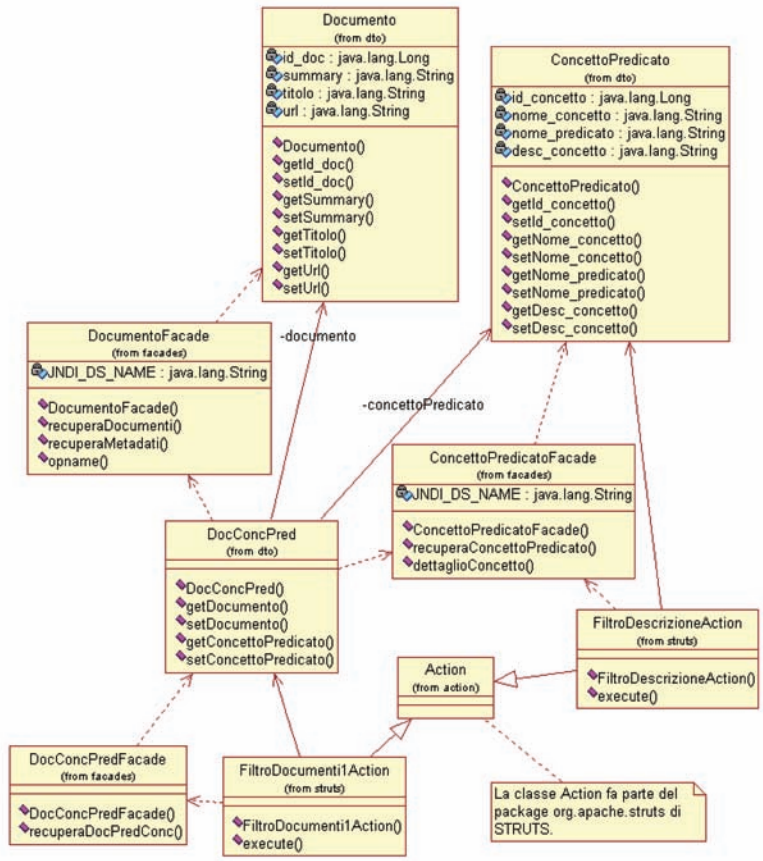


Figure 12. Class diagram for the semantic visualization of the learning modules



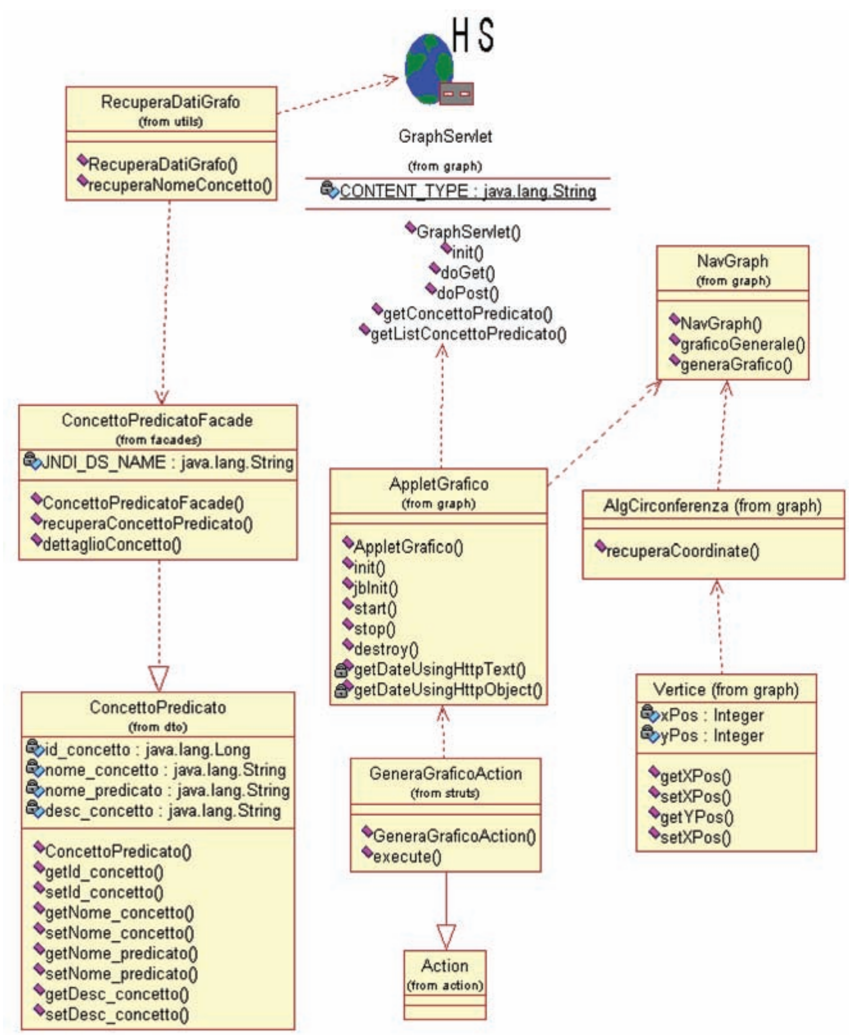
many times and under different conditions each functionality of the system. They also executed a general test for the overall system to evaluate its robustness and the coherence of data management and tracking systems. From the other side, the subject matter experts, after the design phase and the coordination of teams involved into the content creation process, executed a double-layer control: one for the exactness of how each topic was expressed, and one for the semantic link about different topics. Both tests revealed a set of enhancements that have been implemented in the new version of SWELS.

The "Beta test phase" was performed by involving a group of 20 students attending an International master program at the eBMS of Scuola Superiore ISUFI,. They used SWELS (platform and contents) as an additional learning tool during the attendance of the module on "Change Management and Leadership." After one week, at the end of this module, a face-to-face discussion meeting was organized with the participation of an outstanding professor in this field. Final impressions of master participants about SWELS were extremely positive, because they represented a sort of personal assistant to deepen and clarify some difficult concepts of the module and, above all, to have a systemic vision of the general topic.

Future Developments of Swels

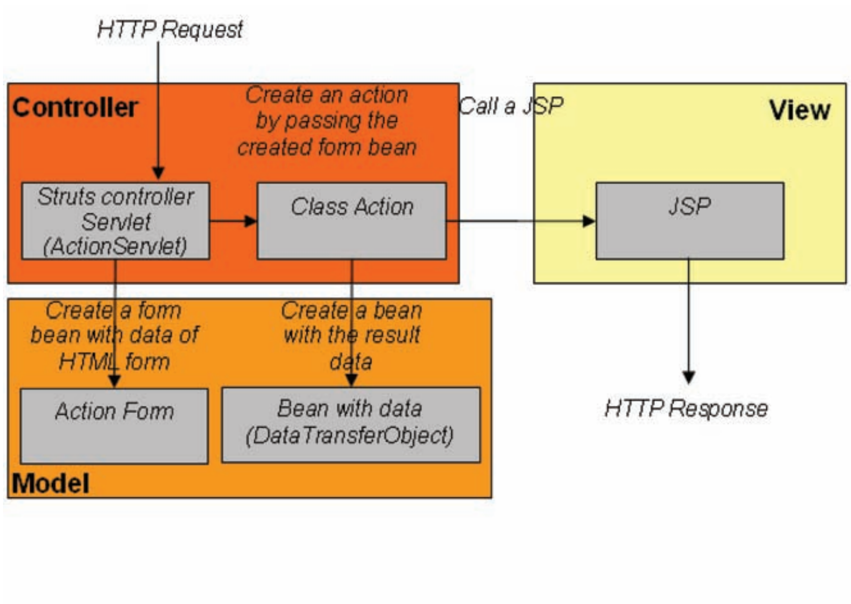
The next steps that we aim to develop in the future are:

Figure 13. Class diagram for the ontology structure visualization



- The implementation of an interactive radial layout layer (i.e., an interactive graphical interface to activate the "conceptual semantic boundary"). In our opinion, this improvement could make SWELS more effective, since that learners can immediately access learning modules, by directly clicking on the concepts shown in the radial graph.
- The ontological representation of two further learning dimensions: the typology and characteristics of the learning resource (i.e., assessment, difficulty level, etc.) and the learner profile expressed in terms of interests and knowledge gaps (by tracking the learning pattern dynamically created by learners).
- The integration of SWELS into a LMS, that means the development of a personal learning agent integrated into a LMS that proactively configures and recommends personalized learning paths to the learners according to their learning profile.

Figure 14. Integrating struts framework in MVC architectures



Finally, a large scale experimentation of the system should be organized in order to evaluate the effectiveness of SWELS and, more in general, of the learning approach embedded in SWELS.

CONCLUSION

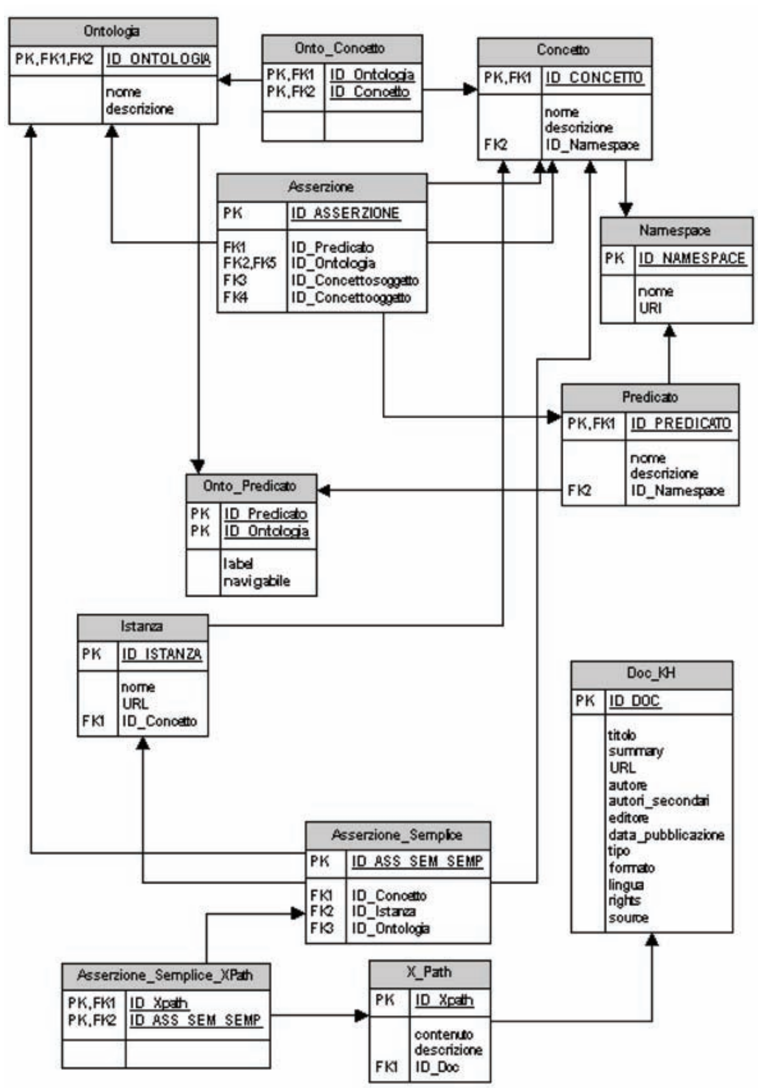
SWELS platform is the result of applying Semantic Web technologies to e-learning. Such a strategic choice allows learners and knowledge workers to increase the effectiveness of their learning process since it enables a personalized access to learning materials as well as a complete and deep understanding of the knowledge domain.Indeed, from the point of view of the final users, the main benefits of using SWELS are:

- The explicitation of tacit knowledge contained in the knowledge base conceptualization process and held in the minds of subject matter experts as well as domain designers;
- The systematization of knowledge through an explicit indexing of knowledge resources through simple and complex semantic assertions;
- A more direct access to the knowledge domain by explicitly navigating and browsing the ontology map;
- A more flexible structure of the learning materials that can be easily recombined and described for other purposes and learning goals in other knowledge domains.

In our beliefs, this approach could provide a new way in which students learn, since it is based on a learner-centric strategy characterized by:

- The role of personal tacit knowledge and learning experiences as the starting point and the knowledge background of future learning patterns;
- A solution-oriented approach for creating just-in-time new learning patterns;
- The possibility to fulfill the personal skill gap by actively participating and self-exploring the knowledge base;

Figure 15. Relational schema of the database implemented



- A stimulus to the “knowledge curiosity” of learners in deepening specific knowledge domains;
- The development of knowledge, skills and attitudes conceived as capacity for effective actions and problem solving;
- A set of customized training curricula consistent to learners’ needs, their own time and place, without compromising its effectiveness (Keegan, 2000);
- A dynamic creation of learning paths, starting from different knowledge resources semantically annotated, according to the learner interests and knowledge needs, expressed by them in real time.

REFERENCES

Adelsberger, H., Bick, M., Körner, F., & Pawlowski, J. M. (2001). Virtual education in business information systems (VAWI)--Facilitating collaborative development processes using the Essen learning model. In H. Hoyer (Ed.), *20th ICDE World Conference on Open Learning and Distance Education. The Future of Learning – Learning for the Future: Shaping the Transition*.

Barnes-Lee, T. (1998), Semantic Web Roadmap. *W3C Design Issues*.

Barnes-Lee, T. (2000). What the Semantic Web can represent.

Brennan, M., Funke, S., & Andersen, C. (2001). The learning content management system--a new e-Learning market segment emerges. *IDC White Paper*.,

Damiani, E., Corallo, A., Elia, G., & Ceravolo, P. (2002, November). *Standard per i learning objects: Interoperabilità ed integrazione nella didattica a distanza*. Paper presented at the International Workshop eLearning: una sfida per l'Università - Strategie Metodi Prospettive, Milan, Italy.

Drucker, P. (2000). Need to know: Integrating e-learning with high velocity value chains. *Delphi Group White Paper*.

Dublin Core Metadata Initiative. (2006). Dublin Core Metadata Terms., Retrieved on May 26, 2003, from http://dublincore.org

Elia, G., Secundo, G., & Taurino, C. (2006), Towards unstructured and just-in-time learning: The "Virtual eBMS" e-Learning system. In A Méndez-Vilas, A. Solano-Martin, J. Mesa González, & J.A. Mesa González (Eds.), *m-ICTE2006: Vol. 2. Current Developments in Technology-Assisted Education* (pp. 1067-1072). FORMATEX, Badajoz, Spain.

Fowler, M., & Scott, K. (Eds.). (1999). *UML Distilled Second Edition--A Brief Guide to the Standard Object Modelling Language*. A. Wesley.

Greenberg, L. (2002). LMS and LCMS: What's the Difference? *Learning Circuits, ASTD*.

Keegan, M. (Ed.). (2000). *e-Learning, The engine of the knowledge economy*. Morgan Keegan & Co.

Kolovski, V., & Galletly, J. (2003). Towards e-learning via the Semantic Web. In B. Rachev, & A. Smrikarov (Eds.), In *Proceedings of the 4th International Conference on Computer Systems and Technologies--CompSysTech'2003* (pp. 591 - 596). ACM New York, NY, USA.

Koper, R. (2004). Use of the Semantic Web to solve some basic problems in education. *Journal of Interactive Media in Education*, (6): 1–23.

Lockwood, F., & Gooley, A. (Eds.). (2001). *Innovation in open and distance learning - successful development of online and Web-based learning*. London, Kogan Page.

McCrea, F., Gay, R. K., & Bacon, R. (2000). *Riding the big waves: a white paper on the B2B e-Learning Industry*. San Francisco/Boston/New York/London: Thomas Weisel Partners LLC.

Mizoguchi, R. (2000). IT revolution in learning technology, In *Proceedings of SchoolNet 2000*, Pusan, Korea (pp. 46-55).

Naeve, A., Nilsson, M., & Palmer, M. (2001), E-learning in the semantic age. CID, Centre For User Oriented It Design. Stockhom, Sweden. Retrieved on September 30, 2006

Nejdl, W. (2001). Learning repositories--Technology and context. In A. Risk (Ed.), *ED-Media 2001 World Conference on education multimedia, Hypermedia and Telecommunications: Vol. 2001, N. 1*.

Rosenberg, M. J. (Ed.). (2006). *Beyond E-Learning: Approaches and technologies to enhance knowledge, learning and performance*. Pfeiffer.

Secundo, G., Corallo, A., Elia, G., & Passiante, G. (2004). An e-Learning system based on semantic Web supporting a Learning in Doing Environment. In *Proceedings of International Conference on Information Technology Based Higher Education and Training--ITHET 2004.* IEEE XPlore,

Stojanovic, L., Staab, S., & Studer, R. (2001,October 23-27). eLearning based on the Semantic Web. In W. A. Lawrence-Fowler & J. Hasebrook (Eds.), In *Proceedings of WebNet 2001--World Conference on the WWW and Internet*, Orlando, Florida. (pp.1174-1183). AACE.

Sun Microsystems, Inc. (2002). Java Blueprints--Model-view-controller.

The Apache Software Foundation. (2006). Apache Struts.

This work was previously published in Semantic Knowledge Management: An Ontology-Based Framework, edited by A. Zilli; E. Damiani; P. Ceravolo; A. Corallo and G. Elia, pp. 120-145, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 3.10
Web Services Discovery with Rough Sets

Maozhen Li
Brunel University, UK

Bin Yu
Level E Limited, UK

Vijay Sahota
Brunel University, UK

Man Qi
Canterbury Christ Church University, UK

ABSTRACT

Web services are emerging as a major technology for building service-oriented distributed systems. Potentially, various resources on the Internet can be virtualized as Web services for a wider use by their communities. Service discovery becomes an issue of vital importance for Web services applications. This article presents ROSSE, a Rough Sets based Search Engine for Web service discovery. One salient feature of ROSSE lies in its capability to deal with uncertainty of service properties when matching services. A use case is presented to demonstrate the use of ROSSE for discovery of car services. ROSSE is evaluated in terms of its accuracy and efficiency in service discovery.

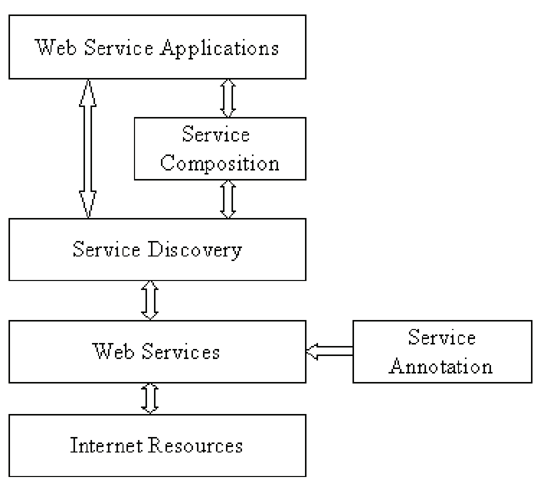
INTRODUCTION

Web services are emerging as a major technology for developing service-oriented distributed systems. Potentially, many resources on the Internet or the World Wide Web can be virtualized as services for a wider use by their communities. Service discovery becomes an issue of vital importance for Web service applications. As shown in Figure 1, discovered services can either be used by Web service applications or they can be composed into composite services using workflow languages such as BPEL4WS (Andrews Curbera, Dholakia, Goland, Klein, Leymann et al., 2003). UDDI (Universal Description, Discovery and Integration, http://www.uddi.org) has been proposed and used for Web service publication and discovery. However, the search mechanism supported by UDDI is limited to keyword matches. With the development of the Semantic Web (Berners-Lee, Hendlet, & Lassila, 2001), services can be annotated with metadata for enhancement of service discovery. The complexity of this metadata can range from simple annotations, to the representation of more complex relationships between services based on first order logic.

One key technology to facilitate this semantic annotation of services is OWL-S (Martin, Paolucci, McIlraith, Burstein, McDermott, McGuinness et al., 2004), an OWL (Web Ontology Language, http://www.w3.org/TR/owl-features/Reference) based ontology for encoding properties of Web services. OWL-S ontology defines a service profile for encoding a service description, a service model for specifying the behavior of a service, and a service grounding for invoking the service. Typically, a service discovery process involves a matching between the profile of a service advertisement and the profile of a service request using domain ontologies described in OWL. The service profile not only describes the functional properties of a service such as its inputs, outputs, pre-conditions, and effects (IOPEs), but also non-functional features including service name, service category, and aspects related to the quality of a service. In addition to OWL-S, another prominent effort on Semantic Web services is WSMO (Roman, Keller, Lausen, Bruijn, Lara, Stollberg et al., 2005), which is built on four key concepts—ontologies, standard WSDL based Web services, goals, and mediators. WSMO stresses the role of a mediator in order to support interoperation between Web services.

However, one challenging work in service discovery is that service matchmaking should be able to deal with uncertainty in service properties when matching service advertisements with service requests. This is because in a large-scale heterogeneous system, service publishers and requestors may use their pre-defined properties to describe services, for example, in the form of OWL-S or WSMO. For a property explicitly used in one service advertisement, it may not be explicitly used by another service advertisement within the same service category. As can be seen from Table 1, the property P_1 used by the service advertisement S_1 does not appear in the service advertisement S_2. When services S_1 and S_2 are matched with a query using properties P_1, P_2 and P_3, the property P_1 becomes an uncertain property when matching S_2. Similarly, the property P_3 becomes an uncertain property when matching

Figure 1. A layered structure for service-oriented systems



S_1. Consequently, both S_1 and S_2 may not be discovered because of the existence of uncertainty of properties even though the two services are relevant to the query.

It is worth noting that properties used in service advertisements may have dependencies, for example, both P_1 and P_3 may be dependent properties of P_2 when describing services S_1 and S_2 respectively. Both S_1 and S_2 can be discovered if P_1 and P_3 (which are uncertain properties in terms of the user query) can be dynamically identified and reduced in the matching process. To increase the accuracy of service discovery, a search engine should be able to deal with uncertainty of properties when matching services.

In this article, we present ROSSE, a Rough Sets (Pawlak, 1982) based Search Engine for Web service discovery. One salient feature of ROSSE lies in its capability to deal with uncertainty in service properties (attributes) when matching service advertisements with service requests. Experiment results show that ROSSE is more effective in service discovery than existing mechanisms such as UDDI keyword matching and OWL-S matchmaking.

The remainder of this article is organized as follows. The ROSSE Design section presents the design details of ROSSE. The ROSSE Case Study section gives a case study to demonstrate the use of ROSSE for discovery of car services. The ROSSE Implementation and Evaluation section evaluates ROSSE from the aspects of accuracy and efficiency in service discovery. The Related Word section discusses some related work, and the Conclusion and Future Work section concludes the article.

ROSSE DESIGN

ROSSE considers input and output properties individually when matching services. For the simplicity of expression, input and output properties used in a service request are generally referred to as service request properties. The same goes to service advertisements.

Figure 2 shows ROSSE components. The Irrelevant Property Reduction component takes a service request as an input (step 1), and then it accesses a set of advertised domain services (step 2) to remove irrelevant service properties using the domain ontology (step 3). Reduced properties will be marked in the set of advertised domain services (step 4). Once invoked (step 5), the Dependent Property Reduction component accesses the advertised domain services (step 6) to discover and reduce indecisive properties which will be marked in advertised domain services (step 7). Invoked by the Dependent Property Reduction component (step 8), the Service Matching and Ranking component accesses the advertised domain services for service matching and ranking (step 9), and finally it produces a list of matched services (step 10).

In the following sections, we describe in depth the design of ROSSE components for service matchmaking and discovery. Firstly, we introduce Rough sets for service discovery.

Rough Sets for Service Discovery

Rough sets method is a mathematic tool that can deal with uncertainty in knowledge discovery. It is based on the concept of an upper and a lower

Table 1. Two service advertisements with uncertain service properties

service advertisements	property	property	property
S_1	P_1	P_2	
S_2		P_2	P_3

approximation of a set as shown in Figure 3. For a given set X, the yellow grids (lighter shading) represent its upper approximation, and the green grids (darker shading) represent its lower approximation. We introduce Rough sets for service discovery in the following way.

Let

- Ω be a domain ontology.
- U be a set of N service advertisements, $U = \{s_1, s_2, ..., s_N\}$, $N \geq 1$.
- P be a set of K properties used in the N service advertisements, $P = \{p_1, p_2, ..., p_K\}$, $K \geq 2$.
- P_A be a set of M properties used in service advertisements which are relevant to a service request R within the domain ontology Ω,
- $P_A = \{p_{A1}, p_{A2}, ..., p_{AM}\}$, $P_A \subseteq P$, $M \geq 1$.
- X be a set of service advertisements relevant to the service request R, $X \subseteq U$.
- $\underline{X}$ be the lower approximation of the set X.
- $\overline{X}$ be the upper approximation of the set X.

According to the Rough sets theory, we have

$$\underline{X} = \{x \in U : [x]_{P_A} \subseteq X\} \qquad (1)$$

$$\overline{X} = \{x \in U : [x]_{P_A} \cap X \neq \varnothing\} \qquad (2)$$

For a property used by a service request $p \in P_A$, we have

- $\forall x \in \underline{X}$, x definitely has property p.
- $\forall x \in \overline{X}$, x possibly has property p.
- $\forall x \in U - \overline{X}$, x absolutely does not have property p.

The use of "definitely," "possibly" and "absolutely" are used to encode properties that cannot be specified in a more exact way. This is a significant addition to existing work, where discovery of services needs to be encoded in a precise way, making it difficult to find services which have an approximate match to a query.

Advertised domain service properties may be irrelevant (having no effect on service matching) or relevant (having an impact on service matching). Certain properties used by advertised services may be redundant which can be reduced without losing essential classificatory information. The concept of the reduct is fundamental for Rough sets theory (Winiarski, 2001). Service property reduction can be considered as a process of finding a smaller (than the original one) set of properties with the same or close classificatory power as

Figure 2. ROSSE components

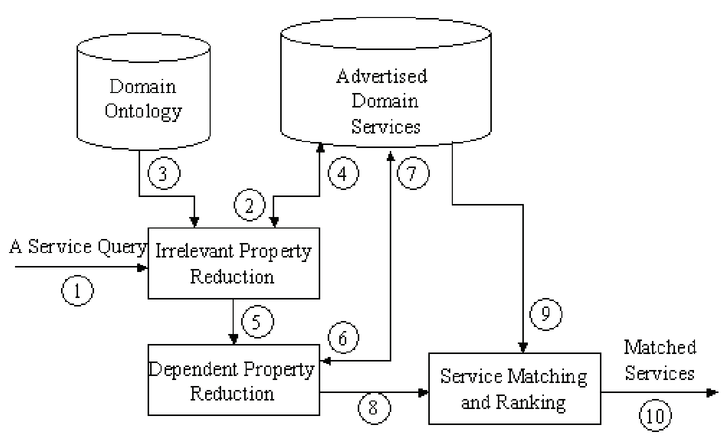
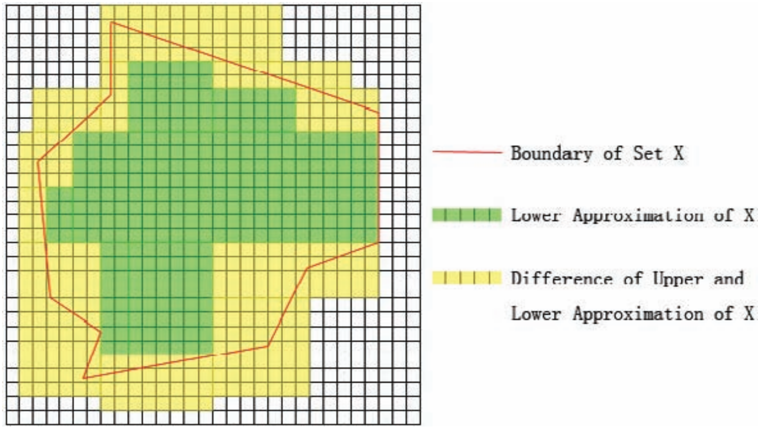


Figure 3. Approximation in Rough sets



the original set. For a service query, the most relevant properties of advertised services can be determined after property reduction.

Reducing Irrelevant Properties

When searching for a service, a service request may employ some properties which are irrelevant to the properties used in a service advertisement within one domain ontology. These irrelevant properties used in service advertisements should be removed before the service matchmaking process is performed.

Let

- p_R be a property used in a service request.
- p_A be a property used in a service advertisement.

Following the work proposed in (Paolucci, Kawamura, Payne, & Sycara, 2002), we define the following relationships between p_R and p_A:

- **exact** match, p_R and p_A are equivalent or p_R is a subclass of p_A.
- **plug-in** match, p_A subsumes p_R.
- **subsume** match, p_R subsumes p_A.
- **nomatch**, no subsumption between p_R and p_A.

For each property used in a service request, the Irrelevant Property Reduction component uses Algorithm 1 to remove irrelevant properties from advertised services. For those properties used in service advertisements that have a nomatch result, they will be treated as irrelevant properties. Service advertisements are organised as service records in a database. Properties are organised in such a way that each property uses one column to ensure the correctness in the following reduction of dependent properties. As a property used in one service advertisement might not be used in another one, some properties may have empty values. For a service request, a property with an empty value in a service record becomes an uncertain property. If a property in an advertised service record is marked as nomatch, the column associated with the property will be marked as nomatch. As a result, all properties within the column including uncertain properties (i.e., properties with empty values) will not be considered in service matchmaking.

Reducing Dependent Properties

Properties used by service advertisements may have dependencies. Dependent properties are indecisive properties which have no effect on

Algorithm 1. Reducing irrelevant properties from service advertisements

```
1: for each property  pA used in service advertisements
2:   for all properties used in a service request
3:       if  pA is nomatch with any  pR
4:          then pA is marked with nomatch;
5:       end if
6:    end for
7: end for
```

service matching. Building on the work proposed in (Jensen, Shen, & Tuson, 2005), we designed Algorithm 2 to reduce dependent properties from advertised services.

Let

- Ω, U, P, P_A be defined as in the Rough Sets for Service Discovery section.
- P_A^D be a set of L_D decisive properties for identifying service advertisements relevant to the service request R in terms of Ω,
- $P_A^D = \{p_{A1}^D, p_{A2}^D, \ldots, p_{AL_D}^D\}$, $P_A^D \subseteq P_A$, $L_D \geq 1$.
- P_A^{IND} be a set of L_{IND} indecisive properties for identifying service advertisements relevant to the service request R in terms of Ω,
- $P_A^{IND} = \{p_{A1}^{IND}, p_{A2}^{IND}, \ldots, p_{AL_{IND}}^{IND}\}$, $P_A^{IND} \subseteq P_A$, $L_{IND} \geq 1$.
- $IND()$ be an indiscernibility relation.
- f be a mapping function from a property to a service advertisement.

Then

$$IND\,(P_A^{IND}) = \{(x, y) \in U : \forall p_{Ai}^{IND} \in P_A^{IND},\ f(x, p_{Ai}^{IND}) = f(y, p_{Ai}^{IND})\} \quad (3)$$

$$P_A^D = P_A^{IND} - P_A \quad (4)$$

For a service request, the Dependent Property Reduction component uses Algorithm 2 to find the decisive properties in service advertisements. Specifically, service advertisements with the maximum number of nonempty property values are used in the algorithm as targets to find indecisive properties. The targeted services can still be uniquely identified without using these indecisive properties. All possible combinations of individual indecisive properties are checked with an aim to maximally remove indecisive properties which may include uncertain properties whose values are empty. In the mean time, the following service discovery process is speeded up due to the reduction of dependent properties.

Computing Match Degrees

The Service Matching and Ranking component uses the decisive properties to compute the match degrees of advertised services related to a service request.

Let

- Ω, U, P, P_A be defined as in the Rough Sets for Service Discovery section.
- P_R be a set of M properties used in a service request R. $P_R = \{P_{R1}, P_{R2}, \ldots, P_{R3}\}$, $M \geq 1$.
- P_A^D be a set of L_D decisive properties for identifying service advertisements relevant to the service request R in terms of Ω,
- $P_A^D = \{p_{A1}^D, p_{A2}^D, \ldots, p_{AL_D}^D\}$, $L_D \geq 1$.
- $m(p_{Ri}, p_{Aj})$ be a match degree between a property P_{Ri} and a property P_{Aj} in terms of Ω, $P_{Ri} \in P_R$, $1 \leq i \leq M$, $P_{Aj} \in P_A^D$, $1 \leq j \leq L_D$.
- $v(P_{Aj})$ be a value of the property P_{Aj}, $P_{Aj} \in P_A^D$, $1 \leq j \leq L_D$.

Algorithm 2. Reducing dependent properties from advertised services

```
S is a set of service advertisements with the maximum number of nonempty
      property values relevant to a service request;
   PA is a set of properties used by the S set of service advertisements;
   PAD is a set of decisive properties, PAD ⊆ PA;
   PAIND is a set of individual indecisive properties, PAIND ⊆ PA ;
   PAIND_Core is a set of combined indecisive properties,
     PAIND_Core ⊆ PAIND;
   PAD = Ø; PAIND = Ø; PAIND_Core = Ø;
for each property p∈ PA
   if p is an indecisive property for identifying the S set of services
     then
    add p into PAIND;
    PAIND_Core = Ø;
    add p into PAIND_Core;
   end if
end for
for i=2 to sizeof(PAIND)-1
   calculate all possible i combinations of the properties in PAIND;
   if any combined i properties are indecisive properties for identifying
          the S set of services
     then
   PAIND_Core = Ø;
   add the i properties into PAIND_Core;
   continue;
     else if any combined i properties are decisive properties
       then break;
   end if
end for
PAD = PA-PAIND_Core;
return PAD;
```

- $S(R, s)$ be a similarity degree between a service advertisement s and the service request R, $s \in U$.

Algorithm 3 shows the rules for calculating a match degree between a property used in a service request and a property used in a service advertisement. A decisive property with an empty value has a match degree of 50% when matching each property used in a service request. A property used in a service advertisement will be given a match degree of 100% if it has an exact match relationship with a property used in a service request. A match degree of 50% will be given if it has a plug-in relationship with a service request property and the relationship is out of five generations. Similarly, a property used in a service advertisement will be given a match degree of 50% if it has a subsume relationship with a service request property and the relationship is out of three generations.

Algorithm 3. The rules for calculating match degrees between properties used in service requests and service advertisements respectively

for each property $p_{Aj} \in P_A^{\ D}$, $v(p_{Aj}) \neq NULL$
for each property $p_{Ri} \in P_R$
if p_{Aj} is an exact match with p_{Ri}
then $m(p_{Ri}, p_{Aj}) = 1$;
else if p_{Aj} is a plug-in match with p_{Ri}
then if p_{Ri} is the kth subclass of p_{Aj} and $2 \leq k \leq 5$
then $m(p_{Ri}, p_{Aj}) = 1-(k-1)\times 10\%$;
else if p_{Ri} is the kth subclass of p_{Aj} and $k>5$
then $m(p_{Ri}, p_{Aj}) = 0.5$;
end if
else if p_{Aj} is a subsume match with p_{Ri}
then if p_{Aj} is the kth subclass of p_{Ri} and $1 \leq k \leq 3$
then $m(p_{Ri}, p_{Aj}) = 0.8-(k-1)\times 10\%$;
else if p_{Aj} is the kth subclass of p_{Ri} and $k>3$
then $m(p_{Ri}, p_{Aj}) = 0.5$;
end if
end if
end for
end for
for each property $p_{Aj} \in P_A^{\ D}$, $v(p_{Aj}) = NULL$
for each property $p_{Ri} \in P_R$
$m(p_{Ri}, p_{Aj}) = 0.5$;
end for
end for

Each decisive property used for identifying service advertisements has a maximum match degree when matching all the properties used in a service request. $S(R, s)$ can be calculated using formula (5).

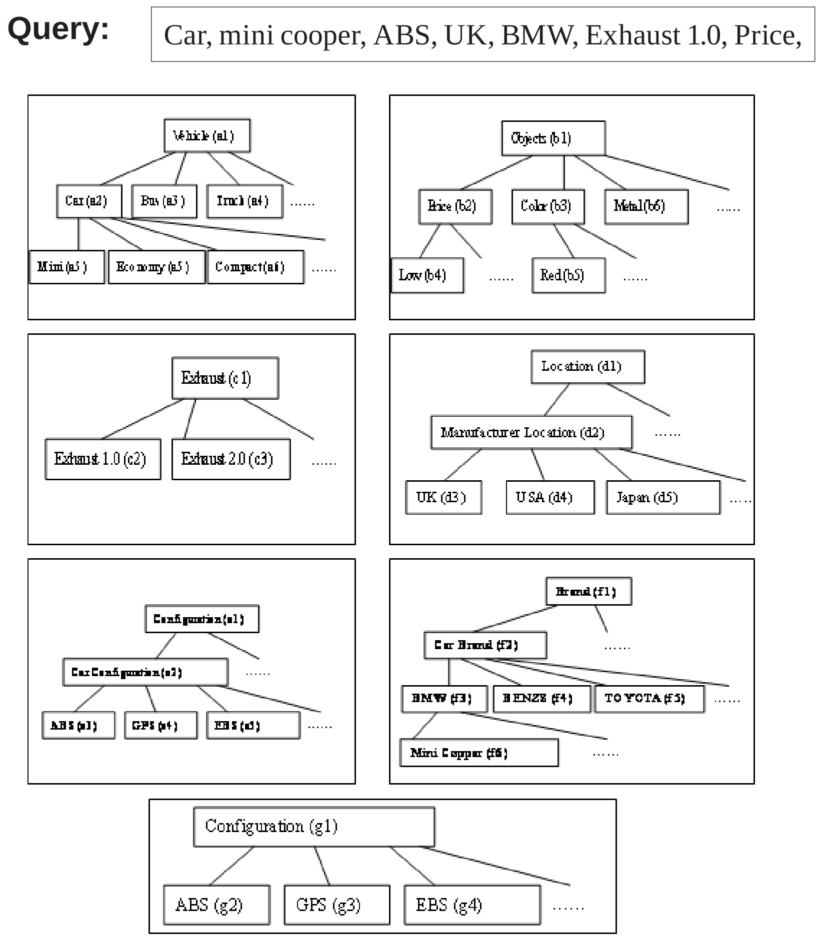
$$S(R,s) = \sum_{j=1}^{L_D} \sum_{i=1}^{M} \max(m(p_{Ri}, p_{Aj})) \Big/ L_D \qquad (5)$$

Using the formula (5), ROSSE calculates a matching degree for each service advertisement related to a service request. The similarity degrees are used to produce a lower and an upper approximation set of discovered services.

ROSSE CASE STUDY

In this section, we present a use case of ROSSE to discover vehicle services. Figure 4 shows the

Figure 4. Ontolgogies used in the search scenario



ontologies used in this scenario defining the classifications of *vehicles, objects, exhausts, locations, configurations, brands* respectively. Two ontologies are used to classify configurations of vehicles represented respectively by *e1-e5* and *g1-g4*. Relevant vehicle services are registered with ROSSE. In the following sections, we describe how services are matched in terms of the following query to search for car services that sell red BMW mini coopers that have an exhaust of 1.0, and are configured with ABS, manufactured in the UK. Price information is also provided by the car services.

Building a Decision Table

A service decision table is used to compute dependent properties among services. As the number of services registered with ROSSE can be tremendous, the decision table is constructed by sampling registered services. For a specific query, ROSSE randomly selects a certain number of services records. A service record is selected as long as one of its properties has a valid relationship with a property used in a service query. The relationship can be *exact*, *plug-in* or *subsume* as defined in algorithm 1 which is described in the Reducing Irrelevant Properties section.

Table 2. A segment of the decision table used for discovery of car services

properties / services	f6	g2	d3	f3	c2	b2	b3	d2	c1	e1/g1	d1	b6
S_1	1	1	1	1	1	0	0	0	0	0	1	1
S_2	0	1	0	1	0	0	0	0	0	0	1	1
S_3	0	1	0	1	0	0	1	1	1	1	0	0
S_4	0	1	0	0	0	1	1	1	0	1	0	0
S_5	0	1	1	0	1	0	0	0	0	1	0	0
S_6	1	1	1	1	1	1	0	0	0	0	0	0
S_7	0	1	0	0	0	0	1	1	1	0	0	0
S_8	1	1	1	1	1	0	0	0	0	1	0	0
S_9	0	1	0	1	0	0	0	1	0	0	1	1
S_{10}	0	1	0	0	0	0	0	0	1	0	1	0
S_{11}	0	1	0	0	1	0	0	1	0	0	0	0
S_{12}	0	1	0	1	0	0	0	0	1	0	1	1
S_{13}	1	1	1	1	1	1	1	0	0	1	0	0

Table 3. Computed dependent properties

properties / services	f6	g2	d3	f3	c2	b2	b3	d2	c1	e1/g1	d1	b6
S_1	1	1	1	1	1	0	0	0	0	0	1	1
S_2	0	1	0	1	0	0	0	0	0	0	1	1
S_3	0	1	0	1	0	0	1	1	1	1	0	0
S_4	0	1	0	0	0	1	1	1	0	1	0	0
S_5	0	1	1	0	1	0	0	0	0	1	0	0
S_6	1	1	1	1	1	1	0	0	0	0	0	0
S_7	0	1	0	0	0	0	1	1	1	0	0	0
S_8	1	1	1	1	1	0	0	0	0	1	0	0
S_9	0	1	0	1	0	0	0	1	0	0	1	1
S_{10}	0	1	0	0	0	0	0	0	1	0	1	0
S_{11}	0	1	0	0	1	0	0	1	0	0	0	0
S_{12}	0	1	0	1	0	0	0	0	1	0	1	1
S_{13}	1	1	1	1	1	1	1	0	0	1	0	0

Table 2 shows a segment of the decision table with 13 service records for discovery of car services. As can be seen from Table 2, properties of advertised services that are relevant to the car service query are *f6*, *g2*, *d3*, *f3*, *c2*, *b2*, *b3*, *d2*, *c1*, *e1/g1*, *d1*, *b6*. If a property in a service record is marked with *1*, this means that the property is used by the service in its advertisement. For example, the service S_1 has properties of *f6*, *g2*, *d3*, *f3*, *c2*, *d1*, and *b6* in its advertisement. A property marked with *0* in a service record means that the service does not have the corresponding property in its advertisement, for example, properties such as *b2*, *b2*, *d2*, *c1*, and *e1/g1* are not used by the service S_1 for advertisement. However, it should be noted that a property marked with 0 in a service record does not necessarily mean this property is not relevant to the service. Such a property might be an inherent property of the service. ROSSE deals with properties marked with 0 as uncertain properties when matching services.

Computing Dependent Properties

Once a service decision table is constructed, the next step is to compute dependent properties. Using the algorithm 2 presented in the Reducing Dependent Properties section, properties *g2*, *d3*, *f3*, and *c2* are indecisive properties which are reduced from the decision table in matching services as shown in Table 3. Table 4 shows the segment of the decision table without dependent properties.

Table 4. The segment of the decision table without dependent properties

properties / services	f6	b2	b3	d2	c1	e1/g1	d1	b6
S_1	1	0	0	0	0	0	1	1
S_2	0	0	0	0	0	0	1	1
S_3	0	0	1	1	1	1	0	0
S_4	0	1	1	1	0	1	0	0
S_5	0	0	0	0	0	1	0	0
S_6	1	1	0	0	0	0	0	0
S_7	0	0	1	1	1	0	0	0
S_8	1	0	0	0	0	1	0	0
S_9	0	0	0	1	0	0	1	1
S_{10}	0	0	0	0	1	0	1	0
S_{11}	0	0	0	1	0	0	0	0
S_{12}	0	0	0	0	1	0	1	1
S_{13}	1	1	1	0	0	1	0	0

Table 5. Computation of matching degrees

Match Degrees	100%					95%	90%	
properties / services	f6	b2	b3	d2	c1	e1/g1	d1	b6
S_1	1	0	0	0	0	0	1	1
S_2	0	0	0	0	0	0	1	1
S_3	0	0	1	1	1	1	0	0
S_4	0	1	1	1	0	1	0	0
S_5	0	0	0	0	0	1	0	0
S_6	1	1	0	0	0	0	0	0
S_7	0	0	1	1	1	0	0	0
S_8	1	0	0	0	0	1	0	0
S_9	0	0	0	1	0	0	1	1
S_{10}	0	0	0	0	1	0	1	0
S_{11}	0	0	0	1	0	0	0	0
S_{12}	0	0	0	0	1	0	1	1
S_{13}	1	1	1	0	0	1	0	0

Computing Match Degrees

Decisive properties are used for computing the similarities between an advertised service and a service request. For each decisive property used in a service advertisement and a property used in the service query, a maximum matching degree can be computed using ontologies defined in Figure 4. Table 5 shows the matching degrees of the decisive properties used in the exemplified 13 service records. It should be noted that both *e1* and *g1* refers to the same property *Configuration*, but they use different ontology definitions as shown in Figure 4. The matching degree of *Configuration* to the *ABS* property used in the query is computed in such way that a mean of two matching degrees using the two ontology definitions (i.e., 100% and 90%) is computed which is 95%.

It is worth noting that for an uncertain property which is marked with the number of 0 in a box of Table, a matching degree of 50% is given. Based on the formula (5) presented in the Comupting Match Degrees section, the similarity degree between an advertised service and a service query can be computed. In the car service query case, for example, service S_1 has a similarity degree of 66.25% and service S_{13} has a similarity degree of 74.375%.

ROSSE IMPLEMENTATION AND EVALUATION

ROSSE is implemented with Java on a Pentium IIII 2.6G with 512M RAM running Red Hat Fedora Linux 3. Figure 5 shows the homepage of ROSSE. It has two registries for service registration, a UDDI registry and an OWL registry. The UDDI registry is used to register services with WSDL interfaces, and the OWL-S registry is used to register services with OWL-S interfaces. The UUID of a WSDL service registered with the UDDI registry is used to uniquely identify semantic annotation records of the registered service. In this way, WSDL services registered with ROSSE can be matched with semantic inferences instead of using keywords only. jUDDI (http://ws.apache.org/juddi) and mySQL (http://www.mysql.com) are used to build the UDDI registry and UDDI4J (http://uddi4j.sourceforge.net/) is used to query the registry. OWL-S API (http://www.mindswap.org/2004/owl-s/api) is used to parse OWL-S documents to register services with OWL-S interfaces with the OWL-S registry in ROSSE.

ROSSE provides graphical user interfaces to register services. Figure 6 shows a page to register a *vehicle* service that has a WSDL Interface, and Figure 7 shows the four steps used to semantically

Figure 5. ROSSE user interface

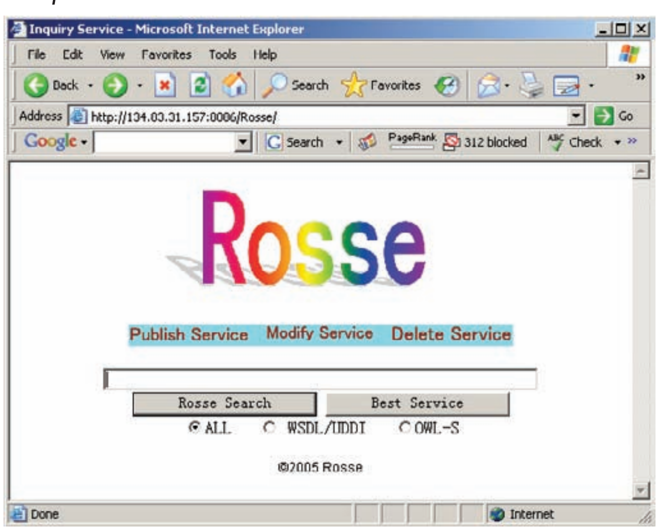


Figure 6. Registering a service that has a WSDL interface

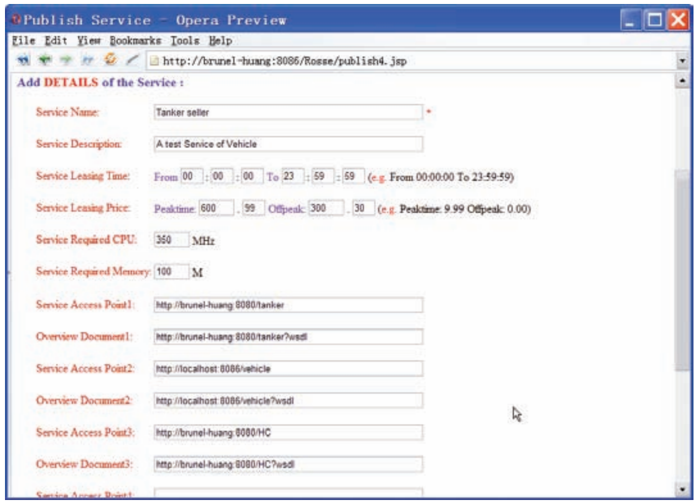
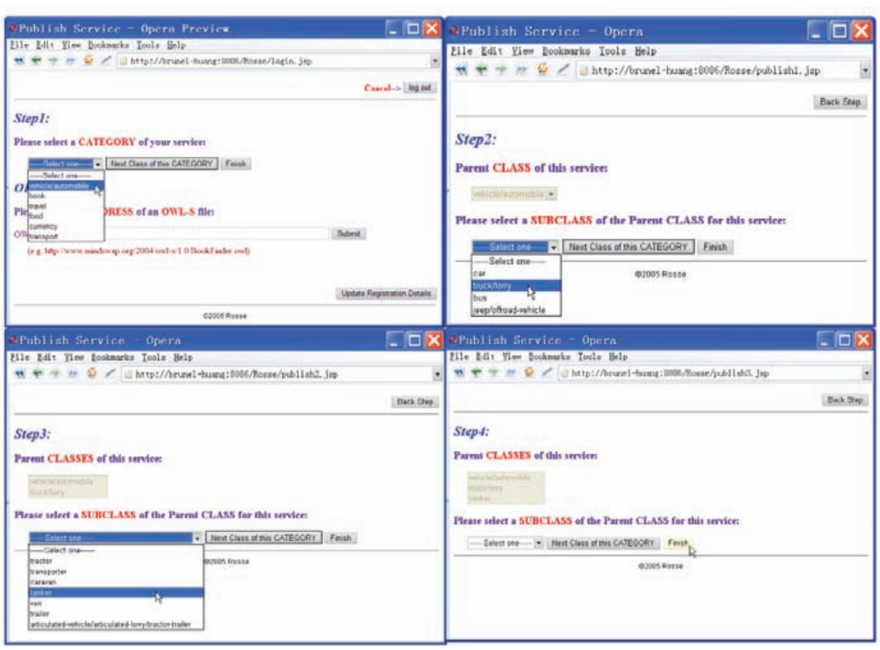


Figure 7. Annotating a vehicle service with semantic information



annotate the *vehicle* service. Figure 8 shows the registration of a zip code finding service with an OWL-S interface in ROSSE.

For a service request, ROSSE computes a matching degree for each service advertisement in terms of its functional input and output properties using formula (5). As shown in Figure 5, ROSSE can discover services with WSDL interfaces or OWL-S interfaces. It can also discover the best service from service advertisements which has

Figure 8. Registering OWL-S services with ROSSE

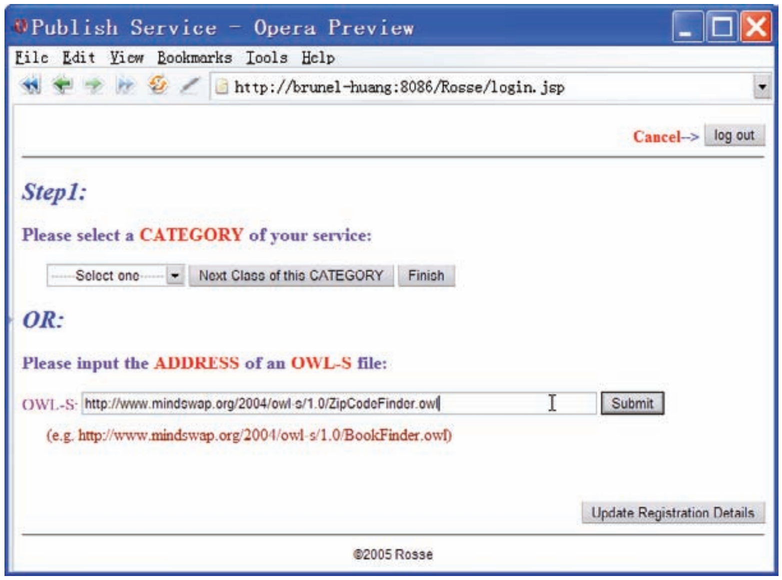
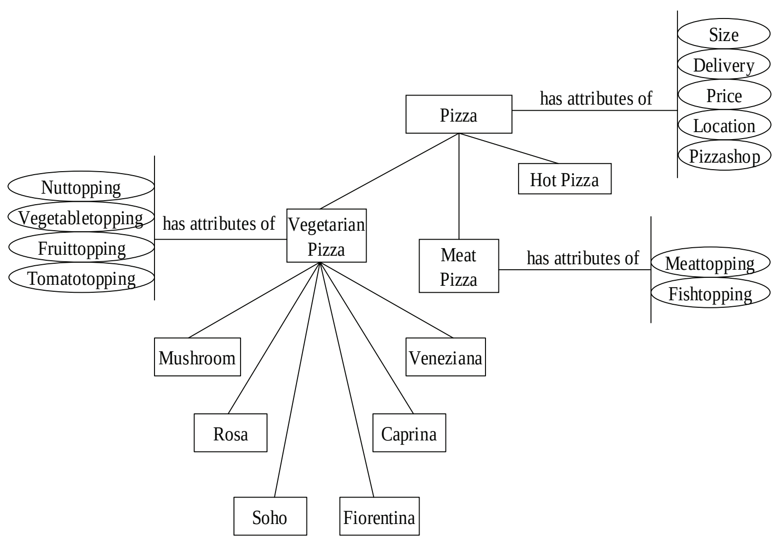


Figure 9. Pizza ontology structure



the highest matching degree related to a service request.

In this section, we evaluate the accuracy and efficiency of ROSSE in service discovery. We compare ROSSE with UDDI and OWL-S respectively. RACER (Haarslev & Möller, 2001) was used by OWL-S to infer the relationships between properties used in service queries and service advertisements. We implemented a light weighted reasoning component in ROSSE to overcome a high overhead incurred by RACER. The component uses the Protégé OWL API (http://protege.stanford.edu/plugins/owl/api/) to parse OWL documents.

We designed Pizza services for the tests using the Pizza ontology defined by http://www.co-ode.org/ontologies/pizza/pizza_20041007.owl. Figure 9 shows the Pizza ontology structure. The approach adopted here can be applied to other domains—where a specific ontology can be specified. The use of service properties needs to be related to a particular application-specific ontology.

ROSSE Accuracy in Service Discovery

Precision and recall are standard measures that have been used in information retrieval for measuring the accuracy of a search method or a search engine (Rijsbergen, 1979). We performed 4 groups of tests to evaluate the precision and recall of ROSSE in service discovery using 10 service records in each group. Each service had 5 properties of which 2 properties were dependent properties. For a service query, each group had 3 relevant services. The 10 services in group 1 did not have uncertain properties, but group 2 had 3 services with uncertain properties, group 3 had 5 services with uncertain properties and group 4 had 7 services with uncertain properties. Properties such as *Size*, *Price*, *Nuttoping*, *Vegetariantopping*, and *Fruittopping* were used by the advertised services. Table 6 shows the evaluation results.

In the tests conducted for group 1, both OWL-S and ROSSE have a precision of 100%. This is because all service advertisements in this group do not have uncertain properties (i.e., properties with empty values). UDDI discovered 4 services, but only 2 services were relevant to the service query with a precision of 50%, and a recall of 66.7%. In the tests of the last 3 groups where advertised services have uncertain properties, OWL-S cannot discover any services producing a precision of 0 and a recall of 0. Although UDDI can still discover some services in these tests, the precision of each group is low. For example, in the tests of group 3 and group 4 where the service property certainty rates are 50% and 30% respectively, UDDI cannot discover any relevant services. ROSSE is more effective than both UDDI and OWL-S in dealing with uncertain properties when matching services. For example, ROSSE is still able to produce a precision of 100% in the tests of the last 3 groups albeit with a low recall which is 33.3%.

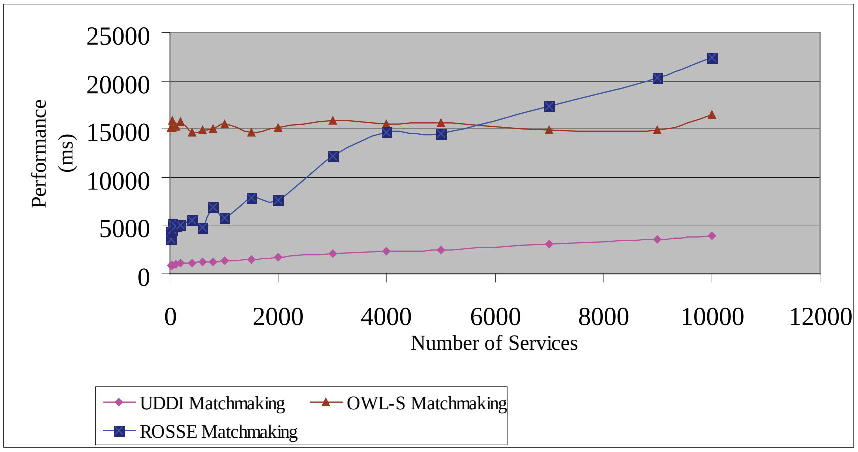
ROSSE Efficiency in Service Discovery

We have registered 10,000 Pizza service records with ROSSE for testing its efficiency in service discovery. Service discovery involves two processes, one is service matchmaking and the other is service accessing (i.e., accessing matched services). We compared the efficiency of ROSSE in matching services with that of UDDI and OWL-S

Table 6. ROSSE accuracy in service discovery

Service Property Certainty Rate	UDDI		OWL-S		ROSSE	
	Precision	Recall	Precision	Recall	Precision	Recall
100%	50%	66.7%	100%	100%	100%	100%
70%	33.3%	33.3%	0	0	100%	33.3%
50%	0	0	0	0	100%	33.3%
30%	0	0	0	0	100%	33.3%

Figure 10. ROSSE efficiency in service matchmaking



respectively, and the evaluation results are plotted in Figure 10. We also compared their efficiency in accessing matched services, and the results are plotted in Figure 11.

From Figure 10 we can see that UDDI has the least overhead in matching services. This is because UDDI only supports keyword based exact matching. UDDI does not support the inference of the relationships between requested service properties and advertised service properties which is a time consuming process. We also observe that ROSSE has a better performance in service matchmaking than OWL-S when the number of advertised services is less than 5500. This is because ROSSE used a simpler reasoning component than RACER which was used by OWL-S for matching services. However, the overhead of ROSSE in service matchmaking increases when the number of services gets larger. This is due to the overhead caused by a reduction of dependent properties. The major overhead of OWL-S in matching services is caused by RACER which is sensitive to the number of service properties instead of the number of services.

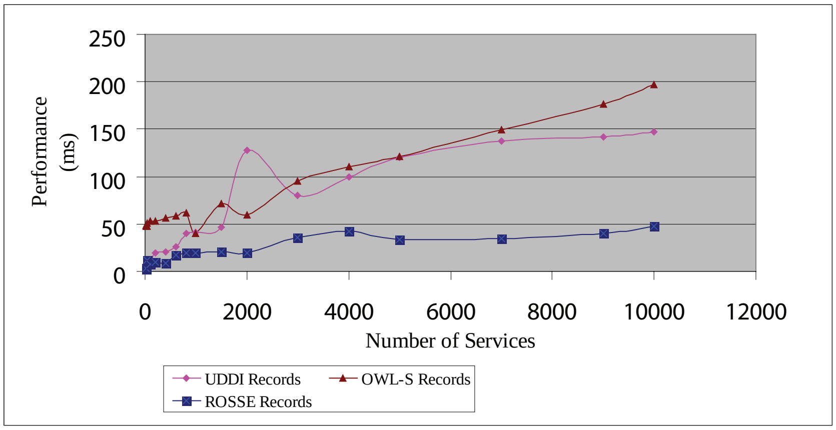
From Figure 11 we can see that the ROSSE matchmaking algorithm is most efficient in accessing matched services due to its reduction of dependent properties. The OWL-S has a similar performance to UDDI in this process.

RELATED WORK

Service matchmaking is becoming an issue of vital importance in service-oriented systems. UDDI has been proposed to support service publication and discovery. However, the search mechanism supported by UDDI is limited to keyword matches and does not support any inference based on the taxonomies referred to by the tModels. Various extensions (Miles, Papay, Dialani, Luck, Decker, Payne et al., 2003; Powles & Krishnaswamy, 2005; ShaikhAli, Rana, Al-Ali, & Walker, 2003) have been proposed to complement UDDI with rich descriptions and powerful match mechanisms in support of service discovery.

Among the extensions, the UDDI-M approach (Miles et al., 2003) is flexible in attaching metadata to various entities associated with a service, but this approach assumes the properties used in service advertisements and in service requests are consistent. Semantic Web service technologies such as OWL-S and WSMO have been proposed to enhance service discovery with semantic annotations. However, the classical OWL-S matching algorithm (Paolucci et al., 2002) cannot deal with uncertainty in service properties when matching service advertisements with service requests. This work has been extended in various ways in applying Semantic Web services for

Figure 11. ROSSE efficiency in accessing matched services



service discovery. For example, Jaeger, Rojec-Goldmann, Mühl, Liebetruth, and Geihs (2005) introduce "contravariance" in matching inputs and outputs between service advertisements and service requests using OWL-S. Li & Horrocks (2004) introduce a "intersection" relationship between a service advertisement and a service request. Majithia, Ali, Rana, and Walker (2004) introduce reputation metrics in matching services. However, these OWL-S based methods still cannot deal with missing (uncertain) properties.

WSMO introduces mediators trying to support distinct ontologies employed by service requests and service advertisements. However, the discovery mechanism (Keller, Lara, Polleres, Toma, Kifer, & Fensel, 2004) proposed in WSMO requires that properties used by both the goals and services should be consistent.

Compared with the work mentioned above, ROSSE matchmaking can deal with uncertain properties in matching services. It takes all service advertisements belonging to one service category into one search space to dynamically identify and reduce irrelevant and dependent properties which may be uncertain properties related to a service request.

CONCLUSION AND FUTURE WORK

In this article we have presented ROSSE for service discovery. ROSSE is novel in its capability to deal with uncertainty of service properties for high accuracy in service discovery. The preliminary experimental results achieved so far are encouraging. However, the following issues need to be considered for ROSSE enhancement:

- It has been shown that finding a minimal reduct in Rough set is a problem of NP-hard when the number of attributes gets large (Skowron & Rauszer, 1992). Heuristic methods need to be investigated to speed up the process in service property reduction.
- Services registered with ROSSE could be tremendous. Scalability is one the issues that need to be tackled. UDDI Version 3 (http://uddi.org/pubs/uddi_v3.htm) provides larger support for multiple registries, but the specification does not specify how these registries should be structured for enhanced scalability in service registration. Distributed Hash Table (DHT) based Peer-to-Peer (P2P) systems such as Chord (Stoica, Morris, Liben-Nowell, Karger, Kaashoek, Dabek et al., 2003) and Pastry (Rowstron &

Druschel, 2001) have shown their efficiency and scalability in content lookup. Scalability in ROSSE can be improved with DHT structured P2P systems.

- Advertised services may be further described in terms of their non-functional properties related to QoS such as reliability and cost. One challenge is how to model such QoS data so that functionally matched services can be evaluated in terms of their QoS properties.
- Currently ROSSE only supports keyword-based queries. It is expected that complex queries to be supported in ROSSE, for example, queries with a range or fuzzy queries.

REFERENCES

Andrews, T., Curbera, F., Dholakia, H., Goland, Y., Klein, J., Leymann, F. et al. (2003). Business Process Execution Language for Web Services version 1.1.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web, *Scientific American*, Vol. 284 (4), pp. 34-43.

Haarslev, V. & Möller, R. (2001). Description of the RACER System and its Applications, *Proc. of 2001 International Workshop on Description Logics (DL-2001)*, Stanford, USA.

Jaeger, M., Rojec-Goldmann, G., Mühl, G., Liebetruth, C. & Geihs, K. (2005). Ranked Matching for Service Descriptions using OWL-S, Proc. of Communication in Distributed Systems (KiVS) 2005, Kaiserslautern, Germany.

Keller, U., Lara, R., Polleres, A., Toma, I., Kifer, M. and Fensel, D. (2004). WSMO Web Service Discovery. Retrieved on XXX from, http://www.wsmo.org/2004/d5/d5.1/v0.1/20041112/d5.1v0.1_20041112.pdf

Li, L. & Horrocks, I. (2004). A software framework for matchmaking based on semantic web technology. *Int. J. of Electronic Commerce, 8*(4), pp. 39-60.

Majithia, S., Ali, A., Rana, O., & Walker, D. (2004). Reputation-Based Semantic Service Discovery, *Proceedings of WETICE 2004*, Italy.

Martin, D., Paolucci, M., McIlraith, S., Burstein, M., McDermott, D., McGuinness, D. et al. (2004). Bringing Semantics to Web Services: The OWL-S Approach, *Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004)*, San Diego, California, USA.

Miles, S., Papay, J., Dialani, D., Luck, M., Decker, K., Payne, T. et al., (2003). Personalised Grid Service Discovery, *IEE Proceedings Software: Special Issue on Performance Engineering, 150*(4), pp. 252-256.

Paolucci, M., Kawamura, T., Payne, T. & Sycara, K. (2002). Semantic Matching of Web Service Capabilities, *Proceedings of the 1st International Semantic Web Conference (ISWC)*, Berlin.

Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Science, 11*(5), pp. 341-356.

Powles, A. & Krishnaswamy, S. (2005). Extending UDDI with Recommendations: An Association Analysis Approach, *Proceedings of WSMDEIS 2005*, Miami, USA.

Roman, D., Keller, U., Lausen, H., Bruijn, J., Lara, R., Stollberg, M. et al., (2005), Web Service Modeling Ontology, *Applied Ontology, 1*(1), pp. 77 - 106.

Rowstron A. & Druschel, P. (2001). Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems, *Proceedings of Middleware 2001*, pp. 329-350, Lecture Notes in Computer Science, Springer.

ShaikhAli, A., Rana, O., Al-Ali, R., & Walker, D. (2003). UDDIe: An Extended Registry for Web Service, Proceedings of SAINT Workshops, Orlando, Florida, USA, 2003.

Skowron, A. & Rauszer, C. (1992). The discernibility matrices and functions in information systems, *Decision Support by Experience - Application of the Rough Sets Theory*, R. Slowinski (ed.), Kluwer Academic Publishers, pp. 331-362.

Stoica, I., Morris, R., Liben-Nowell, D., Karger, D., Kaashoek, M., Dabek et al. (2003). Chord: A Scalable Peer-to-Peer Lookup Protocol for Internet Applications, *IEEE/ACM Transactions on. Networks*, *11*(1), pp. 17-32.

Winiarski, R. (2001). Rough sets methods in Feature Reduction and Classification, *Int. J. Appl. Math. Comput. Sci.*, *11*(3), pp. 565-582.

Jensen, R., Shen, Q., & Tuson, A. (2005). Finding Rough Set Reducts with SAT, *Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC)*, pp. 194-203, Lecture Notes in Computer Science, Springer, Regina, Canada.

Rijsbergen, C. (1979). *Information Retrieval*, 1979, Butterworths: London.

This work was previously published in International Journal of Web Services Research, Vol. 6, Issue 1, edited by L.-J. Zhang, pp. 69-86, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 3.11
Generating Join Queries for Large Databases and Web Services

Sikha Bagui
The University of West Florida, USA

Adam Loggins
Zilliant Inc., USA

ABSTRACT

In this data-centric world, as Web services and service oriented architectures gain momentum and become a standard for data usage, there will be a need for tools to automate data retrieval. In this article the authors propose a tool that automates the generation of joins in a transparent and integrated fashion in heterogeneous large databases as well as Web services. This tool reads metadata information and automatically displays a join path and a SQL join query. This tool will be extremely useful for performing joins to help in the retrieval of information in large databases as well as Web services. [Article copies are available for purchase from InfoSci-on-Demand.com

INTRODUCTION AND RELATED WORKS

As we are working with more and more data, the sizes of databases are getting larger and larger. As businesses are going global, Web services are becoming a standard for sharing data (Srivastava et al., 2006; Resende and Feng, 2007). Enterprises are moving towards service oriented architectures where several large databases may be layered behind Web services, hence databases are having to become adaptable with loosely-coupled, heterogenous systems (Srivastava et al., 2006) too. In such scenarios of Web services and service oriented architectures, which may be dealing with several loosely coupled heterogeneous large databases, it is no longer humanly possible to have handy all the information on all the tables and primary keys in all the large databases. Although considerable work is being done on the challenges associated with Web services addressing the problem of multiple Web services to carry out particular tasks (Florescu et. al., 2003; Ouzzani and Bouguettaya, 2004), most of this work is targeted towards work-flow of applications, rather than coordinating how data can be retrieved from multiple large databases in Web services via SQL (Srivastava et al., 2006). In this article we try to address one aspect of this problem of retrieving data from multiple heterogeneous large databases using SQL. Specifically, we present a tool that automatically formulates joins by reading the metadata of databases in the context of very large databases or in the context of Web services which may employ the use of several large heteregenous databases.

Let us look at an example of a query presented to a Web service: *Suppose a health insurance company needs to verify the salary, health, and travel patterns of a person before determining the amount of health insurance he/she needs to pay. In a Web service, this will require joining of several tables. And, of course, no one person will have knowledge of all the primary key/foreign key relationships between the tables to join in the Web services.*

When databases were smaller, it was possible to have knowledge of most of the tables and primary key/foreign key relationships in databases, and SQL join queries could easily be built by joining tables in databases. But, in large databases layered behind Web services, it will not be possible to have knowledge of all the database schemas.

The join operation, originally defined in the relational data model (Codd 1970, 1972), is a fundamental relational database operation, facilitating the retrieval of information from two relations (tables). Writing efficient joins is simple for small databases since few relations are involved and one has knowledge of the complete database schema. But, writing efficient joins is a challenge in large database scenarios and Web services where it may not be possible to have a complete picture of the database schema and it's relations.

Since joins are one of the most time-consuming and data-intensive operations in relational query processing, joins have been studied discussed extensively in the literature. Mishra and Eich (1992) present a very comprehensive study of works that have been done on joins. Query optimization issues in joins, and devising strategies for distributed join processing have also been discussed by many, for example, Kim et al. (1985), Perrizo et al. (1989), Segev (1986), Swami and Gupta (1988), Yoo and Lafortune(1989), and Yu et al (1985, 1987). These works have to be extended in the context of databases for Web services and service oriented architectures. Srivastava, et. al (2006) addresses the problem of query optimization over Web services on a much broader scale.

In this article we present a tool that we have developed that will: (i) read the meta data of databases, that is, search the database model or schema and discover the relationships between the tables using table indexes defined in the database catalogs; (ii) find efficient join paths between the tables to be joined; and, (iii) generate a SQL join query (in ANSI SQL standard).

This rest of the article is organized as follows: Section two briefly describes relational databases with respect to the join operation; section three presents an architectural overview of our tool; section four presents the configuration details of our tool; section five describes how we tested our tool and presents some results; and section six presents the conclusion. Some relevant code portions are presented in the appendices.

RELATIONAL DATABASES AND THE JOIN OPERATION

In relational databases, data is stored in the form of tables or relations. Each table has information on a particular subject or concept and is composed of a group of "related" attributes. The attributes in a table are all "related" in the sense that they describe the subject or concept of the table. For example, there could be a table called Employee, with attributes emp_lastName, emp_midName, emp_firstName, emp_ssn, emp_birthdate, city, state, homePhone, cellPhone, deptnum, etc. All these attributes describe an Employee. Likewise, there could be another table called Department, with attributes, dept_Name, dept_Number, dept_manager, dept_location, etc. Here again, all these attributes describe a Department. Now, if we want information that is partly in the Employee table and partly in the Department table, for example, if we want to know which employee is working for a department located in LA, we have to perform a "join" of the Employee table and the Department table on some common attribute (usually the primary key field of one table and the foreign key field of the other table). In this case we would perform this join with a simple SQL query where Employee.deptnum = Department.dept_Number.

Usually, when a join query is composed, one has to determine which tables contain the information needed to answer the query, and has to join those tables by the key fields. This is possible if there are few tables and one has a conceptual idea of the databases. But how do we compose joins when there are hundred of tables, with an unknown (large number) of attributes per table in a database – the scenario for Web services. Moreover, the conceptual schema of the databases could be constantly evolving or changing.

So, the join operation is used to combine related tuples from two relations into single tuples that are stored in a resulting new relation. The desired relationship between the tuples or some attributes in the tuples is specified in terms of the join condition. In its simplest form, the join of two tables or relations, R and S is written as:

$$R \; |X|_{r(a)\ \Theta\ s(b)} \; S$$

where r(a) Θ s(b) defines the join condition; a and b are the attributes, usually the key fields of the respective tables R and S (usually indexed); and Θ defines the join condition that must hold true between the attributes a and b of R and S, respectively. The Θ operation can be any one of the following: $=$, $\neq$, $>$, $<$, $\geq$ or $\leq$. The join condition also includes multiple simple conditions of the form shown above connected with the logical connective AND (Earp and Bagui (2000), Elmasri and Navathe (2007)):

condition AND condition AND condition

AN ARCHITECTURAL OVERVIEW OF OUR TOOL

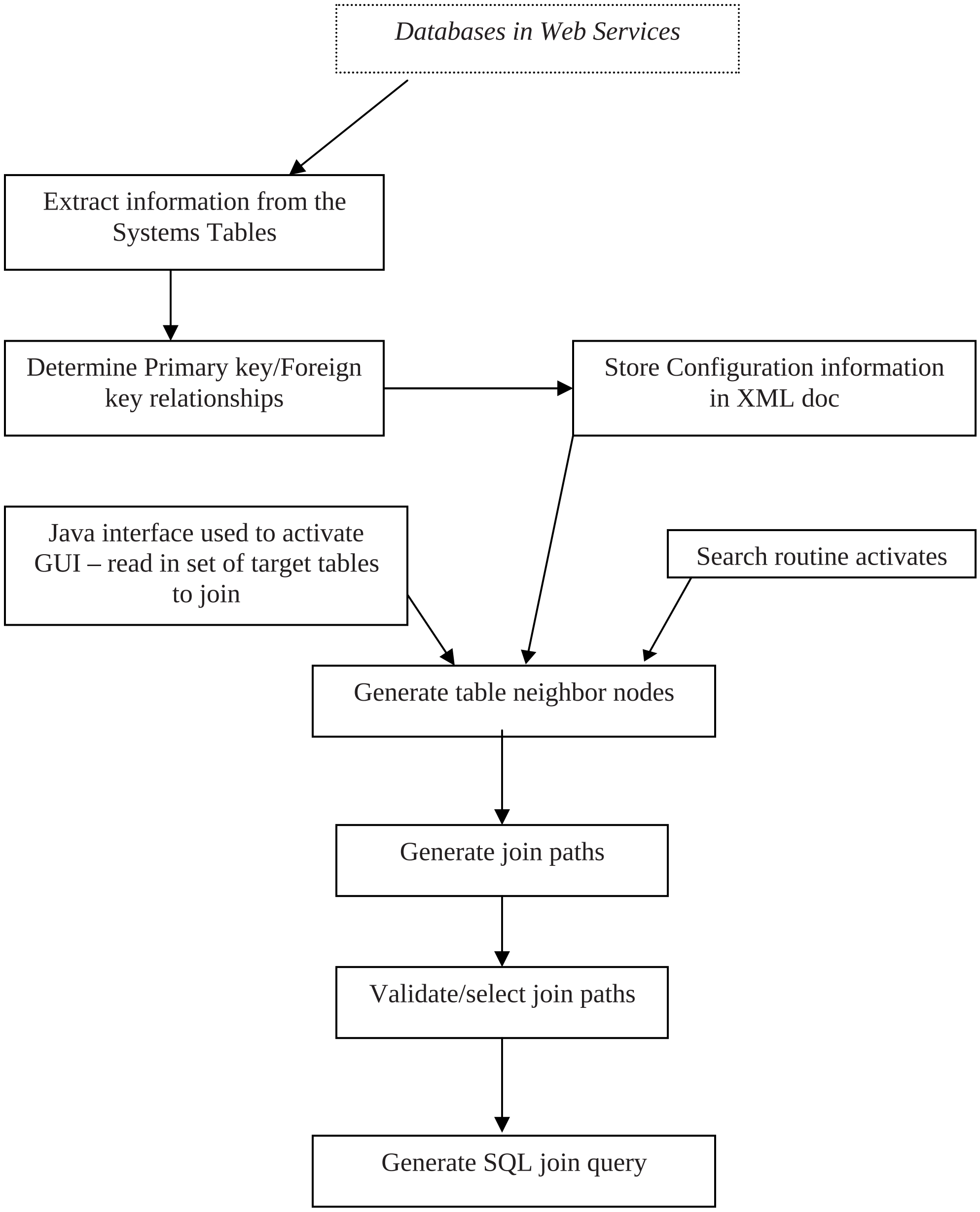
We developed a tool that extracts configuration information from system tables in databases in Web services. The primary keys of the tables in the databases are extracted. Primary key/foreign key relationships are determined. A list of table indexes is also constructed – this information is also obtained from the metadata of the databases. This information is then stored in an XML configuration document. A Java interface is then used to activate a GUI that takes, as input parameters, the names of tables that need to be joined. Then, a search routine is called that generates the table's neighbor nodes (tables) – that is, information on which table is linked to which table. From here, join paths are generated, from where a valid join path is then selected, and the final product is a SQL join query generated from the valid join path.

Below we present the algorithm of our tool.

Algorithm of our Tool

Input: *Activate GUI, input parameters, that is, the names of the tables that have the final information that is required (tables to be joined).*

Figure 1. Architectural overview of our tool



Output: *SQL Query.*

Method:

1. *For all the metadata in the required database catalogs in Web services*
   a. *Read the primary keys of the tables*
   b. *Check for primary key/foreign key relationships and generate primary key/foreign key relationship table*
2. *Call Search Routine*
3. *Generate join paths*
4. *Generate SQL query*

Figure 2. UML for tool

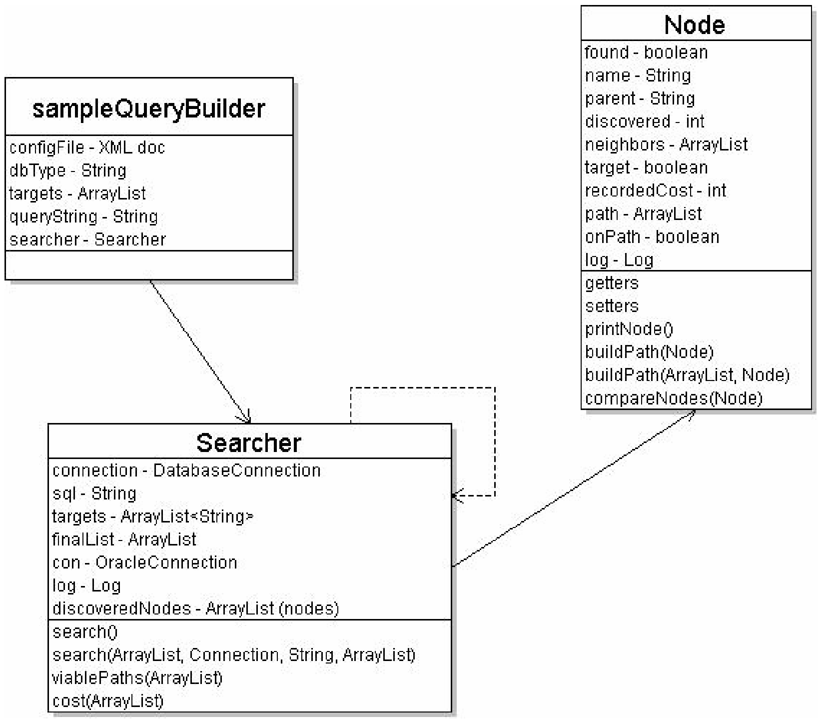


Figure 1 presents the architectural overview of our tool.

TOOL CONFIGURATION DETAILS

The UML class diagram of the tool is presented in Figure 2.

The first step is to extract the list of tables in the database from the information available in the Systems Table of the databases. This is done using the following SQL query:

```
SELECT table_name
FROM System_tables
WHERE table_type = 'TABLE'
```

The next step is to determine the primary key/foreign key relationships between the tables and indexes used for the relationships. This is done using the following SQL queries:

```
SELECT DISTINCT scr.PKTABLE_NAME AS localTable,
scr.PKCOLUMN_NAME AS localCol,
scr.FKTABLE_NAME AS foriegnTable,
scr.FKCOLUMN_NAME AS foriegnCol,
scr.FK_NAME AS key
FROM System_CROSSREFERENCE scr
WHERE scr.PKTABLE_NAME = '-tablename-'
```

And,

```
SELECT DISTINCT scr.PKCOLUMN_NAME AS localCol,
scr.FKCOLUMN_NAME AS foriegnCol
FROM System_CROSSREFERENCE scr
WHERE scr.PKTABLE_NAME = '-tablename-'
AND scr.FKTABLE_NAME = '-tablename2-'
```

This configuration information is stored using an XML document, presented in appendix 1. The XML document creates a table of the form:

Table(localTable, localColumn, foreignTable, foreignColumn)

Then, a search routine, as can be seen from UML diagram (Figure 2), reads, as input parameters, a list of table names to find relationships between, and searches for the relationships between the tables by looking up the primary key/foreign key relationships between the tables. The code of the Searcher class is presented in appendix 2. The sampleQueryBuilder class develops the GUI. This class takes in the database type (in this case, hypersonic) as the input parameter, the XML configuration file (presented in appendix 1), the table names, and calls the Searcher class, which then creates the nodes.

The Search Routine

The search routine reads, as input parameters, a list of table names to find relationship between, and searches for relationships between the tables. Below we present the algorithm of our search routine.

Algorithm of Search Routine

Until all targets are found

a. Read in the tables to be joined

b. Determine the relationships between the tables

c. For all tables

- *i. Find all neighbors*
- *ii. If targets are found stop else find neighbors of new tables*

For example, let us assume that we have the following tables in a database schema:

{A, B, C, D, E, F, G, H, I, J, K, L}

And assume that you do not know the primary key/foreign key relationships between the tables.

Now, we need information that is partly in table A, partly in table D and partly in table I. That is, we want to see if tables A and D can be joined and if tables A and I can be joined. We need to determine if the following links, as shown in Figure 3, exist.

The algorithm takes the first table: A, and finds all its neighbors. A's neighbors are all the tables that A links to. Now suppose for example, it was found that A's neighbors are B, C, and D, that is, A links to B, C, and D, as in Figure 4.

That is, A's primary key is in tables B, C and D as the foreign key, as shown in Figure 5.

So, one of the targeted links have been found, A-D. But, all the targeted links have not been found, so the algorithm keeps running until all the targeted links have been found. So now, the algorithm stores the links A-B, A-C, A-D, as shown in Figure 4.

Next the routine finds the neighbors of tables B, C, and D. Now suppose for example, B's neighbor is E, C's neighbor is F and D's neighbor is G and H. So now we have found the following links A-B, A-C, A-D, A-B-E, A-C-F, A-D-G, A-D-H, as shown in Figure 6.

Next search routine finds the neighbors of E, F, G and H. Now suppose that E links to I and J, and F links to K and L, as shown in Figure 7.

The algorithm now stops since the targeted tables have been found: A links to D directly: A-D. And, A links to B which links to E which links to I: A-B-E-I. The algorithm now keeps only these two paths as the valid join paths. Using these join paths, the next step was to generate a SQL join query.

Generating the SQL Query

Our algorithm to create the SQL join query goes down the shortest join path first. The shortest path is A-B-E-I. So, the joins will be in the form (see Box 1).

As shown in Figure 8.

TESTING THE TOOL

We tested our tool using the hypersonic database and the Java Business Process Management (JBPM) in the context of Web services. The hypersonic database (HSQLDB), freely available on the Web at http://www.hsqldb.org/, is a leading SQL relational database engine written in

Figure 3. Are these tables linked?

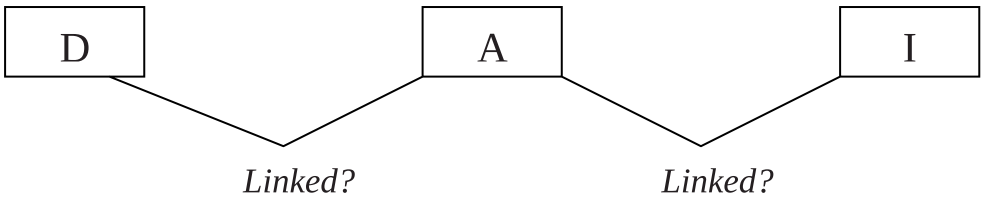


Figure 4. Neighbors of Table A

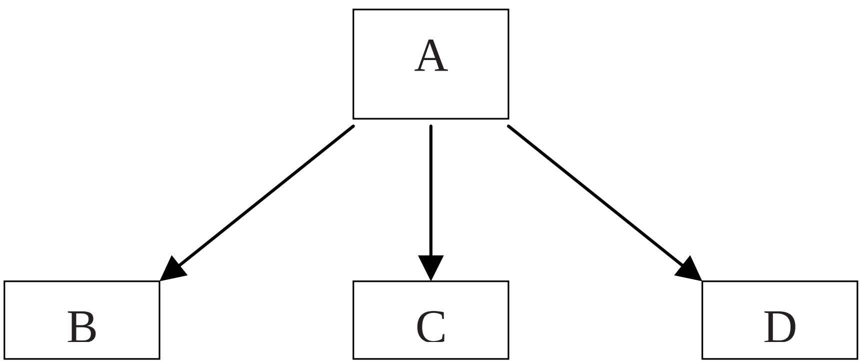


Figure 5. Primary key/Foreign key relationships

TableA(PrimaryKeyOfA, AttributeA1, AttributeA2)
TableB(PrimaryKeyOfB, AttributeB1, PrimaryKeyOfA)
TableC(PrimaryKeyOfC, AttributeC1, PrimaryKeyOfA)
TableD(PrimaryKeyOfD, AttributeD1, PrimaryKeyOfA)

Figure 6. Neighbors of Tables B, C, and D

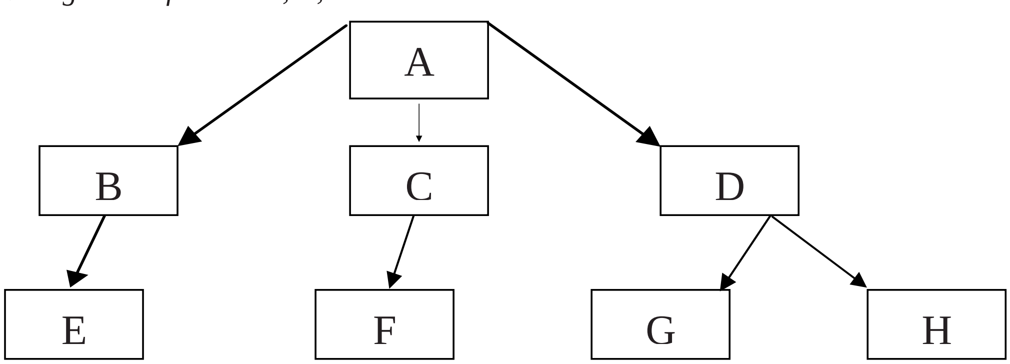
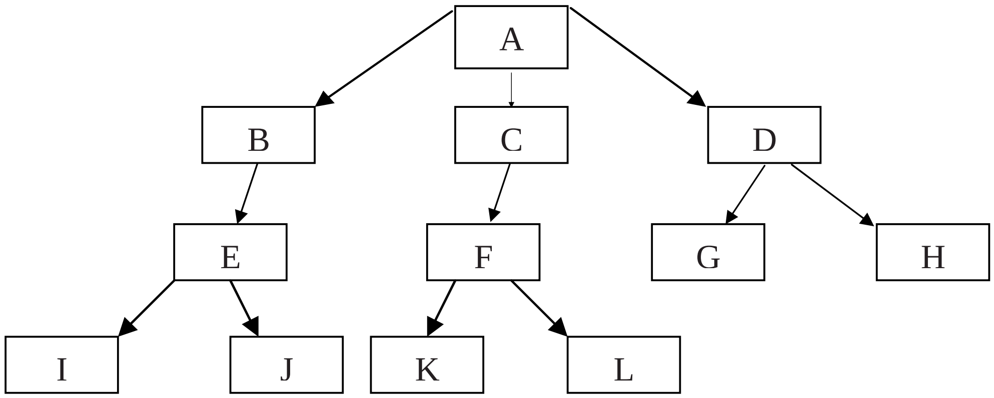


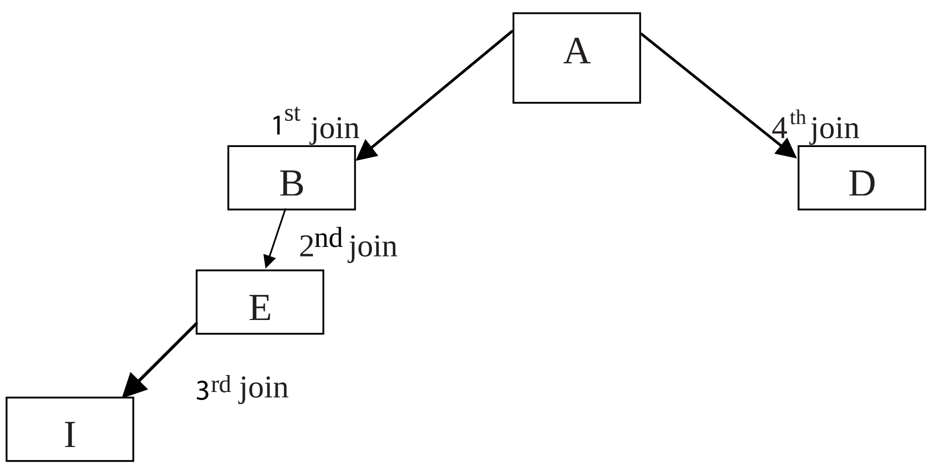
Figure 7. Neighbors of Tables E, F, G, and H



Box 1.

Table A	INNER JOIN	Table B	ON	A.X = B.X
Table B	INNER JOIN	Table E	ON	B.X = E.X
Table E	INNER JOIN	Table I	ON	E.X = I.X
Table A	INNER JOIN	Table D	ON	A.X = D.X

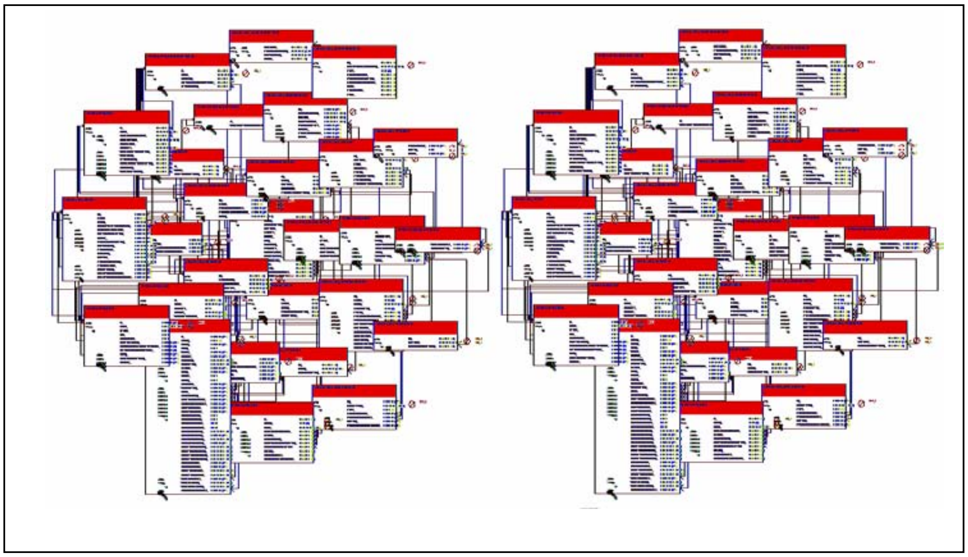
Figure 8. Join paths



Java. The HSQLDB database engine offers both in-memory and disk-based tables and supports embedded and server nodes.

The JBPM data model, available at http://www.jboss.com/products/jbpm, is a business friendly open source piece of software with an architecture that will run standalone or can be embedded with any Java application. JBPM (or JBOSS JBPM) presently has a fully-formed table schema of 38 tables with multi-table relationships. Figure 9 shows a snapshot of a portion of the complicated multi-table schema of the JBPM data model. These tables house data that record the state of a process as it progresses through its life cycle. The

Figure 9. Schema of the JBPM data model



JBPM_ProcessInstance table, for, example, stores the process instance id along with start and end dates; the JBPM_ProcessDefinition table stores the various definitions that are present for use by the application along with their versions.

Running the Application

Step 1: *From the GUI interface (shown in Figure 10), the user selects a data source (of the database). We selected JBOSS, as shown in Figure 10.*

Step 2: *The next step will be to select the tables that you want to join from that data source (the schema of the JBOSS data source is given in Figure 9). We selected the JBPM_ACTION and JBPM_BYTEBLOCK tables. This was an arbitrary selection, and the user can select any table or any number of tables by selecting the* **select another** *tab (shown in Figure 11). We will illustrate this software by using two tables. After the user selects the tables, as shown in Figure 11, the user clicks the* **find relations** *tab.*

Step 3: *Once you click* **find relations**, *you will get join path displayed in Figure 12.*

Figure 10. The GUI interface

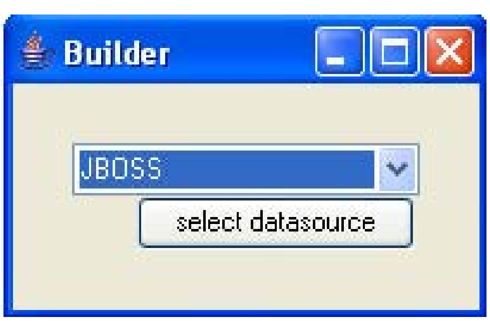
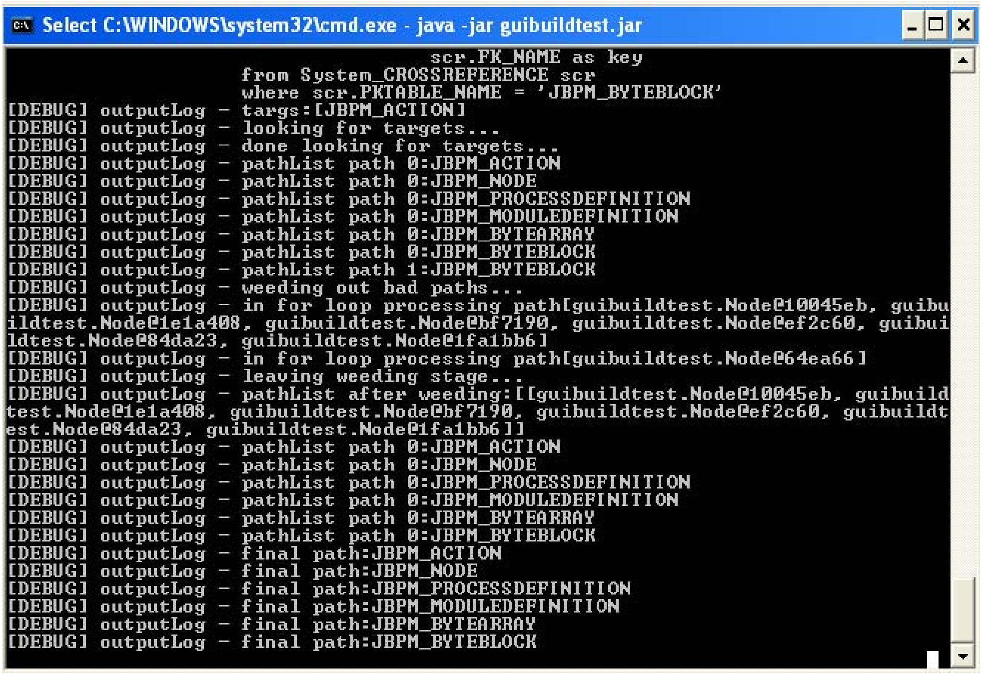


Figure 11. Selecting tables using the GUI

Figure 12. The join path



From Figure 12's output screen we can see that there is join path from the JBPM_ACTION table to the JBPM_NODE table to the JBPM_PROCESS-DEFINITION table to the JBPM_MODULEDEF-INITION table to the JBPM_BYTEARRAY table to the JBPM_BYTEBLOCK table.

This join path generates the ANSI SQL join query shown in Figure 13.

Results

Using the large database, JBPM, we ran several tests to test our software. Below we present the results of some of the test runs. The first column shows the tables that needed to be joined. The second column shows the number of tables that the algorithm needed for the output. The third column shows the resulting number of 2-table joins that the algorithm required for the output. And the last column shows the number of milliseconds it took to produce the output for this particular join.

On the average, six tables needed to be joined to obtain the results of 2 table joins, and the time averaged 53.2 mil seconds; 4 tables needed to be joined to obtain the results of 3 table joins, and

Figure 13. ANSI SQL join query generated

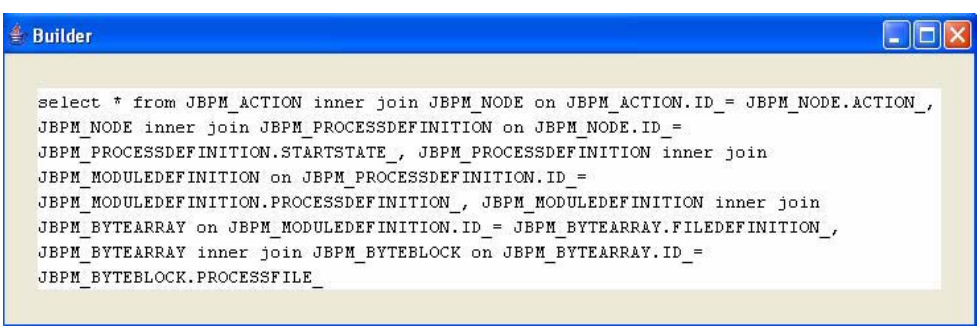


Table 1.

Tables to be joined	Number of tables to be joined	Number of 2-table joins required	Total Number of different tables joined to get result	Number of mil seconds it took to build the query
JBPM_ACTION, JBPM_BYTEARRAY	2	4	8	94
JBPM_ACTION, JBPM_RUNTIMEACTION	2	2	4	32
JBPM_NODE, JBPM_ACTION, JBPM_EVENT	3	2	4	47
JBPM_TASKINSTANCE, JBPM_SWIMLANEINSTANCE, JBPM_SWIMLANE	3	4	8	47
JBPM_LOG,JBPM_COMMENT,JBPM_PROCESSINSTANCE	3	4	8	31
JBPM_VARIABLEINSTANCE, JBPM_LOG, JBPM_TOKEN, JBPM_TOKENVARIABLEMAP	4	3	6	31
JBPM_ACTION, JBPM_PROCESSDEFINITION, JBPM_TRANSITION	3	2	4	47
JBPM_PROCESSDEFINITION,JBPM_VARIABLEINSTANCE, JBPM_PROCESSINSTANCE	3	2	4	94
JBPM_PROCESSINSTANCE, JBPM_NODE	2	2	4	78
JBPM_LOG, JBPM_VARIABLEINSTANCE	2	1	2	31
JBPM_VARIABLEINSTANCE, JBPM_PROCESSDEFINITION	2	2	4	47
JBPM_ID_PERMISSIONS, JBPM_ID_GROUP	2	-	-	16
JBPM_POOLEDACTOR,JBPM_PROCESSINSTANCE	2	3	6	62
JBPM_TIMER, JBPM_TASK	2	2	4	31
JBPM_ACTION, JBPM_PROCESSINSTANCE	2	3	6	125
JBPM_PROCESSDEFINITION, JBPM_RUNTIMEACTION	2	2	4	47
JBPM_PROCESSINSTANCE, JBPM_COMMENT	2	2	4	47
JBPM_SWIMLANE, JBPM_DELEGATION	2	1	2	140
JBPM_ACTION,JBPM_EVENT,JBPM_LOG, JBPM_SWIMLANE, JBPM_SWIMLANEINSTANCE	5	5	6	204
JBPM_PROCESSDEFINITION, JBPM_PROCESSINSTANCE, JBPM_NODE, JBPM_TOKEN, JBPM_TASKINSTANCE	5	5	6	266

Figure 14. Number of mil seconds to build a SQL join query for 2 tables

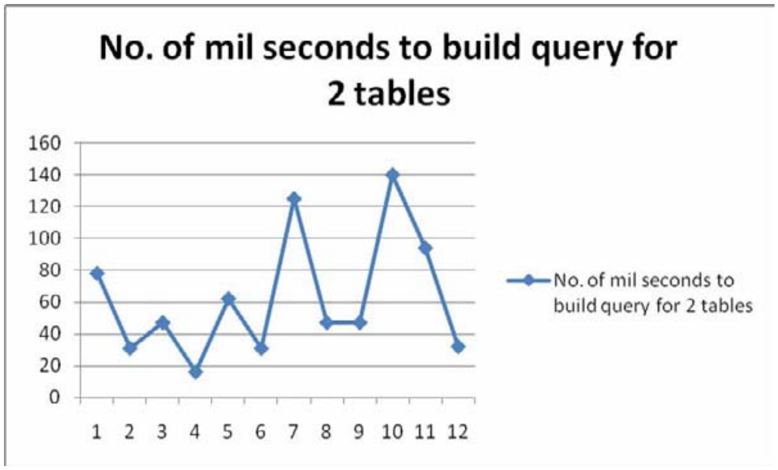
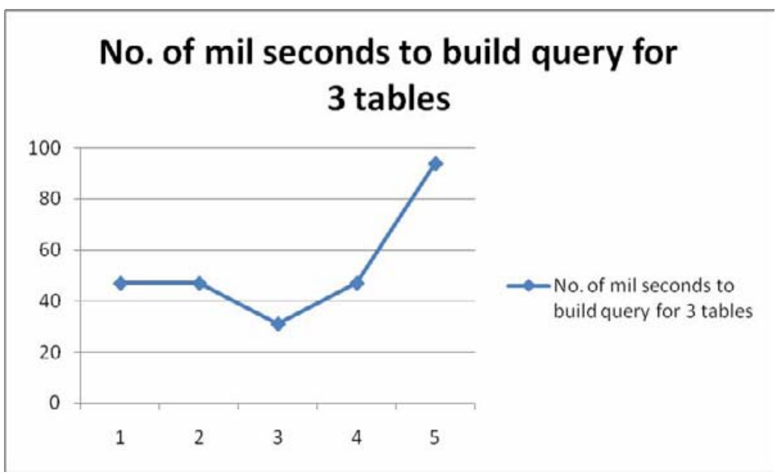


Figure 15. Number of mil seconds to build a SQL join query for 3 tables



the time averaged 62.5 seconds; six tables needed to be joined to obtain the results of 4 table joins, and the time averaged 46.5 mil seconds; six tables needed to be joined to obtain the results of 5 table joins, and the time averaged 235mil seconds.

Figure 14 presents a graphical representation of the number of milliseconds it took to build a query for two table joins, and Figure 15 presents a graphical representation of the number of milliseconds it took to build a query for three table joins.

CONCLUSION

This tool will generate join paths and a SQL join query from large databases as well as large databases in Web services. The user only needs to know which tables he/she wants to join, but does not have to know the join path needed to join them (the primary key/foreign key relationships between the tables). The software finds the join path between the tables selected, and displays a SQL join query. This software is a very important step forward in the process of retrieving information efficiently from large databases as well as large databases used in Web Services and Service Oriented Architectures.

ACKNOWLEDGMENT

The authors would like to thank the Editor, Dr. Ghazi Alkhatib, and the referees of this article for their constructive suggestions, which led to the present improved version of this article.

REFERENCES

Codd, E. (1970). A Relational Model for Large Shared Data Banks, *CACM*, *13*(6).

Codd, E. (1972). Further Normalization of the Data Base Relational Model. *Data Base Systems*. Prentice Hall.

Earp, R., & Bagui, S. (2000). Oracle's Joins, *Oracle Internals*, *2*(3), 6-14.

Elmasri, R., & Navathe, S.B. (2007). *Fundamentals of Database Systems*, (5th ed.). Boston, MA: Pearson Education.

Florescu, D., Grunhagen, A., & Kossmann, D. (2003). XL: A platform for Web services. *Proceedings of First Biennial Conference on Innovative Data Systems Research (CIDR)*.

Kim, W., Reiner, D. S., & Batory, D. S. (1985). *Query Processing in Database System*. New York: Springer-Verlag.

Mishra, P., & Eich, M.H. (1992). Join Processing in Relational Databases. *ACM Computing Surveys*, *24*(1), 63-113.

Ouzzani, M., & Bouguettaya (2004). A. Efficient access to Web services. *IEEE Internet Computing*, *8*(2), 34-44.

Perrizo, W., Lin, J.Y.Y., & Hoffman, W. (1989, June). Algorithms for distributed query processing in broadcast local area networks. *IEEE Trans. Know. Data Eng.*, *1*(2), 215-225.

Resende, L., & Feng, R. (2007, June 12-14). Handling Heterogeneous Data Sources in a SOA Environment with Service Data Objects (SDO). *SIGMOD'07*, Beijing, China.

Segev, A. (1986, March). Optimization of join operations in horizontally partitioned database systems. *ACM Trans. Database Systems*, *11*(1), 48-80.

Srivastava, U., Munagala, K., Widom, J., & Motwani, R. (2006, September 12-15). Query Optimization over Web Services. *VLDB* '06, Seoul Korea.

Swami, A., & Gupta, A. (1988). Optimizing large join query. *Proceedings of SIGMOD*, 8-17.

Yoo, H., & Lafortune, S. (1989, June). An intelligent search method for query optimization by semi-joins. *IEEE Trans. Knowl. Data Eng.*, *1*(2), 226-237.

Yu, C. T., Chang, C. C., Templeton, M., Brill, D., & Lund, E. (1985). Query Processing in a fragmented relational database system: Mermaid. *IEEE Trans. Software Eng. SE-11*, 8, 795-810.

Yu, C. T., Guh, K. C., Zhang, W. Templeton, M., Brill, D., & Chen, A. L. P. (1987). Algorithms to process distributed queries in fast local networks. *IEEE Trans. Comput. C-36*, *10*, 1153-1164.

APPENDIX A

Storing the configuration information using an XML document

```
<?xml version="1.0" encoding="UTF-8"?>
<!--
   Document   : configDoc.xml
   Author     : Owner
-->
<config>
    <datasource id="TEST1">
        <driver>org.postgresql.Driver</driver>
        <url>jdbc:postgresql://localhost:5432/TEST1</url>
        <username>postgres</username>
        <password>****</password>
                 <type>postgres</type>
    </datasource>
    <datasource id="school">
        <driver>oracle.jdbc.driver.OracleDriver</driver>
                <url>jdbc:oracle:thin:@unix.cslab.uwf.edu:1521:STUDENT_COURSE</url>
                <username></username>
                <password></password>
    </datasource>
        <datasource id="JBOSS">
        <driver>org.hsqldb.jdbcDriver</driver>
                <url>jdbc:hsqldb:E:\Graduate_Project\guibuildTest\hypersonic\localDB</url>
                <username>sa</username>
                <password></password>
                          <type>hypersonic</type>
        </datasource>
    <DBType id="oracle">
            <query id="tables">
                     select table_name
                     from ALL_IND_COLUMNS
                 </query>
        <query id="table_relation">
                     select con.table_name as localtable,
                             concol.column_name as localColumn,
                             icol.TABLE_NAME as foreignTable,
                             icol.Column_name as foreignColumn,
                             con.Constraint_name as key
                     from all_constraints con,
                          all_ind_columns icol,
                          all_cons_columns concol
```

```
            where con.constraint _ type = 'R'
            and con.R _ constrant _ name = icol.Index _ Name
            and con.constraint _ name = concol.Constraint _ Name
        </query>
</DBType>
<DBType id="hypersonic">
    <query id="tables">
            select table _ name
            from System _ tables
            where table _ type = 'TABLE'
    </query>
    <query id="table _ relation">
            select distinct scr.PKTABLE _ NAME as localTable,
                            scr.PKCOLUMN _ NAME as localCol,
                            scr.FKTABLE _ NAME as foriegnTable,
                            scr.FKCOLUMN _ NAME as foriegnCol,
                            scr.FK _ NAME as key
            from System _ CROSSREFERENCE scr
            where scr.PKTABLE _ NAME = '-tablename-'
    </query>
</DBType>
<DBType id="postgres">
    <query id="tables">
    select distinct tabs.tablename as table _ name
    from pg _ catalog.pg _ tables tabs
    where tabs.schemaname = 'public'
    </query>
    <query id="table _ relation">
    select localTab.tablename as localTable,
           locCol.column _ name as localColumn,
           refTab.tablename as foriegnTable,
           refCol.column _ name as foriegnColumn,
           keys.FK as key
    from pg _ catalog.pg _ indexes localTab,
         constraint _ column locCol,cvs
         constraint _ column refCol,
         pg _ catalog.pg _ indexes refTab,
         (select locTab.conname as localTabIndex,
                 refTab.conname as foriegnTabIndex,
                 con.conname as FK
         from pg _ catalog.pg _ constraint con,
              pg _ catalog.pg _ constraint locTab,
              pg _ catalog.pg _ constraint refTab
         where con.contype = 'f'
```

```
              and con.conrelid = locTab.conrelid
              and con.confrelid = refTab.conrelid ) keys
        where localTab.indexname = keys.localTabIndex
        and refTab.indexname = keys.foriegnTabIndex
        and locCol.constraint_name = keys.localTabIndex
        and refCol.constraint_name = keys.foriegnTabIndex
        and localTab.tablename = locCol.table_name
        and refTab.tablename = refCol.table_name
        </query>
</config>
```

This work was previously published in International Journal of Information Technology and Web Engineering, Vol. 4, Issue 2, edited by G. I. Alkhatib; D. C. Rine, pp. 45-60, copyright 2004 by IGI Publishing (an imprint of IGI Global).

Section IV
Utilization and Application

This section introduces and discusses the utilization and application of Web technologies. These particular selections highlight, among other topics, the application of semantic Web technologies to e-tourism, e-banking, and in car repairs as well as the adoption of Web services in digital libraries. Contributions included in this section provide excellent coverage of today's online environment and insight into how Web technologies impact the fabric of our present-day global village.

Chapter 4.1
The Role of Web Services: A Balance Scorecard Perspective

Pauline Ratnasingam
University of Central Missouri, USA

ABSTRACT

This chapter aims to examine the extent of Web services usage and quality, applying the balanced scorecard methodology in a small business firm as an exploratory case study. This chapter contributes to guidelines and lessons learned that will inform, educate, and promote small businesses on the importance of maintaining the quality of Web services.

INTRODUCTION

The Internet, a rapidly expanding global computer and communication infrastructure, has facilitated the emergence of digitization and globalization that in turn has permitted businesses to extensively engage in foreign investments. Forrester Research suggests that e-commerce in the U.S. will grow 19%, reaching $230 billion by 2008. Today, firms are attempting to attain their value chain goals by offering to sell products and services. Web services have become a significant part of small business, as they are used to facilitate the seamless flow of business transactions and are known to offer many benefits.

However, studies also show that the lack of effective use, quality, and security in Web service applications is one of the main reasons why firms fail to realize the full potential of their IT investments (Benko & McFarlan, 2003). It is imperative that small businesses focus on the quality of Web services and their operations given the extent to which Web service applications are used in business processes in this fast changing market conditions. Enforcing and maintaining the quality of Web services does not only involve a set of security analyses and audit procedures that most firms conduct periodically, but rather it is a continual process that needs to align a rigorous methodology. Such methodology is the balanced scorecard, which is a set of quantifiable measures that aim to monitor and manage a firm's strategic performance. This chapter aims to examine the extent of Web services usage and quality by applying the balance scorecard methodology in a small business firm.

The balanced scorecard is needed to align, monitor, and adjust the factors that impact the quality

DOI: 10.4018/978-1-60566-042-4.ch016

of Web services. Previous studies applying the balanced scorecard in the context of Web services and quality is limited. Only 10% of the organizations executed their implementation strategy to apply the balanced scorecard methodology because they experienced barriers in formulating a vision, allocating resources (i.e., human resources), and managing change (Niven, 2003). This chapter aims to examine the extent of Web services usage and quality, applying the balanced scorecard methodology in a small business firm as an exploratory case study. The next section discusses the theory of balanced scorecard and Web services followed by the development of a framework which integrates the critical success factors. Then we discuss the research method and provide a description of the background information of the firm. We then test the framework via an exploratory case study and report the findings. The findings contribute to guidelines and lessons learned that will inform, educate, and promote small businesses on the importance of maintaining the quality of Web services. Finally, we conclude the chapter with contributions and directions for future research.

BACKGROUND INFORMATION: THE BALANCED SCOREBOARD

The balanced scorecard deployed to measure the effective use and quality of Web services among small businesses focuses on a system that enforces measurement and feedback, thereby imposing quality, continuous improvement, employee empowerment, and strategic performance that aim to sustain the competitive and strategic objectives. The balanced scorecard measures the performance of Web services in a small business firm from four perspectives, namely, learning and growth, internal business processes, customer, and financial perspectives, which are discussed below. Each of these four perspectives is further categorized by their objectives (as in what are their outcomes?) measures (as in how to achieve their outcomes?) targets, that is, accountability (as in how do we know that we have achieved it?), and initiatives (as in what actions to take?). Further, the balanced scorecard is based on three time dimensional timelines, namely, yesterday, today, and tomorrow. The next section presents a discussion of the four perspectives.

1. The ***learning and growth perspective*** aims to measure the human, information, and organizational capital. Human capital includes the skills, knowledge, expertise, the extent of training given to employees, and the business cultural attitudes. Do small business employees have the skills/competencies to operate the Web service application and align it with their internal business processes effectively in order to meet their customers' objectives of using Web services? Information capital aims to measure effective communication and information sharing. Do small business employees possess the information required to achieve objectives? Organizational capital aims to monitor the soft areas of the employees, such as, learning and growth, culture, leadership, knowledge sharing, and teamwork. Do small businesses have the ability to sustain growth and change that in turn enhances the quality of Web services?
2. The ***internal business process perspective*** aims to measure performance that permits small businesses to be aware of the quality of their products and services. Web services, considered as system quality, are defined as "the conformance to explicitly stated functional and performance requirements, explicitly stated development standards, and implicit characteristics that are expected of all professionally developed software" (Solano, De Ovalles, Rojas, Padua, & Morales, 2003, p. 67). Similarly, Ortega, Perez, and Rojas (2000) suggest that product effectiveness should include characteristics

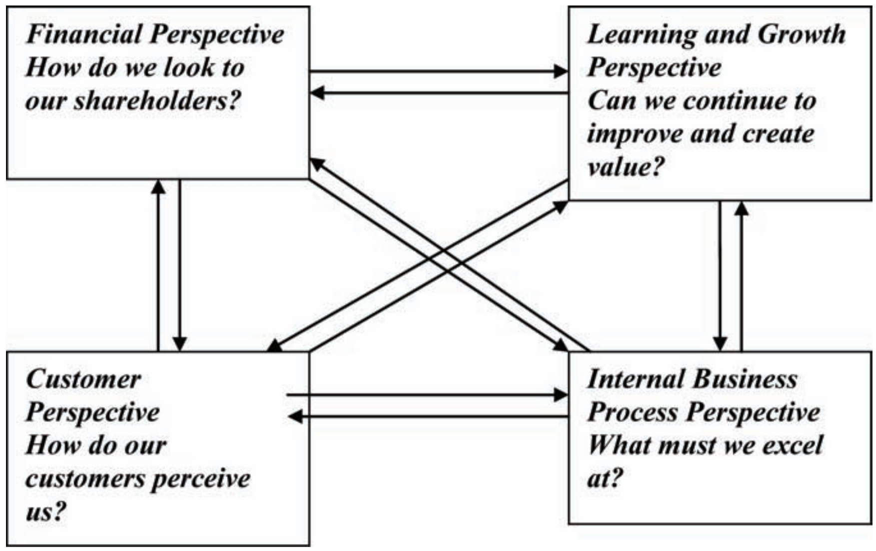
such as timeliness, functionality, reliability, usability, efficiency, maintainability, and probability. Small businesses need to be aware of the following questions when assessing the quality of their Web services performance. Does our internal business processes applying Web services conform to the mission of small businesses? Does the internal business processes meet our customer requirements? There are two types of processes under strategic management. First, the mission oriented-process focuses on the strategic goals of small businesses, and second, the support processes are more repetitive and are used in the daily operations that in turn enforce benchmarking. The balanced scorecard provides a diagnostic feedback into the various internal processes, thereby guiding and improving the business processes involved in the use of Web services on a continuous basis. What must small businesses do well internally in order to achieve the objectives they set forth to achieve quality in Web services? Where does the Web services "process" start, and where does it end?

3. The ***customer perspective*** focuses on meeting the needs of the customers, retaining existing customers, and gaining customer satisfaction. What do customers expect or demand from the use of Web services? Dimensions of customers experience include time, quality, price or cost, accessibility, reputation, and relationship. Who do we define as our customers? How do our customers see us? How do Web services create value for our customers?
4. The ***financial perspective*** aims to provide timely and accurate financial information. By implementing a centralized database, it is hoped that processing can be standardized and automated. Further, both risk assessment and cost benefit analysis can be easily conducted in order to ensure that the bottom line of small businesses is achieved. What accountability do small businesses that use Web services have to financial stakeholders? In many respects, the financial perspective represents "end in mind" of the small business strategic vision. Small business managers are able to examine the outcomes of their business performance that provide strategic financial feedback and show the trends of their business performance using Web services overtime. Figure 1 presents the perspectives of the balanced scorecard.

WEB SERVICES

Web services are creating a service-oriented architecture that provides technical solutions and e-collaborations among value chain partners (Chan, Kellen, & Nair, 2004; Yen, 2004). Web services refer to a new breed of Web applications as self-contained, self-describing modular applications that can be published, located, and invoked across the Web. Businesses use existing software components specified as services to perform business operations in a "service-oriented architecture." Similarly, Web services refer to a set of software applications or components developed using a specific set of application programming interface (API) standards and Internet-based communication protocols. The objective is to enable these applications or components to invoke function calls and exchange data among themselves over the standard Internet infrastructure. We define Web services as "modular Internet-based business functions that perform specific business tasks to facilitate business interactions within and beyond the organization." Further, Web services generate sustainable competitive advantage for firms supporting their core competencies and adding value to the execution of the corporate strategy. The technology's fullest potential will not be realized if it is used only for improving the operational efficiency of existing business processes. Therefore,

Figure 1. The balanced scorecard perspectives



we focus on the quality of Web services usage by applying a rigorous methodology called the balanced scorecard methodology as it measures the effective use and quality of Web services from four perspectives, namely, learning and growth, internal business processes, customer, and financial perspectives, thereby providing a holistic view.

The Web Services Architecture

The Web services architecture is made up of three layers of technology. At the foundation (the bottom layer) is the software standards and communication protocols that provide a common language for Web services and enable applications to be connected. Standards and protocols are often cited as the strategic value that Web services bring (Hagel, 2002; Hagel & Brown, 2001). The service grid (the middle layer) provides a set of shared utilities from security to third-party auditing, to billing and payment so that critical business functions and transactions over the Internet can be conducted. This layer publishes the services that serve as entry points for queries to find service descriptions. In short, the service grid plays two roles. First, it helps the Web service requestors and providers to find and connect with one another; secondly, it creates a trusted environment essential for carrying out mission-critical business activities, thereby contributing to technology trust. Finally, the top layer consists of a diverse array of application services. It is in this top layer where the day-to-day operations will be most visible to employees, customers, and trading partners. The top layer performs the service binding and invocation. Similarly, three are layers of Web services, namely, basic services, composite services, and managed services. While basic services manage the publication, discovery, selection, and binding, composite services facilitate the coordination, conformance, monitoring, and quality of service. Finally, managed services provide the market certification rating and operations support.

Table 1 presents the three layers of the Web services architecture adapted from Hagel and Brown (2001).

Table 1. The Web services architecture

Top Layer – Application Service Web services runs on any application platform as long as it has a Web server connected to the Internet.
Middle Layer – Service Grid The service grid layer provides four types of utilities: (1) Service management utilities (i.e., provisioning, monitoring, ensuring quality of service, synchronization, conflict resolution) (2) Resource knowledge management utilities (i.e., directories, brokers, registries, repositories, data transformation) (3) Transport management utilities (i.e., message queuing, filtering, metering, monitoring, routing, resource orchestration) (4) Shared utilities (i.e., security, auditing, and assessment of third party performance, billing and payment)
Bottom Layer – Software Standards and Communication Protocols Software standards include: (1) Web service description language (WSDL) to describe the Web service (2) Universal, description, discovery and integration (UDDI) to advertise, syndicate as a central organization for registering, finding, and using Web services) (3) Web services flow language (WSFL) to define work flows (4) XML (format for data exchange and description) (5) Communication protocols including simple object access protocol (SOAP) to communicate that are for calling Web services, HTTP, and TCP/IP)

FRAMEWORK OF THE BALANCED SCORECARD FOR WEB SERVICES

The framework of balanced scorecard for Web services was developed by integrating the theory of balanced scorecard and Web services. The framework consists of critical success factors or indicators that make up the objectives, measures, targets, and initiatives. The goal of these critical success factors is to evaluate the effective use and quality of Web service applications. Table 2 below illustrates the framework of the balanced scorecard in Web services, which serves as a measurement tool, thereby ensuring the quality of Web services.

RESEARCH METHOD

Case studies were chosen as an appropriate method to evaluate the effective use and quality of Web services among small businesses as it elicited subtle and rich data needed, thereby increasing our understanding in the use and quality of Web services applying the balanced scorecard methodology (Yin, 1994).

We attempted to identify the critical success factors in the effective use and quality of Web services by deploying the balanced scorecard framework for small businesses based on the objectives (what are the outcomes?) measures (how to achieve the outcomes?), targets, that is, accountability and initiatives (how do we know that we have achieved it?), and initiatives (what actions to take?) for all the four perspectives (i.e., learning and growth, internal business processes, customer, and financial perspectives).

In this study we interviewed the managers of the firm in the agricultural industry. Initial entry into the case site was obtained by making telephone calls to key representatives in the organization. A brief description and purpose of the study was discussed before requesting them to participate. The telephone conversation was followed by an e-mail to the respondents with an attachment of the file describing the purpose of the study. Once confirmation of their willingness to participate in the study was received, appointment dates for interview sessions were arranged with the managers in the firm. Evidence for the exploratory case study data came from the hand written notes taken during the two (90 minutes) face-to-face interviews and the tape recorded data. In addition, analysis of existing documents relating to Web service applications, day-to-day

Table 2. Framework of the balanced scorecard in Web services

Balanced Scorecard Perspectives	Relationship to Effective use and Quality of Web Services via the Critical Success Factors (or Indicators)
(1) ***(1) Learning and growth perspective*** *How can we continue to improve and create value in the use of Web services?*	***Objectives:*** employees must be well trained and are expected to perform their day-to-day business operations applying Web services ***Measures:*** provide online training, user manuals, standard operating procedures, help desk, and reward employees with high productivity ***Targets:*** fewer customers and stakeholder complaints ***Initiatives:*** ongoing monitoring of employees performance, focus on the culture, climate, and commitment of the organization.
(2) ***Internal business process perspective*** *What processes do we need to excel further when using Web services?*	***Objectives:*** to achieve high quality and productivity of the services provided via Web services ***Measures:*** apply best business practices, reliable, accurate, and timely information. Focus on the usability, interoperability of the system ***Targets:*** increased profit and fewer customers and stakeholders complaints ***Initiatives:*** ongoing audit and applying the quality assurance plan on a regular basis
(3) ***Customer perspective*** *How can we enhance our business reputation with our customers via Web services?*	***Objectives:*** satisfaction of customers and stakeholders, reputation of the firm and the quality of their products and services ***Measures:*** open, frequent communications, providing excellent quality services, provide value for money and reliable operations ***Targets:*** increase in profit and sales ***Initiatives:*** training employees, ongoing weekly meetings with employees; regular review and reflection of the goals and mission of the company
(4) ***Financial perspective*** *How are we perceived by our shareholders and other stakeholders invested in our firm?*	***Objectives:*** increase in profits, rate of return ***Measures:*** increased productivity, increase quality of services provided, apply the return on capital employed, economic value added and free cash flow ***Targets:*** profit figures, increased shareholders value ***Initiatives:*** advertising their company products, attending trade shows, business conferences and seminars, word of mouth, control operating costs, maximize potential use of Web services

interactions, operating procedures, Web sites, and internal security policies were analyzed. The tape recorded data was transcribed to identify concepts via pattern matching to the balanced scorecard framework.

FINDINGS: BACKGROUND INFORMATION OF FIRM A

Firm A is a small business seed manufacturer and seller of wild flower and bird seeds located in Kingsville. It is a family owned business with 10 employees and the owner has been in this business for 23 years. They sell a variety of wild flowers seeds including the annual mix, shade, and suns. They also supply bird seeds to regular stores, residential customers, and fulfill large bids from the government for beautifying the land, such as for the city of Blue Springs and the Tiffany Springs highway. Their main form of Web service application is their business-to-consumer (B2C) Web site implemented in 2003, which has embedded IT solutions, e-mail, and fax. Table 3 presents the background information of Firm A.

In this section we report the findings of the exploratory case study from the hand written notes taken during the face-to-face interview with the manager of Firm A.

Web Services Quality in the Learning and Growth Perspective

Firm A found that the learning capability in applying best business practices was important for their business performance. Although only six

Table 3. Background information of firm A

Background Characteristics	Firm A
Type of industry	Agricultural industry
Size of Firm	Small business
Type of ownership	Family owned
Number of employees	10
Type of Web service application	B2C Web site & embedded IT solutions
Type and number of customers/business partners	Regular store customers, government bid contracts, Web customers
Annual revenue in US$	$1.5 million
How long have they being in business?	23 years
Mode of attracting customers	Advertise in the local newspaper, Web page

employees were assigned to operate their computer systems, they had to abide to best business practices that included changing their passwords every 10 days and not disclosing pricing information to any other employees other than the manager and two other employees in the accounts department. Each employee was given limited access rights and was unable to see the detail information of the transaction.

The manager noted, "*We continually improve our employee skills by providing training to our employees in order to increase the potential use of Web services.*"

Web services were deployed to facilitate information sharing and collaboration among employees and business units.

Web Services Quality in Improving Internal Business Processes

The internal business process perspective focused on the quality and use of Web services in activities such as supply chain management, customer relationship management, and research and development. The balanced scorecard provided a systemic quality approach to assess Web services as it allowed the software processes to be efficiently and effectively balanced. The manager noted, "*We applied the quality assurance plan which included the following critical success factors: timeliness of obtaining information or processing the transaction, accuracy as in achieving integrity in the content of the message, confidentiality, access rights, non-repudiation, reusability, portability, reliability, security and efficiency of the Web service applications were considered in this perspective.*"

The manager indicated that, "*Further, with industry-accepted standards and protocols, Web services provided a standard interface allowing integration among heterogeneous platforms, thus facilitating efficient and effective collaboration between departments that use different IT systems. Finally, Web services' service-oriented architecture allows firms to build a flexible IT infrastructure that enables faster decision-making and response to market changes.*"

Hence, through the orchestration of modular, loosely coupled software components, Web services enable an "assemble line approach" to software development, resulting in a responsive IT infrastructure for designing and building faster application development and enterprise applications.

We argue that Web services technology, with its industry-accepted standards and protocols, can enhance internal business operations by enabling process automation, increasing interoperability and reducing integration complexity, and improving process design.

Web Services Quality in Improving Customer Retention and Relationships

The customer perspective is the core of the business strategy. The manger noted, *"Our Web services and IT solutions offered us with unique business processes and customer value propositions that determined the correct business processes thereby creating better customer relationships."*

The manager also noted, *"We have key attributes of Web services that create customer value propositions such as; enhanced customer intimacy via open communications, improved customer retention, and better customer value. Beyond the quality and specifications of its products and services we try to satisfy our customers by meeting their needs and offering quality goods and services."*

These attributes that serve as critical success factors were derived from the balanced scorecard methodology which comprises of objectives, measures, targets, and initiatives.

The manager indicated, *"Web services made our firm's IT infrastructure more flexible and adaptable, affording the organizational agility to meet the ongoing customers' changing requirements."*

Web Services Quality in Improving Financial Position

The use of the balanced scorecard methodology allowed for improved capability of learning and innovation, better internal business processes, and the enhanced customer value that in turn served as performance drivers for increased financial returns.

The manager noted, *"Web services has directly influenced our shareholder value as it influenced our firm's financial strategy, productivity and revenue growth."*

Further, he added, *"For example, in the financial perspective we aimed to create value for the shareholders and there is a balance between growth and productivity. Further, return on capital employed and economic value indicators are added."*

LESSONS LEARNED

The findings suggests a cyclic process that was created with the use of the balanced scorecard approach to evaluate the quality of Web services applications and in order to integrate quality, provide a strategic map, and indicate how information will be disseminated so that the potential use of Web services can be attained. The processes adapted from Kaplan and Norton (2000) included:

- **Analysis of the sector, its development and role of company:** Refers to the identifications of the key goals in the use of Web services and establishing the characteristics and requirements for the industry.
- **Establishing or confirming the company's strategic plan:** Refers to the establishment or confirmation of a strategic plan, intensifying the internal and external analysis of the earlier processes, and ensuring that agreements are arrived towards the quality of Web services.
- **Translating the strategy into operational terms:** Refers to the actual actions taken to ensure that best business practices, standards, and quality procedures that were followed in the use of Web services. For example, in the financial perspective they aim to create value for the shareholders and there is a balance between growth and productivity. Return on capital employed and economic value indicators are added. Likewise, in the customer's perspective, the growth in terms of volume generated from customers was examined. Further, segments that value quality and innovation were emphasized. In the internal business

process perspective they try to differentiate between basic operating processes and operating excellence in the support services via the use of Web services. The product and service quality were measured through the product quality index indicator using the market share in order to gain profit from the investment in the financial perspective. Finally, in the learning and growth perspective, three strategic objectives were identified, namely, basic competencies and skills (referring to the skills expected of the employees), technology (referring to the Web services applications used in the value chain), and climate for action (referring to organization commitments that must be implemented by the human resources department).

- **Aligning the organization with the strategy:** Refers to the alignment of business unit goals with the organization's goal in the use of Web services.
- **Making the strategy everyone's daily job:** Refers to the linking of the employees with the business unit and the organization's strategy.
- **Making the strategy an ongoing process:** Refers to proposing a process for managing the strategy by integrating the tactical management and the strategic management in the use of Web services.
- **Promoting change through management leadership:** Refers to the involvement of the management team in the change process.

CONCLUSION

In this chapter we discussed the role of Web services usage and quality, applying the balanced scorecard methodology. We developed a framework that presented the balanced scorecard's four business perspectives (i.e., learning and growth, internal business processes, customer, and financial perspectives) and tested it via an exploratory case study of a small business firm within the agricultural industry. Then we reported the findings based on the impact of Web services quality on the four perspectives. The findings suggest that the lessons learned evolve over a set of processes that are aimed at integrating quality into the potential use of Web services.

The contribution of this study was attributed to the framework, which provided guidelines for measuring Web services usage and quality, applying the key Web services features that gave rise to business values that were matched with strategic initiatives that small businesses can have in each of the business perspectives. Web services further drive the quality of IT strategy as it is fully aligned with overall business strategy, thereby focusing on the key organizational capabilities and sustainable competitive advantage.

This study contributes to the theory as it extends the current literature of Web services to include the balanced scorecard framework of measure to its usage and quality. Further, the study contributes to practice as small businesses using Web services will benefit from the balanced scorecard as they will have a system which provides them with timely, cost-effective, scalable, manageable and reliable feedback on their strategic performance. The business issues including effective organizational performance and successful strategy implementation will be greatly enhanced. Further, the balanced scorecard gives a holistic view of the small business by simultaneously examining its performance from four perspectives. It is able to translate an organization's vision and strategy into specific objectives, monitored through a coherent set of performance indicators. Future research should aim to apply the balanced scorecard framework for Web services via a survey or multiple case studies with firms across different sections in the industry so that the findings can contribute to a generic framework.

REFERENCES

Benko, C., & McFarlan, F. W. (2003). *Connecting the dots: Aligning projects with objectives in unpredictable times*. Boston: Harvard Business School Press.

Chan, S. S., Kellen, V., & Nair, C. (2004). *Adopting Web services for B2B e-collaboration: Research directions.* Paper presented at the International Resources Management Association Conference, Innovations through Information Technology (pp. 1191-1193).

Hagel, J. (2002). Edging into Web services. *The McKinsey Quarterly, 4.*

Hagel, J. III, & Brown, J. S. (2001). Your next IT strategy. *Harvard Business Review*, *79*(9), 105–115.

Kaplan, R., & Norton, D. (2000). *The strategy focused organization.* Harvard Business School Press.

Niven, P. R. (2003). *Balanced scorecard. Step-by-step for government and non-profit agencies.* John Wiley & Sons.

Ortega, M., Perez, M., & Rojas, T. (2000). *A model for software product quality with a system focus.* Paper presented at the 4th World Multi Conference on Systematic Cybernatics.

Solano, J., De Ovalles, M. P., Rojas, T., Padua, A. G., & Morales, L. M. (2003, Winter). Integration of systemic quality and the balanced scorecard, privacy and security in e-business. *Information Systems Management*, 66–81. doi:10.1201/1078/43203.20.1.20031201/40086.9

Yen, V. C. (2004). *Applications development with Web services.* Paper presented at the International Resources Management Association Conference, Innovations through Information Technology (pp. 875-876).

Yin, R. K. (1994). *Case study research: Design and methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

APPENDIX: CASE STUDY QUESTIONNAIRE

A) Demographic Section

1. Name of your organization?
2. Your job title?
3. How long have you worked in your present organization?
   A. Less than a year
   B. Between 1-5 years
   C. Between 6-10 years
   D. Between 11-20 years
   E. Over 20 years
4. How long have you had your present position?
   A. Less than a year
   B. Between 1-2 years
   C. Between 3-5 years
   D. Between over 5 years
5. What is your organization's reach?
   A. Local
   B. Regional
   C. National
   D. Global
6. What is your organization's size?
   A. Large > 500 employees
   B. Small-Medium-Enterprise 1-499 employees
7. Please name the type of industry your organization is or provide the NAICS code?
8. Please indicate which of the following terms best describes your organization's main business activity
   A. Retail/wholesale trade
   B. Manufacturing/distribution
   C. Computers/communications
   D. Financial services
   E. Education
   F. Health
   G. Government services
   H. Other services
   I. Other --
9. What is the main role of your organization?
   A. Buyer
   B. Seller
   C. Manufacturer
   D. Supplier
10. Please indicate which of the following types of business transactions are actively supported by the IT systems in your organization?

A. Purchase orders
B. Invoices
C. Advance shipping notices
D. Product information
E. Payment transactions

11. Please indicate the type of business applications and tools your organization implemented, or will be implementing?
    A. EDI – Value-Added-Network
    B. Internet-based EDI
    C. Extranets
    D. Intranets
    E. E-mail
    F. B2C or B2B shopping cart
    G. Other types of Web service applications
    H. Other ---
12. Please indicate the number of business partners your organization has? (business partners refer to those who are contracted legally to do business with the firm)
    A. 1-20
    B. 21-50
    C. 51-100
    D. 101-499
    E. Over 500
13. Approximately what is your annual revenue, in millions? (please estimate)
    A. 0-1m
    B. 1-10m
    C. 10-100m
    D. 100-500,
    E. 500-1,000m
    F. Over 1,000m
14. Approximately how much does your organization spend annually on information technology? (please estimate)
    A. 0-100,000
    B. 100,000-500,000
    C. 500,000-1m
    D. 1-5m
    E. Over 5m
15. Please indicate how your organization chose your business partners?
    A. Advertising on your Web page
    B. Screening of business partners
    C. Based on past reputations
    D. Other --
16. When did your organization start implementing a balanced scorecard system? Please indicate the year.

17. Please indicate the importance of implementing a balanced scorecard system in your organization's current business strategy
    A. No importance/not considered
    B. Small consideration
    C. Nominally part of strategy
    D. Substantially part of strategy
    E. Crucial to strategy
18. On average how often does your organization meet face-to-face with your business partners?
    A. At least once per week
    B. At least once per month
    C. More than once a month
    D. Other ---
19. Please indicate how your organization maintains your business partners
    A. Verbal agreements
    B. Legal business partner agreements (written)
    C. Screening of business partners (based on a performance assessment)
    D. Other --

Questions on your Firm's Application of the Balanced Scorecard on Web Services

How did your firm implement a balanced scorecard methodology?What is the vision of your organization towards the use of Web services?Do your employees have a solid understanding of your firm's use of Web services?Why was the balanced scorecard important in measuring your business performance?What benefits did your firm experienced from the quality of Web service applications?What risks/challenges did your firm experienced from using Web services?

Learning and Growth Perspective

Do small business employees have the skills/competencies to operate the Web service application and align it with their internal business processes effectively in order to meet their customers' objectives of using Web services?

Do small business employees possess the information required to achieve objectives?

Do small businesses have the ability to sustain growth and change that in turn enhances the quality of Web services?

How does the balanced scorecard impact the quality of Web services in the learning and growth perspective?

How was the data gathered, analyzed, evaluated for feedback so that appropriate actions can be taken to improve the learning and growth perspective?

What skills do your employees possess regarding performance management and more specifically, developing performance measures in the balanced scorecard framework?

(This question relates to what qualification requirements do the senior employees who undertake the role of a balanced scorecard auditor or quality control manager has? The purpose is to ensure that they have the adequate skills to completely measure the firm's performance).

Do your employees have the right skills/competencies to operate the Web service applications?

Do your employees have the tools and information they require to make effective decisions that impact customer outcomes?

Do your employees possess the information required to achieve objectives?

Does your organization have the ability to sustain growth and change?

Do your employees face difficulties when accessing critical information needed to serve customers?

Internal Business Process Perspective

Does the internal business processes applying Web services conform to the mission of small businesses?

Does the internal business processes meet their customer requirements?

What must small businesses do well internally in order to achieve the objectives they set forth to achieve quality in Web services?

Where does the Web services "process" start, and where does it end?

How does the balanced scorecard impact the quality of Web services in the internal business processes? How were the data gathered, analyzed, evaluated for feedback so that appropriate actions can be taken to improve the internal business process perspective?

Do your employees know how their day to day actions contribute to the organization's success? (how and why)

What must your firm do well internally in order to achieve the objectives set forth in the customer perspective?

Customer Perspective

What do customers expect or demand from the use of Web services? Dimensions of customers experience include; time, quality, price or cost, accessibility, reputation, and relationship.

Who do we define as our customers?

How do our customers see us? How does Web services create value for our customers?

How does the balanced scorecard impact the quality of Web services and its relationship with customers' perspectives? What factors are evaluated in the customers' perspective?

How the data was gathered, analyzed, evaluated for feedback so that appropriate actions can be taken to improve the customer perspective?

Who does your firm define as your customers?

What do your customers expect or demand from your firm? (factors pertaining to time, quality, price or cost, accessibility, reputation, relationship and image).

Do you require anything from your customers? (in order to meet your customers demands – is there anything you need from them?)

You have a unique value proposition for customers (for example cost, technical superiority, customer intimacy).

What factors does your firm excel in as in evidence of past successes thereby providing accountability to your stakeholders that you are satisfying customer expectations?

(This question relates to what factors do the firm excel in as in evidence of past successes thereby providing accountability)

Financial Perspective

What accountability do small businesses that use Web services have to financial stakeholders?

How does the balanced scorecard impact the quality of Web services performance from a financial perspective? What factors are evaluated in the financial perspective?

How are the data gathered, analyzed, evaluated for feedback so that appropriate actions can be taken to improve the financial perspective?

What reporting mechanisms are deployed in your firm in order to enforce accountability?

Does your firm create significant value from intangible assets such as employee knowledge and innovation, customer relationships, and a strong culture. (how and why)

Does your senior management team spend time together discussing variances from plan and other financial issues? (How often?)

Does your organization clearly define the performance targets for both financial and non-financial indicators? (how and why)

This work was previously published in Managing Web Service Quality: Measuring Outcomes and Effectiveness, edited by K. Khan, pp. 342-357, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 4.2
Semantic Web Take-Off in a European Industry Perspective

Alain Léger
France Telecom R&D, France

Johannes Heinecke
France Telecom R&D, France

Lyndon J.B. Nixon
Freie Universität Berlin, Germany

Pavel Shvaiko
University of Trento, Italy

Jean Charlet
STIM, DPA/AP-Hopitaux Paris & Université Paris 6, France

Paola Hobson
Motorola Labs, UK

François Goasdoué
LRI, CNRS et Université Paris Sud XI, France

ABSTRACT

Semantic Web technology is being increasingly applied in a large spectrum of applications in which domain knowledge is conceptualized and formalized (e.g., by means of an ontology) in order to support diversified and automated knowledge processing (e.g., reasoning) performed by a machine. Moreover, through an optimal combination of (cognitive) human reasoning and (automated) machine processing (mimicking reasoning); it becomes possible for humans and machines to share more and more complementary tasks. The spectrum of applications is extremely large and to name a few: corporate portals and knowledge management, e-commerce, e-work, e-business, healthcare, e-government, natural language understanding and automated translation, information search, data and services integration, social networks and collaborative filtering, knowledge mining, business intelligence and so on. From a social and economic perspective, this emerging technology should contribute to growth in economic wealth, but it must also show clear cut value for everyday activities through technological transparency and efficiency. The penetration of Semantic Web technology in industry and in services

DOI: 10.4018/978-1-60566-066-0.ch001

is progressing slowly but accelerating as new success stories are reported. In this chapter we present ongoing work in the cross-fertilization between industry and academia. In particular, we present a collection of application fields and use cases from enterprises which are interested in the promises of Semantic Web technology. The use cases are focused on the key knowledge processing components that will unlock the deployment of the technology in industry. The chapter ends with the presentation of the current state of the technology and future trends as seen by prominent actors in the field.

CURRENT SITUATION

As a result of the pervasive and user-friendly digital technologies emerging within our information society, Web content availability is increasing at an incredible rate but at the cost of being extremely multiform, inconsistent and very dynamic. Such content is totally unsuitable for machine processing, and so necessitates too much human interpretation and its respective costs in time and effort for both individuals and companies. To remedy this, approaches aim at abstracting from this complexity (i.e., by using ontologies) and offering new and enriched services able to process those abstractions (i.e., by mechanized reasoning) in a fully – and trusted - automated way. This abstraction layer is the subject of a very dynamic activity in research, industry and standardization which is usually called "Semantic Web" (see for example, DARPA, European IST Research Framework Program, W3C initiative, OASIS). The initial application of Semantic Web technology has focused on Information Retrieval (IR) where access through semantically annotated content, instead of classical (even sophisticated) statistical analysis, aimed to give far better results (in terms of precision and recall indicators). The next natural extension was to apply IR in the integration of enterprise legacy databases in order to leverage existing company information in new ways. Present research has turned to focusing on the seamless integration of heterogeneous and distributed applications and services (both intra- and inter-enterprise) through Semantic Web Services, and respectful of the legacy systems already in place, with the expectation of a fast return on investment (ROI) and improved efficiency in e-work and e-business.

This new technology takes its roots in the cognitive sciences, machine learning, natural language processing, multi-agents systems, knowledge acquisition, automated reasoning, logics and decision theory. It can be separated into two distinct – but cooperating fields - one adopting a formal and algorithmic approach for common sense automated reasoning (automated Web), and the other one "keeping the human being in the loop" for a socio-cognitive Semantic Web (automated social Web) which is gaining momentum today with the Web 2.0 paradigm[1].

On a large scale, industry awareness of Semantic Web technology has started at the EC level with the IST-FP5 thematic network Ontoweb[2] [2001-2004] which brought together around 50 motivated companies worldwide. Based on this experience, within IST-FP6, the Network of Excellence Knowledge Web[3] [2004-2008] made an in-depth analysis of the concrete industry needs in key economic sectors, and in a complementary way the IST-FP6 Network of Excellence REWERSE[4] was tasked with providing Europe with leadership in reasoning languages, also in view of a successful technology transfer and awareness activities targeted at the European industry for advanced Web systems and applications. This impetus will continue and grow up in the EU IST-FP7 [2007-2013][5].

The rest of the chapter is organized as follows. Four prototypical application fields are presented in Section 2, namely (1) healthcare and biotechnologies, (2) knowledge management (KM), (3) e-commerce and e-business, and finally, (4) multimedia and audiovisual services. Finally, Section

3 reports on a current vision of the achievements and some perspectives are given.

Overall Business Needs and Key Knowledge Processing Requirements

Use Case Collection

In order to support a large spectrum of application fields, two EU FP6 Networks of Excellence NoE-Knowledge Web and NoE-REWERSE are tasked with promoting transfer of best-of-the-art knowledge-based technology from academia to industry. The networks are made up of leading European Semantic Web research institutions that co-ordinate their research efforts while parallel efforts are made in Semantic Web education to increase the availability of skilled young researchers and practitioners and last but not the least, in pushing the take-up in Business and Industry.

In order to accelerate the transfer from research to industry, the objective of an Industry-Research co-operation is to establish a working relationship between Semantic Web researchers and an industry partner, in which research results being produced in an area of Semantic Web research will be prototypically applied to the industry partner's selected business case. The co-operation not only seeks to achieve an individual success story in terms of some specific research and a given business case, but also to establish the value of Semantic Web technologies to industrial application in a more general sense. It achieves this by demonstrating the use of Semantic Web technology in a business setting, exploring their usefulness in solving business problems and ensuring future applicability by directing researchers towards meeting industrial requirements in their work.

In NoE-Knowledge Web, an Industry Board was formed at the beginning of the network to bring together potential early adopters of Semantic Web technologies from across a wide spread of industry sectors. Industry Board members have been involved in many initiatives of the Knowledge Web Industry Area, including the collection of business use cases and their evaluation. In order to more directly achieve close co-operation between researchers and industry, each research activity in the network was invited to select a use case whose requirements closely correlated to what would be achieved in their research work. Results have been collected and reported in July 2007[6].

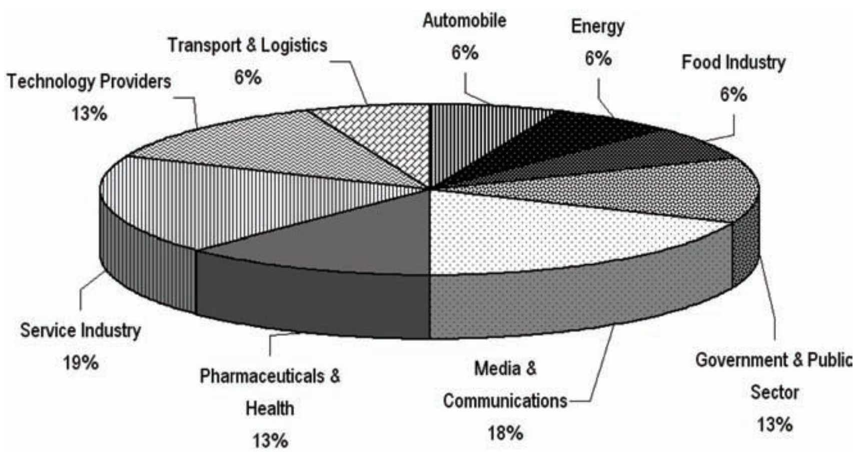
Currently in 2007, this Industry Board consisted of about 50 members (e.g., France Telecom, British Telecom, Institut Français du Pétrole, Illy Caffè, Trenitalia, Daimler AG, Thalès, EADS, …) from across 14 nations and 13 economic sectors (e.g., telecoms, energy, food, logistics, automotive,...).

The companies were requested to provide illustrative examples of actual or hypothetical deployment of Semantic Web technology in their business settings. This was followed up with face-to-face meetings between researchers and industry experts from the companies to gain additional information about the provided use cases. Thus, in 2004, a total of 16 use cases were collected from 12 companies. In 2007, through many workshops and Industry forum sessions at major Semantic Web conferences, more than a hundred use cases were available or illustrative of the current trend to introduce Semantic Web technology in the main stream.

As shown in Figure 1, where the use cases are broken down according to the industry sector, collected cases are from 9 industry sectors, with the highest number of the use cases coming from the service industry (19%) and media & communications (18%) respectively. This initial collection of use cases can be found in (Nixon L. et al., 2004), and an updated selection is available on the Knowledge Web Industry portal[7].

The co-operations have been a very challenging activity, given the early state of much cutting edge Semantic Web research and the differences in perspective between academia and business. However, successes have been reported, not only

Figure 1. Breakdown of use cases by industry sector



in the production of some prototypical solutions and demos which can be shown to industry and business top managers, but also in making researchers more aware of the importance of their work to solving business problems and the earlier recognition by academics of industry needs and expectations and so integrating them to their research agenda.

Hence, the Industry-Research co-operations in NoE-Knowledge Web and NoE-REWERSE must be seen as a significant first attempt to align the ambitious cutting edge work on Semantic Web technologies done by leading researchers in Europe and the real world business problems encountered by the European industry which may find a potential solution in those same Semantic Web technologies. Given a continued rise in awareness among Semantic Web researchers of the applicability of their work to industry and the continued rise in awareness among industry of the potential of the work of Semantic Web researchers, which has been begun in IST-NoEs, in IST-R&D projects, but also clearly in industry (SMEs and large companies), the technology transfer is gradually evolving.

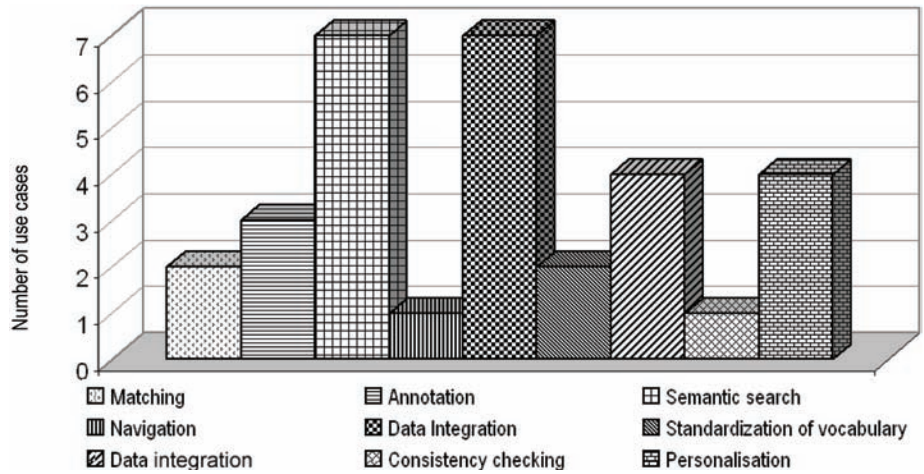
Use Case Analysis

A preliminary analysis of the use cases has been carried out in order to obtain a first vision (end of 2004) of the current industrial needs and to estimate the expectations from knowledge-based technology with respect to those needs. The industry experts were asked to indicate the existing legacy solutions in their use cases, the service functionalities they would be offered and the technological locks they encountered, and eventually how they expected that Semantic Web technology could resolve those locks. As a result, this analysis has provided an overview of:

- Types of business or service problems where the knowledge-based technology is considered to bring a plausible solution.
- Types of technological issues (and the corresponding research challenges) which knowledge based technology is expected to overcome.

Figure 2 shows a breakdown of the areas in which the industry experts thought Semantic Web technology could provide a solution. For example, for nearly half of the collected use cases, data integration and semantic search were areas

Figure 2. Breakdown of use cases by industry sector

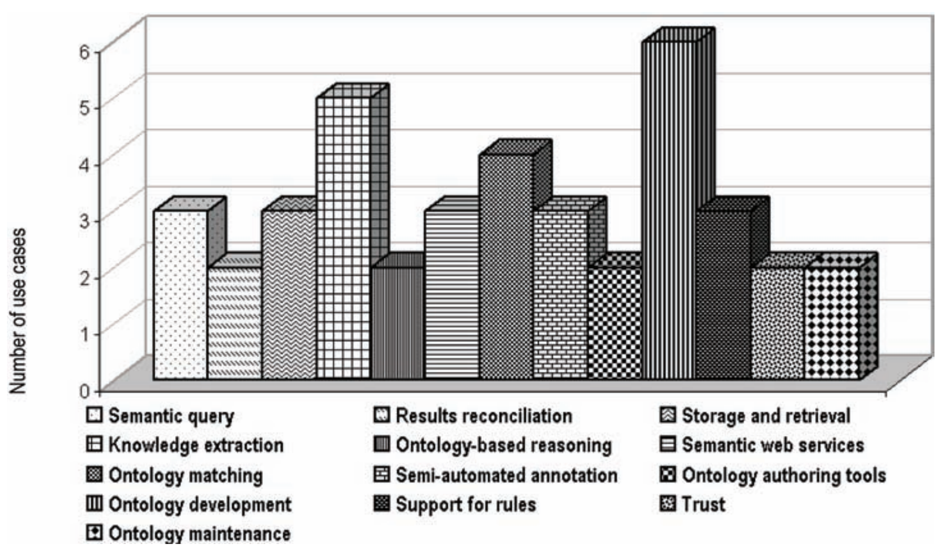


where industry was looking for knowledge-based solutions. Other areas mentioned, in a quarter of use cases, were solutions to data management and personalization.

Figure 3 shows a breakdown of the technology locks identified in the use cases. There are three technology locks which occur the most often in the collected use cases. These are: (1) ontology development, i.e., modeling of a business domain, authoring, reusing existing ontologies; (2) knowledge extraction, i.e., populating ontologies by extracting data from legacy systems; and (3) ontology mapping, i.e., resolving semantic heterogeneity among multiple ontologies.

Below, an illustration is given, with the help of a use case from the collection, how a concrete business problem can be used to indicate the technology locks for which knowledge-based solutions potentially might be useful. This use case addresses the problem of intelligent search of

Figure 3. Preliminary vision of technology locks in use case



documents in the corporate data of an international coffee company.

The company generates a large amount of internal data and its employees encounter difficulties in finding the data they need for the research and development of new solutions. The aim is to improve the quality of the document retrieval and to enable personalization services for individual users when searching or viewing the corporate data. As technology locks, the expert mentioned here the corporate domain ontology development and maintenance, and semantic querying.

Eventually, this analysis (by experts estimations) has provided with a preliminary understanding of scope of the current industrial needs and the current concrete technology locks where knowledge-based technology is expected to provide a plausible solution. However, to be able to answer specific industrial requirements, there is the need to conduct further a detailed technical analysis of the use cases, thereby associating to each technology lock a concrete knowledge processing task and a component realizing its functionalities.

Knowledge Processing Tasks and Components

Based on the information processing needs identified during the technical use cases analysis (Shvaiko P. et al., 2004), a typology of common knowledge processing tasks and a library of high level components for realizing those tasks, was built, see Table 1. The first tentative typology includes twelve processing tasks. Let us discuss knowledge processing tasks and components of Table 1 in more detail:

- *Ontology management, ontology merging and ontology manager*. These tasks and component are in charge of ontology maintenance (e.g., reorganizing taxonomies, resolving name conflicts, browsing ontologies, editing concepts) and merging multiple ontologies (e.g., by taking the union of the axioms) with respect to evolving business case requirements, see (Dou D. et al., 2005) (McGuiness D. et al., 2000) (Protégé[8]), OAEI-2007 Ontology Alignment Evaluation Initiative[9], NeOn[10] (Networked Evolving Ontologies) and Ontology Matching survey site[11].
- *Ontology matching, matching results analysis, producing explanations and match manager*. These tasks and component are in charge of (on-the-fly and semi-automatic) determination of semantic mappings between the entities of multiple schemas and ontologies, see (Rahm E. et al., 2001) (Shvaiko P. and Euzenat, 2005), (Euzenat J. and Shvaiko P., 2007). Mappings are typically specified with the help of a similarity relation which can be either in the form of a coefficient rating match quality in the (0,1] range (i.e., the higher the coefficient, the higher the similarity between the entities, see (Billig A. et al., 2002) (Ehrig M. et al., 2004) (Euzenat J. et al., 2004) (Do H. H. et al., 2002) (Zhong J. et al., 2002) or in the form of a logical relation (e.g., equivalence, subsumption), see (Giunchiglia F. et al., 2003) (Giunchiglia F. et al., 2004). The mappings might need to be ordered according to some criteria, see (Di Noia T. et al., 2003) (Do H. H. et al., 2002).

Finally, explanations of the mappings might be also required, see (Dhamankar R. et al., 2004) (Shvaiko P. et al., 2005). Matching systems may produce mappings that may not be intuitively obvious to human users. In order for users to trust the mappings (and thus use them), they need information about them. They need access to the sources that were used to determine semantic correspondences between terms and potentially they need to understand how deductions and manipulations are performed. The issue here is to present explanations in a simple and clear way to the user.

- *Data translation and wrapper*. This task and component is in charge of automatic manipulation (e.g., translation, exchange) of instances between heterogeneous information sources storing their data in different formats (e.g., RDF, SQL DDL, XML …), see (Hull R. 1997) (Petrini J. et al., 2004) (Velegrakis Y. et al., 2005) (Halevy A. et al., 2006). Here, mappings are taken as input (for example, from the match manager component) and are the support for generating query expressions that perform the required semantic and syntactical manipulations with data instances coming from heterogeneous environment.
- *Results reconciliation and results reconciler*. This task and component is in charge of determining an optimal solution, in terms of contents (no information duplication, etc.) and routing performance, for returning results from the queried information sources, see (Preguica N. et al., 2003).
- *Composition of Web services and planner*. This task and component is in charge of automated composition of Web services into executable processes (Orchestration). Composed Web services perform new functionalities by specific on demand interaction with pre-existing services that are published on the Web, see surveys from (Chan et al., 2007) (Berardi et al., 2005) (Hull et al., 2005) (Pistore et al., 2005) (Roman et al., 2005) (Traverso P. et al., 2004) (Cardoso et al., 2003) (McIlraith et al., 2001). From a business viewpoint, it remains a key challenge to be overcome, as the businesses react very positively to the need for a very effective integration technology and for more agility in a very competitive worldwide economy. In the meantime, reducing interoperability problems will open opportunities for easier innovative solutions and for the increase in cooperation between enterprises. This should result in re-combinations of businesses the technology provides and so will have a profound impact on business and economic workflows.
- *Content annotation and annotation manager*. This task and component is in charge of automatic production of metadata for the contents, see aceMedia[12] for multimedia annotation. Annotation manager takes as input the (pre-processed) contents and domain knowledge and produces as output a database of content annotations. In addition to the automatic production of content metadata, prompt mechanisms offer the user the possibility to enrich the content annotation by adding extra information (e.g., title, name of a location, title of an event, names of people) that could not be automatically detected.
- *Automated reasoning*. This task and component is in charge of providing logical reasoning services (e.g., subsumption, concept satisfiability, instance checking tests), see (Haarslev V. et al., 1999-2007). For example, when dealing with multimedia annotations, logical reasoning can be exploited in order to check consistency of the annotations against the set of spatial (e.g., left, right, above, adjacent, overlaps) and temporal (e.g., before, after, during, co-start, co-end) constraints. This can certify that the objects detected in the multimedia content correspond semantically to the concepts defined in the domain ontology. For example, in the racing domain, the automated reasoner should check whether a car is located on a road or whether the grass and sand are adjacent to the road.
- *Semantic query processing and query processor*. This task and component is in charge of rewriting a query posed by a human being or a machine, by using terms which are explicitly specified in the model of domain knowledge in order to provide

Table 1. Typology of knowledge processing tasks & components

N	Knowledge processing tasks	Components
1	Ontology Management	Ontology Manager
2	Ontology Matching	Match Manager
3	Ontology Matching results Analysis	Match Manager
4	Data Translation	Wrapper
5	Results Reconciliation	Results Reconciler
6	Composition of Web Services	Planner
7	Content Annotation	Annotation manager
8	Reasoning	Reasoner
9	Semantic Query Processing	Query Processor
10	Ontology Merging	Ontology Manager
11	Producing explanations	Match Manager
12	Personalization	Profiler

semantics preserving query answering, see (Mena E. et al., 1996) (Halevy et al., 2001) (Calvanese et al., 2002) (IST-IP aceMedia 2004). Examples of queries are "Give me all the games played on grass" or "Give me all the games of double players", in the tennis domain. Finally, users should be able to query by sample content e.g. an image. In this case, the system should perform an intelligent search of images and videos (e.g., by using semantic annotations) where, for example, the same event or type of activity takes place.

- *Personalization and profiler*. This task and component is in charge of tailoring services available from the system to the specificity of each user, see (Antoniou G. et al., 2004). For example, generation and updating of user profiles, recommendation generation, inferring user preferences, and so on. For example users might want to share annotations within trusted user networks, thus having services of personal metadata management and contacts recommendation. Also, a particular form of personalization, which is media adaptation, may require knowledge-based technology and consistent delivery of the contents to a broad range user terminals (e.g., PDA, mobile phone, portable PC).

KEY APPLICATION SECTORS AND TYPICAL TECHNOLOGY PROBLEMS

Healthcare and Biotechnologies

The medical domain is a favourite target for Semantic Web applications just as the expert system was for artificial intelligence applications 20 years ago. The medical domain is very complex: medical knowledge is difficult to represent in a computer format, making the sharing of information even more difficult. Semantic Web solutions become very promising in this context.

One of the main mechanisms of the Semantic Web - resource description using annotation principles - is of major importance in the medical informatics (or sometimes called bioinformatics) domain, especially as regards the sharing of these resources (e.g. medical knowledge on the Web or genomic database). Through the years, the IR area has been developed by medicine: medical thesauri are enormous (e.g., more than 1,600,000 terms in Unified Medical Language

System, UMLS[13]) and are principally used for bibliographic indexation. Nevertheless, the MeSh thesaurus (Medical Subject Heading) or UMLS have been used to provide data semantics with varying degrees of difficulty. Finally, the Web services technology allows us to imagine some solutions to the interoperability problem, which is substantial in medical informatics. Below, we will describe current research, results and expected perspectives in these biomedical informatics topics in the context of Semantic Web.

Biosciences Resources Sharing

In the functional genomics domain, it is necessary to have access to several databases and knowledge bases which are accessible separately on the Web but are heterogeneous in their structure as well as in their terminology. Among such resources, we can mention SWISSPROT[14] where the gene products are annotated by the Gene Ontology[15], Gen-Bank[16], etc. When comparing these resources it is easy to see that they propose the same information in different formats. The XML language, which acts as a common data structure for the different knowledge bases, provides at most a syntactic Document Type Definition (DTD) which does not resolve the semantic interoperability problem.

One of the solutions comes from the Semantic Web with a mediator approach (Wiederhold G., 1992) which allows for the accessing of different resources with an ontology used as the Interlingua pivot. For example and in another domain than that of genomics, the NEUROBASE project (Barillot C. et al., 2003) attempts to federate different neuro-imagery information bases situated in different clinical or research areas. The proposal consists of defining an architecture that allows the access to and the sharing of experimental results or data treatment methodologies. It would be possible to search in the various data bases for similar results or for images with peculiarities or to perform data mining analysis between several databases. The mediator of NEUROBASE has been tested on decision support systems in epilepsy surgery.

Ontologies for Coding Systems

The main usage of ontologies in medical domain is as index of coding system: after using thesauri for indexing medical bibliography (PubMed with the Mesh[17]), the goal is to index Electronic Health records with medical concept in order to enhance information retrieval or to allow epidemiological studies. For that purpose, several countries intend to use the SNOMED, an international classification of concepts organized in eight axes (Spackman et al., 2002). Except the problem of languages, this classification exists in two versions: a classification of 160,000 concepts (SNOMED-I V3.5) and an ontology, which is the evolution of the preceding one, of 330,000 concepts, SNOMED CT. The use of ontologies of such a size is difficult. Some authors describe them as *Reference Ontology* which cannot be accessed without an *interface ontology* (Rosenbloom et al., 2006). Notwithstanding, UK national health system (NHS) is integrating SNOMED CT and it will be interesting to examine the results of this industrial deployment[18].

Web Services for Interoperability

The Web services technology can propose some solutions to the interoperability problematic. We describe now a new approach based on a "patient envelope" and we conclude with the implementation of this envelope based on the Web services technology.

The patient envelope is a proposition of the Electronic Data Interchange for Healthcare group (EDI-Santé[19]) with an active contribution from the ETIAM[20] society. The objective of the work is on filling the gap between "free" communication, using standard and generic Internet tools, and "totally structured" communication as promoted by CEN (in the Working Group IV "Technology for Interoperability"[21]) or HL7[22]. After the worldwide analysis of existing standards, the proposal consists of an "intermediate" structure of information, related to one patient, and storing the minimum amount of data (i.e. exclusively useful data) to facilitate the interoperability between communicating peers. The "free" or the "structured" information is grouped into a folder and transmitted in a secure way over the existing communication networks (Cordonnier E. et al., 2003). This proposal has reached widespread adoption with the distribution by Cegetel.rss of a new medical messaging service, called "Sentinelle", fully supporting the patient envelope protocol and adapted tools.

After this milestone, EDI-Santé is promoting further developments based on ebXML and SOAP (Simple Object Access Protocol) in specifying exchange (see items 1 and 2 below) and medical (see items 3 and 4 below) properties:

1. *Separate what is mandatory* to the transport and management of the message (e.g., patient identification from what constitutes the "job" part of the message.
2. *Provide a "container" for the message*, collecting the different elements, texts, pictures, videos, etc.
3. *Consider the patient as the unique object of the transaction.* Such an exchange cannot be anonymous. It concerns a sender and an addressee who are involved in the exchange and who are responsible. A patient can demand to know the content of the exchange in which (s)he is the object, which implies a data structure which is unique in the form of a triple {sender, addressee, patient}.
4. *The conservation of the exchange semantics.* The information about a patient is multiple in the sense that it comes from multiple sources and has multiple forms and supporting data (e.g., database, free textual document, semi-structured textual document, pictures). It can be fundamental to maintain the existing links between elements, to transmit them together, e.g., a scanner and the associated report, and to be able to prove it.

The interest of such an approach is that it prepares the evolution of the transmitted document from a free form document (from proprietary ones to normalized ones as XML) to elements respecting HL7v3 or EHRCOM data types. In France, the GIP-DMP[23] retains such an approach (in conjunction with the Clinical Document Architecture of HL7[24]) for the implementation of the exchanges of the *Dossier Médical Personnel* (a future national electronic health record).

What is Next in the Healthcare Domain?

These different projects and applications highlight the main consequence of the Semantic Web being expected by the medical communities: the sharing and integration of heterogeneous information or knowledge. The answers to the different issues are the use of mediators, a knowledge-based system, and ontologies, which should be based in the mid term on normalized languages such as RDF, OWL but also in addition to come OWL-S, SAWSDL, WSML, SWRL, or RuleML. The work of the Semantic Web community must take into account these expectations, see for example the FP6 projects[25,26,27]. Finally, it is interesting to note that the Semantic Web is an integrated vision of the medical community's problems (thesauri, ontologies, indexation, and inference) and provides a real opportunity to synthesize and reactivate some research directions (Charlet J. et al., 2002).

Knowledge Management

Leveraging Knowledge Assets in Companies

Knowledge is one of the key success factors for enterprises, both today and in the future. Therefore, company knowledge management (KM) has been identified as a strategic tool. However, if for KM, information technology is one of the foundational elements, KM in turn, is also interdisciplinary by its nature. In particular, it includes human resource management as well as enterprise organization and culture[28]. KM is viewed as the management of the knowledge arising from business activities, aiming at leveraging both the use and the creation of that knowledge for two main objectives: capitalization of corporate knowledge and durable innovation fully aligned with the strategic objectives of the organization.

Conscious of this key factor of productivity in an ever faster changing ecosystem, the European KM Framework (CEN/ISSS[29], KnowledgeBoard[30]) has been designed to support a common European understanding of KM, to show the value of this emerging approach and to help organizations towards its successful implementation. The Framework is based on empirical research and practical experiences in this field from all over Europe and the rest of the world. The European KM Framework addresses all of the relevant elements of a KM solution and serves as a reference basis for all types of organizations, which aim to improve their performance by handling knowledge in a better way.

Benefits of Knowledge-Based KM

The knowledge backbone is made up of ontologies that define a shared conceptualization of an application domain and provide the basis for defining metadata that have precisely defined semantics, and are therefore machine-interpretable. Although the first KM approaches and solutions have shown the benefits of ontologies and related methods, a large number of open research issues still exist that have to be addressed in order to make Semantic Web technology a complete success for KM solutions:

- Industrial KM applications *have to avoid any kind of overhead as far as possible. A seamless integration of knowledge creation* (i.e., content and metadata specification) and knowledge access (i.e., querying or browsing) into the working environment is required. Strategies and methods are needed to support the creation of knowledge, as side effects of activities that are carried out anyway. These requirements mean emergent semantics that can be supported through ontology learning, which should reduce the current time consuming task of building-up and maintaining ontologies.
- *Access to as well as presentation of knowledge has to be context-dependent*. Since the context is setup by the current business task, and thus by the business process being handled, a tight integration of business process management and knowledge management is required. KM approaches can provide a promising starting point for smart push services that will proactively deliver relevant knowledge for carrying out the task at hand more effectively.
- *Conceptualization has to be supplemented by personalization*. On the one hand, taking into account the experience of the user and his/her personal needs is a prerequisite in order to avoid information overload, and on the other hand, to deliver knowledge at the right level of granularity and from the right perspective at the right time.

The development of knowledge portals serving the needs of companies or communities is still a manual process. Ontologies and related metadata provide a promising conceptual basis for generating parts of such knowledge portals. Obviously, among others, conceptual models of the domain, of the users and of the tasks are needed. The *generation of knowledge portals* has to be supplemented with the (semi-) automated evolution of portals. As business environments and strategies change rather rapidly, *KM portals have to be kept up-to-date in this fast changing environment.* Evolution of portals should also include some mechanisms *to 'forget' outdated knowledge.*

KM solutions will be based on a combination of intranet-based functionalities and mobile functionalities in the very near future. Semantic technologies are a promising approach to meet the needs of mobile environments, like location-aware personalization and adaptation of the presentation to the specific needs of mobile devices, i.e., the presentation of the required information at an appropriate level of granularity. In essence, employees should have access to the KM application *anywhere* and *anytime*.

Peer-to-peer computing (P2P), social networking (W2.0), combined with Semantic Web technology, will be a strong move towards getting rid of the more centralized KM approaches that are currently used in ontology-based solutions. W2.0 scenarios open up the way to derive consensual conceptualizations among employees within an enterprise in a bottom-up manner.

Virtual organizations are becoming more and more important in business scenarios, mainly due to decentralization and globalization. Obviously, semantic interoperability between different knowledge sources, as well as trust, is necessary in inter-organizational KM applications.

The integration of KM applications with *e-learning* (e.g., skill management in companies) is an important field that enables a lot of synergy between these two areas. KM solutions and e-learning must be integrated from both an organizational and an IT point of view. Clearly, interoperability and integration of (metadata) standards are needed to realize such integration.

Knowledge management is obviously a very promising area for exploiting Semantic Web technology. Document-based portals KM solutions have already reached their limits, whereas semantic technology opens the way to meet KM requirements in the future.

Knowledge-Based KM Applications

In the context of geographical team dispersion, multilingualism and business unit autonomy, usually a company wants a solution allowing for the identification of strategic information, the secured distribution of this information and the creation of transverse working groups. Some applicative solutions allowed for the deployment of an Intranet intended for all the marketing departments of the company worldwide, allowing for a better division of and a greater accessibility to information, but also capitalisation on the total knowledge. There are four crucial points that aim at easing the work of the various marketing teams in a company: (1) Business intelligence, (2) Skill and team management[31], (3) Process management[32] and (4) Rich document access and management[33].

Thus, a system connects the "strategic ontologies" of the company group (brands, competitors, geographical areas, etc.) with the users, via the automation of related processes (research, classification, distribution, knowledge representation). The result is a dynamic semantic system of navigation (research, classification) and collaborative features.

At the end from a functional point of view, a KM system organises skill and knowledge management within a company in order to improve interactivity, collaboration and information sharing. This constitutes a virtual workspace which facilitates work between employees that speak different languages, automates the creation of work groups, organises and capitalises structured and unstructured, explicit or tacit data of the company, and offers advanced features of capitalisation

(Bonifacio M. et al., 2005) (Brunschweig B. et al., 2005) (Nordheim D. et al., 2005).

Eventually, the semantic backbone makes possible to cross a qualitative gap by providing cross-lingual data.

E-Commerce and E-Business

Electronic commerce is mainly based on the exchange of information between involved stakeholders using a telecommunication infrastructure. There are two main scenarios: business-to-customer (B2C) and business-to-business (B2B).

B2C applications enable service providers to promote their offers, and for customers to find offers which match their demands. By providing unified access to a large collection of frequently updated offers and customers, an electronic marketplace can match the demand and supply processes within a commercial mediation environment.

B2B applications have a long history of using electronic messaging to exchange information related to services previously agreed among two or more businesses. Early plain-text telex communication systems were followed by electronic data interchange (EDI) systems based on terse, highly codified, well structured, messages. A new generation of B2B systems is being developed under the ebXML (electronic business in XML) heading. These will use classification schemes to identify the context in which messages have been, or should be, exchanged. They will also introduce new techniques for the formal recording of business processes, and for the linking of business processes through the exchange of well-structured business messages. ebXML will also develop techniques that will allow businesses to identify new suppliers through the use of registries that allow users to identify which services a supplier can offer. ebXML needs to include well managed multilingual ontologies that can be used to help users to match needs expressed in their own language with those expressed in the service providers language(s) see (Guarino N. 1999) (Zyl J. et al., 200) (Lehtola A. et al., 2003) (Heinecke J. et al., 2003) (Benatallah B et al., 2005).

Knowledge-Based E-Commerce and E-Business Value

At present, ontology and more generally knowledge-based systems appear as a central issue for the development of efficient and profitable e-commerce and e-business solutions. However, because of the actual situation i.e. the partial standardization of business models, processes, and knowledge architectures, it is currently difficult for companies to achieve the promised ROI from knowledge-based e-commerce and e-business.

Moreover, a technical barrier exists that is delaying the emergence of e-commerce, lying in the need for applications to *meaningfully share information*, taking into account the lack of reliability, security and eventually trust in the Internet. This fact may be explained by the variety of e-commerce and e-business systems employed by businesses and the various ways these systems are configured and used. As an important remark, such *interoperability problems* become particularly severe when a large number of trading partners attempt to agree and define the standards for interoperation, which is precisely a main condition for maximizing the ROI indicator.

Although it is useful to strive for the adoption of a single common domain-specific standard for content and transactions, such a task is often difficult to achieve, particularly in cross-industry initiatives, where companies co-operate and compete with one another. Some examples of the difficulties are:

- *Commercial practices* may vary widely, and consequently, cannot always be aligned for a variety of technical, practical, organizational and political reasons.

- *The complexity of a global description* of the organizations themselves: their products and services (independently or in combination), and the interactions between them remain a formidable task.
- It is not always possible to establish *a priori rules* (technical or procedural) governing participation in an electronic marketplace.

Adoption of a single common standard may limit business models which could be adopted by trading partners, and therefore, potentially reduce their ability to fully participate in e-commerce.

A knowledge-based approach has the potential to significantly accelerate the penetration of electronic commerce within vertical industry sectors, by *enabling interoperability at the business leve*l. This will enable services to adapt to the rapidly changing business ecosystem.

Knowledge-Based E-Commerce and E-Business Applications

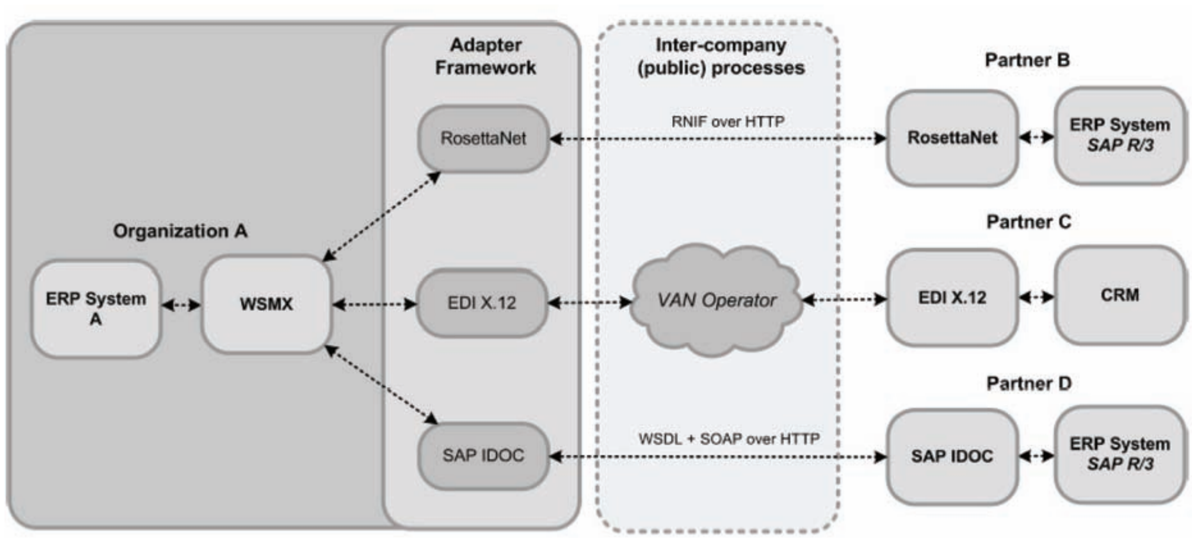
The Semantic Web brings opportunities to industry to create new services[34], extend markets, and even develop new businesses since it enables the inherent meaning of the data available in the Internet to be accessible to systems and devices able to interpret and reason at the knowledge level. This in turn leads to new revenue opportunities, since information becomes more readily accessible and usable. For example, a catering company whose Web site simply lists the menus available is less likely to achieve orders compared to one whose menus are associated with related metadata about the contents of the dishes, their origin (e.g., organic, non-genetically modified, made with local produce), links to alternative dishes for special diets, personalised ordering where a user profile can be established which automatically proposes certain menu combinations depending on the occasion (e.g., wedding banquet, business lunch). The latter case assumes that both provider-side knowledge generation and knowledge management tools are available, such that the asset owner can readily enhance their data with semantic meaning, and client-side tools are available to enable machine interpretation of the semantic descriptions related to the products being offered, such that the end user can benefit from the available and mined knowledge. Examples of some possible application areas were studied by the Agent Cities project[35].

In the e-business area Semantic Web technology can improve standard business process management tools. One prototypical case is in the area of logistics. The application of knowledge technology on top of today's business management tools enables the automation of major tasks of business process management[36] see (Semantic Web Case Studies for eBusiness 2005).

In one of the Knowledge Web Industry-Research co-operations, a number of scenarios within the **B2B integration scenario** were identified, involving data mediation, discovery, and composition of services. All of these use cases have been evaluated according to a community-agreed methodology defined by the SWS challenge methodology with satisfying success levels defined by the methodology. This is an important step when proving the added value of the Semantic Web service technology applied to B2B integration domain. In addition, the standardization process has been partially finalized within the OASIS Semantic Execution Environment Technical Committee (OASIS SEE TC) and W3C Semantic Annotations for WSDL and XML Schema (W3C SAWSDL WG). However, the standardization process in both groups is still ongoing, but under business pressure has concluded respectively on SAWSDL in 2007 and SESA framework early 2008.

The Industry-Research co-operation has *demonstrably solved a business case from the B2B domain*. We have shown how the technology deals with requirements from B2B domain and how this technology reacts to changes in back-end systems which might occur over the system's lifetime.

Figure 4. Semantic Web services integration in B2B



The research is not yet ready for industry. It must be shown how the technology is layered on the existing infrastructure and how it interacts with existing systems. For this purpose some parts of the technology need to be standardized (such as grounding mechanisms built on SAWSDL or the architecture). In particular, the grounding mechanism built on SAWSDL provides a "common interface" between semantic descriptions and non-semantic descriptions (in our case WSDL). The standardization is still ongoing while at the same time, the alignment of service semantics with this grounding mechanism must be further finalised. While it has been demonstrated how this is possible to be done and what the added value of this approach is, *the complexity of business standards still needs to be addressed.*

In addition, a prototype is available[37] and has been provided to NoE-Knowledge Web industry partners see Figure 4.The following scenarios have been realised as part of the Semantic Web Services Challenge:

- **Mediation Scenario** (http://sws-challenge.org/wiki/index.php/Workshop_Budva). Addressing the mediation scenario for B2B integration when proprietary back-end systems of one company needed to be integrated with a partner using RosettaNet standard. Whole scenario has been successfully addressed.
- **Discovery Scenario** (http://sws-challenge.org/wiki/index.php/Workshop_Athens). Addressing discovery scenario when a supplier needed to be discovered and selected from suitable ones. Whole scenario has been successfully addressed.
- **Composition Scenario** (http://sws-challenge.org/wiki/index.php/Workshop_Innsbruck). Addressing composition scenario when more services can satisfy the user need. Whole scenario has been successfully addressed.

Work will continue and the co-operation plans to address additional scenarios of the SWS challenge, namely scenarios when services can be filtered based on non-functional properties (QoS, financial, etc.). In addition, a tutorial was given on SWS in the context of business process management at ICWS'07 conference, and the authors co-organize the workshop on service composition and SWS challenge held at the Web Intelligence conference[38] (Vitvar T. et al., 2007a) (Vitvar T. et al., 2007b) (Hasselwanter T. et al., 2007).

Multimedia and Audiovisual Services

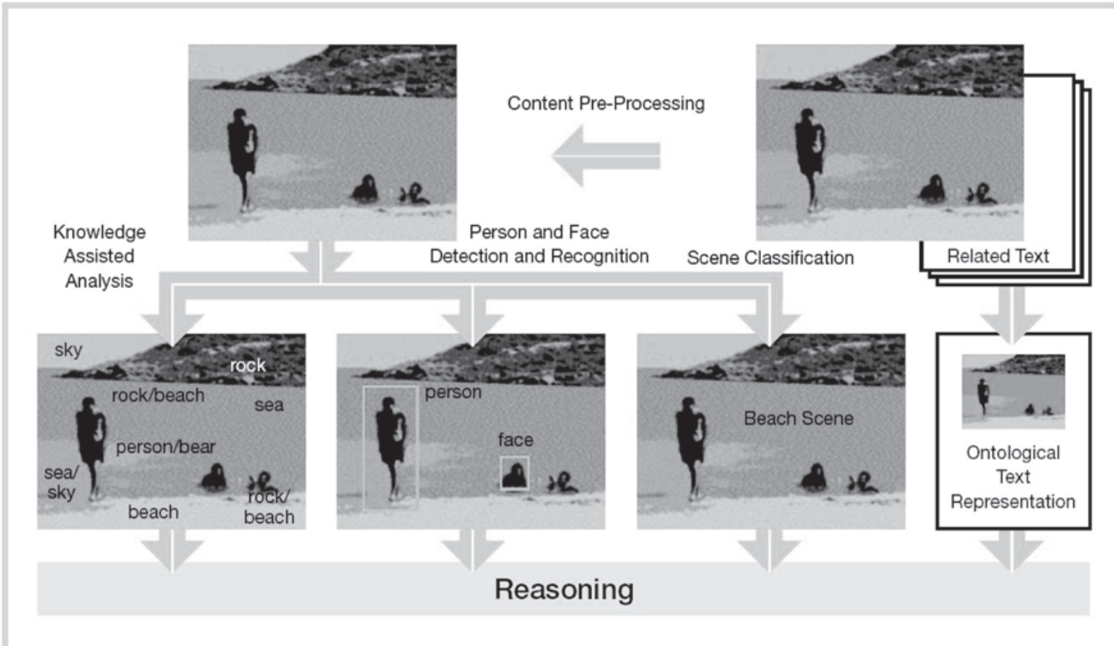
Practical realisation of the Semantic Web vision is actively being researched by a number of experts, some of them within European collaborative projects, and others within company specific initiatives. Earlier projects such as SEKT[39] and DIP, mainly focused on enhancing text based applications from a knowledge engineering perspective. Although significant benefits in unlocking access to valuable knowledge assets are realised via these projects, in various domains such as digital libraries, enterprise applications, and financial services, it was soon recognised that there was a challenging and potentially highly profitable area of research into the integration of multimedia and Semantic Web technologies for multimedia content based applications. Projects such as aceMedia, BOEMIE, and MESH are examples of initiatives aiming to advance the use of semantics and reasoning for improved multimedia applications such as automated annotation, content summarisation, and personalised content syndication.

The drive for application of semantic technologies in the multimedia and content domains comes from a proliferation of audiovisual devices and services which have led to an exponential growth in available content. Users express dissatisfaction at not being able to find what they want, and content owners are unable to make full use of their assets. Service providers seek means to differentiate their offerings by making them more targeted toward the individual needs of their customers. Semantic Web technology can address these issues. It has the potential to reduce complexity, enhance choice, and put the user at the center of the application or service, and with today's fast mobile data services and availability of wifi, such benefits can be enjoyed by consumers and professional users in all environments using all their personal devices, in the home, at work, in the car and on the go.

Semantic Web technologies can enhance multimedia based products to increase the value of multimedia assets such as content items which are themselves the articles for sale (songs, music videos, sports clips, news summaries, etc) or where they are used as supporting sales of other goods (e.g. promotional images, movie trailers etc). Value is added in search applications, such that returned items more closely match the user's context, interests, tasks, preference history etc, as well as in proactive push applications such as personalised content delivery and recommendation systems, and even personalised advertising. However, applications such as content personalisation, where a system matches available content to the user's stated and learned preferences, thereby enabling content offerings to be closely targeted to the user's wishes, rely on the availability of semantic metadata describing the content in order to make the match. Currently, metadata generation is mostly manual, which is costly and time consuming. Multimedia analysis techniques which go beyond the signal level approach to a semantic analysis have the potential to create automatic annotation of content, thereby opening up new applications which can unlock the commercial value of content archives (Stamou et al., 2006) (Stamou et al., 2005).

Automated multimedia analysis tools are important enablers in making a wider range of information more accessible to intelligent search engines, real-time personalisation tools, and user-friendly content delivery systems. Such automated multimedia analysis tools, which add the semantic information to the content, are critical in realising the value of commercial assets e.g. sports, music and film clip services, where manual annotation of multimedia content would not be economically viable, and are also applicable to users' personal content (e.g. acquired from video camera or mobile phone) where the user does not have time, or a suitable user interface, to annotate all their content.

Figure 5. Automated semantic annotation in aceMedia



Multimedia ontologies are needed to structure and make accessible the knowledge inherent in the multimedia content, and reasoning tools are needed to assist with identification of relevant content in an automated fashion. Although textual analysis and reasoning tools have been well researched, and despite the projects funded by the European Commission in the 6th framework, fewer tools are available for semantic multimedia analysis, since the problem domain is very challenging. However, automated multimedia content analysis tools such as those being studied within aceMedia[40] are a first step in making a wider range of information more accessible to intelligent search engines, real-time personalisation tools, and user-friendly content delivery systems.

Furthermore, interoperability of multimedia tools is important in enabling a wide variety of applications and services on multiple platforms for diverse domains. The W3C Multimedia Semantics Incubator Group reported on interoperability issues[41] and it is clear that a common framework using Semantic Web languages tools is essential for full exploitation of the potential of multimedia assets. Interoperability is essential in achieving commercial success with semantic multimedia applications, since it enables multiple manufacturers, content providers and service providers to participate in the market. This in turn enables consumer confidence to be achieved, and a viable ecosystem to be developed.

Knowledge Enhanced Multimedia Services

In aceMedia the main technological objectives are to discover and exploit knowledge inherent in multimedia content in order to make content more relevant to the user; to automate annotation at all levels (see Figure 5) ; and to add functionality to ease content creation, transmission, search, access, consumption and re-use.

Users access multimedia content using a variety of devices, such as mobile phones and set-top-boxes, as well as via broadband cable or wireless to their PC. Through exploitation of Semantic Web tools, aceMedia has created a system which provides new and intuitive ways for users to enjoy multimedia content, such as intelligent search and retrieval, self-organising content, and self-adapting content. For example, using aceMedia's automatic metadata generation, a user can annotate content taken with her mobile phone, then seamlessly upload it to her PC where further automatic metadata generation takes place. aceMedia tools enables the content to be automatically organised into thematic categories, according to the user's preferences, and using extensions to DLNA/UPnP (networked digital home) standards, the content can be automatically pushed to other users (as specified by the content owner) according to chosen rules. For example, our user might automatically receive new pictures of herself on her mobile phone or PC which were acquired and annotated on the device of one of her friends or family.

The aceMedia use case highlighted a number of future direction, issues and new challenges with respect to semantic multimedia content analysis and manipulation within a Semantic Web framework. Apart from the requirements with respect to formal uncertainty representations and more effective reasoning and management tools support, two dimensions of significant importance include:

- *Cross-media analysis*, where additional requirements are posed due to the multimodality of knowledge considered, and their semantic modelling and integration, and
- *Non-standard approaches to reasoning*, as purely deductive reasoning alone proves not sufficient

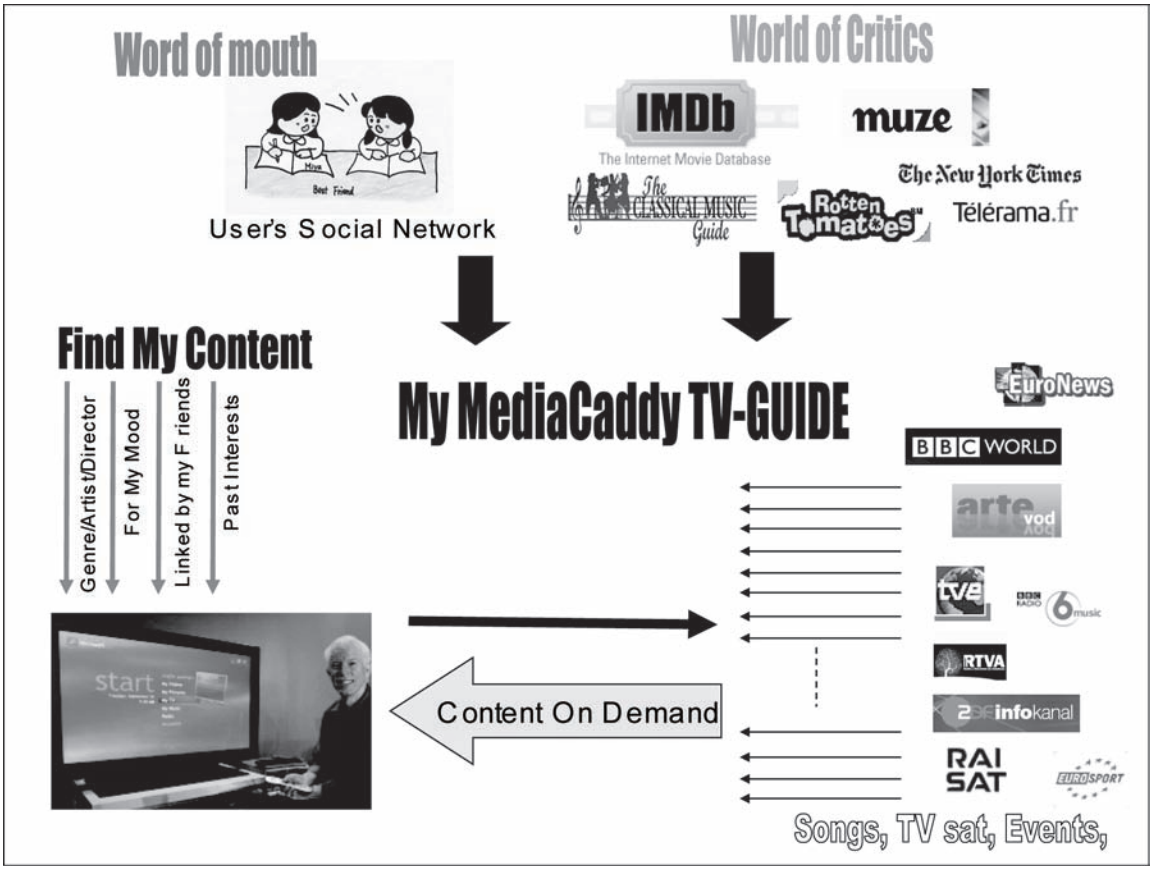
Other projects which can use the results of this co-operation: particularly K-Space[42], X-Media[43], BOEMIE[44] and MESH[45] constitute research consortiums working on the same topic. As, in the case of aceMedia, the main research directions focus on the exploitation of formal explicit knowledge and (possibly extended) inference services for the extraction of semantic descriptions from multimedia content. Additional aspects include among other scalability, logic programming and DL-based reasoning integration for non-standard inference support, and ontology evolution (Stoilos G. et al, 2005) (Petridis K. et al., 2006) (Dasiopoulou S. et al., 2007).

Leveraging Social Network Knowledge for Movie Recommendations

Another interesting reported multimedia experiment is MediaCaddy (Garg S. et al., 2005) aiming at providing *movie or music recommendations* based on published online *critics*, *user experience and social networks*. Indeed, for the entertainment industry, traditional approaches to delivering meta-content about movies, music, TV shows, etc. were through reviews and articles that were done and published in traditional media such as newspapers, magazines and TV shows. With the introduction of the Internet, non-traditional forms of delivering entertainment started surfacing. The third quarter of 2003 in the U.S was the best ever for broadband penetration bringing such services as content on-demand and mobile multimedia. As of today more than 5000 movies and 2,500,000 songs are available on line. In the next couple of years this figure is expected to grow in leaps and bounds. With such a phenomenal rise in content over IP, a new need for secondary metacontent related to the movies/music emerged. Initially this was through movie reviews or music reviews published on Web portals such as Yahoo, MSN and online magazine portals as well as entertainment sales sites such as Netflix.com and Amazon.com.

Most consumers today get information about media content primarily from reviews/articles in entertainment/news magazines, their social network of friends (one user recommends a song or movie to a friend), acquaintances and advertisements. In most of the cases, one or all

Figure 6. Conceptual model of content navigation system from the MediaCaddy project



of the above influence user's opinion about any content (s)he chooses to consume. In addition, a new breed of customizable meta-content portal has emerged, which specifically targets the entertainment industry. Examples of such portals include Rotten Tomatoes[46] and IMDB[47]. However, these services today are typically accessed via Web portals thereby limiting the interactions and access to the information for a user in a non-PC environment.

MediaCaddy is a recommendation and aggregation service built around a self-learning engine, which analyzes a click stream generated by user's interaction and actions with meta-content displayed through a UI. This meta-content (Music /Movies/ TV reviews/ article/ synopsis/ production notes) is accessed from multiple Internet sources and structured as an ontology using a semantic inferencing platform. Figure 6 illustrates the conceptual model of MediaCaddy

This provides multiple benefits, both allowing for a uniform mechanism for aggregating disparate sources of content, and on the other hand, also allowing for complex queries to be executed in a timely and accurate manner. The platform allows this information to be accessed via Web Services APIs, making integration simpler with multiple devices and UI formats. Another feature that sets MediaCaddy apart is its ability to achieve a high level of personalization by analyzing content consumption behaviour in the user's personal Movie/Music Domain and his or her social network and using this information to generate music and movie recommendations.

Prominent Applications

Finally we list some excellent illustrations of the applications of Semantic Web technology, as selected from a worldwide competition[48] which offers participants the opportunity to show the best of the art.

- Multimedia e-culture demonstrator, is to demonstrate how novel semantic-Web and presentation technologies can be deployed to provide better indexing and search support within large virtual collections of cultural heritage resources, 1st Prize 2006, http://e-culture.multimedian.nl/demo/search
- CONFOTO, Essen, Germany. CONFOTO is an online service which facilitates browsing, annotating and re-purposing of photo, conference, and people descriptions. 1st Prize 2005: http://www.confoto.org/
- FungalWeb, Concordia University, Canada. "Ontology, the Semantic Web, Intelligent Systems for Fungal Genomics".2nd Prize 2005: http://www.cs.concordia.ca/FungalWeb/
- Bibster – A semantics-based Bibliographic P2P system. http://bibster.semanticweb.org
- CS AKTive space – Semantic data integration. http://cs.aktivespace.org (Winner 2003 Semantic Web challenge)
- Flink: SemWeb for analysis of Social Networks. http://www.cs.vu.nl/~pmika (Winner 2004 Semantic Web challenge)
- Museum Finland: Sem. Web for cultural portal. http://museosuomi.cs.helsinki.fi (2nd prize 2004 Semantic Web challenge)
- Also see Applications and Demos at W3C SWG BPD. http://esw.w3.org/mt/esw/archives/cat_applications_and_demos.html

CONCLUSION AND FUTURE TRENDS

In 2000, three prominent authors in the Semantic Web activity expounded in a seminal Scientific American paper (Berners-Lee T. et al., 2001) the Semantic Web vision. In the time since then, the Semantic Web has become real. Currently, there are hundreds of millions of RDF triples, on tens of thousands of Web pages, and thousands of ontology pages have been published using RDF schema and OWL, with a growing level of industrial support. This very active support from industry is witnessed at worldwide key conference[49] very focused on the applications of the Semantic Web Technology. Indeed, about 100 talks on industry experience in testing and deploying the technology and about 30 technology showcases and 10 workshops or tutorials were actively followed by hundreds of attendees (300 at STC 05, 700 at STC 06, 730 at STC 07 and 210 at the 1st ESTC 2007) mostly from the industry.

However, the Semantic Web is still in its early days and there are many exciting innovations on the horizon.

A keynote speech[50] foresaw (Hendler J. & Lassila O., 2006) a "re-birth of AI" (or the end of the AI-winter) thanks to big-AI applications (Deep Blue, Mars Rover, Deep Space 1, Sachem-Usinor) and Web AI (IR, NLP, Machine Learning, Services, Grid, Agents, social nets) needed due to the tremendous amount of data continuously available on the Web and the emergence of new ways of doing things (loose coupling of distributed applications or services, new business models, etc.).

From 2000 to 2007, three major endeavours have paved the way for the future: DARPA, W3C and EU IST where DARPA and EU IST funded projects particularly were clearly forces towards production of recommendations to W3C (RDF-S, OWL, Rules, …), for fast adoption in industry. In the meantime, 2003 saw early government adoption and emerging corporate interest, in 2005 the emergence of commercial tools, lots of open source software and even good progress in the problem of scalability (tractable reasoning over 10 million triples has already been claimed by Oracle[51]).So, *significant corporate activity is clearly noticeable today compared to 7 years ago*:

- Semantic (Web) technology companies are starting and growing: Cerebra, Siderean, SandPiper, SiberLogic, Ontology Works, Intellidimension, Intellisophic, TopQuadrant, Data Grid, Software AG, OntoText, Autonomy, FAST, Exalead, iSoco, Jouve, Mondeca, Sirma, Pertim, Biovista, etc.
- Semantic Web technology appears in the strategic plans of large corporations: Nokia, SAP AG, IBM, HP, Adobe, Cisco, Oracle, Sun, Vodaphone, Renault, AGFA, Cable and Wireless, Novartis, JP Morgan Chase Bank, Wells Fargo, Swiss Life, Boeing, Audi, Elsevier etc.
- Outreach to industry is also demonstrated through a newly launched W3C initiative (2007): "Semantic Web Education and Outreach Interest Group - Case Studies and Use Cases". Case studies include descriptions of systems that have been deployed within an organization, and are now being used within a production environment[52].
- Government projects in and across agencies: US, EU, Japan, Korea, China, FAO, etc.
- Life sciences/pharma is an increasingly important market, e.g. the Health Care and Life Sciences Interest Group at W3C[53]
- Many open source tools are available: Kowari, RDFLib, Jena, Sesame, Protégé, SWOOP, Wilbur etc. see the W3C SWAD initiative[54]
- Related technologies are taking off: Google Base (taxonomies for resource descriptions), Web 2.0 initiatives for mash-up applications, etc.
- Enterprise Web 2.0 can be the catalyst for a more collaborative business environment[55]. The BBC World Service had done a lot of work to try to create a more collaborative work environment. As it turned out, the BBC's internal forums, which only cost the company about 200 pounds, got the company to be more collaborative than the more formal initiatives did.

Then, it is also witnessed that adding a few semantics to current Web applications - meaning "not harnessing the full picture at once but step by step" – gives a significant push in applications: richer metadata, data harvesting and visualization, Web-based social network, digital asset management, scientific portals, tools for developers, and so gradually closing the semantic gap.

What has been Learned from AI?

- Cross-breeding with AI succeeded, standalone AI did not!
- Tools are hard to sell (needed too much skill and education)
- Reasoners are a means, not an end (a key component but not the end)
- Knowledge engineering bottleneck (Ontology development and management)

What has been Learned from the Web?

- The magic word: Distribute, interconnect and Share Roadmap!
- PC era [1980-1990] – autonomous computing and Ethernet
- Internet 1st generation [1990-2000] - Web 1.0), "read-only Web", Web sites and Companies' portals
- Social Networks [2000-2010] - Web 2.0, corporate Knowledge Management and social nets

- Semantic Web [2007 – 2020] - Web 3.0 – merging social net with automated Semantic Web
- Web OS [2020-2030] - Web 4.0

However, it must be clear that there are still **key technology locks** identified today that needs academic research and R&D investments for a full uptake of the technology (Cuel et al., 2007):

- **Ontology and reasoning:**

° **The development of medium size to large ontologies is a challenging task:** e.g. modelling of business domains, unified and industry-grade methodology, best practices and guidelines, re-use of existing ontologies and simple tools to use.

° **Automated update of ontologies and knowledge bases:** e.g. ontology maintenance by extraction of new knowledge facts and concept reasoning (abduction, learning), knowledge base population from legacy databases, data warehouse and data on the Web, consistency checking.

° **Ontologies interoperability:** Overcome inevitable heterogeneity in spite of KR standards via e.g. automated mapping (concepts, properties, rules, graphs, …) in particular in the open context of the Web and the social Web (Web 2.0).

° **Scalability:** Be capable to process business and real applications needs e.g. approximate reasoning, reasoning on inconsistent ontologies, ontology and reasoning modularization, distribution and cooperation.

° **KR expressivity and tractability trade-off:** Maintaining the just needed KR expressivity to avoid tractability and decidability issues, there are many open problems in this area. Look for reasoning algorithm optimizations (!), measure experimental complexity and lastly may be relax the completeness property.

° **Rules - Logic Programming and KR:** Moving towards a deeper and broader automation of business process intra- and inter-enterprise require the addition of semantic rules technology. e.g. Rules in communicating applications, Rules to describe / represent service process models, Rules for policies and contracting, etc. (see e.g. RuleML W3C[56])

- **Semantic Web Services and sevices oriented computing** (Papazoglou et al., 2006):

° **Discovery:** Automated service discovery, reasoning on non functional parameters like QoS and cost.

° **Composition:** Business and industrial processes automated. I/O signature and behavioural composition (Inputs, Outputs, pre-conditions, effects and message protocols). Robustness in a versatile and inconsistent context. Composition driven by high level business needs (Business Rules).

° **Management:** Web services supervision, self-healing service, self-optimizing service, self-configuring service, self-protecting service.

° **Design tools:** Unified design principles for engineering service applications, associating a service design methodology with standard software development and business process modelling techniques, service governance, test and proof checks.

° **Pilots and standard platforms**. The most prominent (2007):

WSMX[57] (Fensel et al., 2005) probably the most complete architecture to date, experimented on business cases and in transfer to OASIS

SAWSDL[58]: some running prototypes, industrial pilots and transfer to W3C (Sivashanmugam, 2003) (METEOR-S[59])

OWL-S[60] (Ankolenkar, 2004) (OWL-S MX[61])

SWSF[62]

In summary, the Semantic Web is "an interoperability technology", "a architecture for interconnected communities and vocabularies" and "a set of interoperable standards for knowledge exchange"[63]. Firstly, layman users facing the unmanageable growth of data and information, and secondly companies facing the huge amounts and volatility of data, applications and services, all require urgently automated means that master this growing complexity. In such de-facto context, no individual is able to identify knowledge patterns in their heads, no company (and employees!) is able to always shorter its products and service lifecycle and self adapt rapidly enough to survive.

The performance of semantic technologies clearly shows efficiency gain, effectiveness gain and strategic edge. Those facts are based on a survey[64] of about 200 business entities engaged in semantic technology R&D for development of products and services to deliver solutions and also recently witnesses at the ESTC 2007 industry oriented major event. From an academic and technological viewpoint, most things that have been predicted have happened - the semantic chasm is closing. Some things happened faster than anticipated like – triple store scaling, cooperation tools, enhanced SOA middleware, meta-data tagging tools, semantically-aware development environments and last but not the least, the unpredicted huge rise of Web 2.0 user-oriented technology[65] – and others still need to be realized: ontologies are there (but very little interlinking and the need is huge especially in the healthcare domain), public information sources and public re-usable ontologies (as RDF, OWL, etc.), standard Rules (SWRL, WSML, etc.) and Logic Programming integration to Ontology languages, scalable and robust reasoners, technology transparency for the final user and the practitioners, and these technologies must mature into enterprise-class products, etc.

Pervasive computing is just emerging.

ACKNOWLEDGMENT

This work has been possible thanks to the three large European consortia REWERSE, Knowledge Web and aceMedia. Acknowledgements are also for the large gathering of international conferences mixing research results and prospects from academia and industry: ESWC, ISWC, ASWC, ICWS, WWW, STC etc. Lastly, credits go also directly to the numerous people, in research labs in academia and in industry who are contributing so strongly to make semantic technology a real momentum in industry.

IST-REWERSE is a NoE supported by the European Commission under contract FP6-506779 http://rewerse.net

IST-Knowledge Web is a NoE supported by the European Commission under contract FP6-507482 http://knowledgeweb.semanticweb.org

IST-aceMedia is an Integrated Project supported by the European Commission under contract FP6-001765. http://www.acemedia.org

REFERENCES

Ankolenkar A., Paolucci M., Srinivasan N., and Sycara K., (2004). *The owl services coalition, owl-s 1.1 beta release*. Technical report, July 2004.

Antoniou, G., Baldoni, M., Baroglio, C., Baumgartner, R., Bry, F., & Eiter, T. (2004). Reasoning Methods for Personalization on the Semantic Web. *Annals of Mathematics . Computing & Telefinformatics*, *2*(1), 1–24.

Barillot, C., Amsaleg, L., Aubry, F., Bazin, J.-P., Benali, H., & Cointepas, Y. (2003). Neurobase: Management of distributed knowledge and data bases in neuroimaging. [New-York, NY.]. *Human Brain Mapping*, *19*, 726–726.

Benatallah, B., Hacid, M. S., Léger, A., Rey, C., & Toumani, F. (2005). On automating Web services discovery. *The VLDB Journal*, *14*(1), 84–96. doi:10.1007/s00778-003-0117-x

Berardi, D., Calvanese, D., De Giacomo, G., Lenzerini, M., & Mecella, M. (2005). Automatic Service Composition based on Behavioral Descriptions. *International Journal of Cooperative Information Systems*, *14*(4), 333–376. doi:10.1142/S0218843005001201

Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web. *Scientific American Journal* (May 2001).

Billig A. and Sandkuhl K. (2002). Match-making based on semantic nets: The xml-based approach of baseweb. *1st workshop on XML-Technologien fur das Semantic Web*, 39–51.

Bonifacio, M., & Molani, A. (2005). Managing Knowledge needs at Trenitalia. *Proceedings of the Second European Semantic Web Conference* ESWC 2005, Industry Forum proceedings.

Brunschweig, B., & Rainaud, J. F. (2005). Semantic Web applications for the Petroleum industry. *Proceedings of the 2nd European Semantic Web Conference,* ESWC 2005 Industry Forum

Cardoso, J., & Sheth, A. (2003). Semantic e-workflow composition. [Technical description of the composition of a workflow.]. *Journal of Intelligent Information Systems*, *21*(3), 191–225. doi:10.1023/A:1025542915514

Chan, M. K. S., Judith, B., & Luciano, B. (2006). *Survey and Comparison of Planning Techniques for Web Services Composition*. Technical Report.

Charlet J., Cordonnier E., Gibaud B. (2002). Interopérabilité en médecine: quand le contenu interroge le contenant et l'organisation. *Revue Information, interaction, intelligence*, 2(2), 37-62.

Cordonnier, E., Croci, S., Laurent, J.-F., & Gibaud, B. (2003). Interoperability and Medical Communication Using "Patient Envelope"-Based Secure Messaging. *Medical Informatics Europe Congress*, *95*, 230–235.

Cuel R., Deleteil A., Louis V., Rizzi C. (2008). *Knowledge Web Technology Roadmap: The Technology Roadmap of the Semantic Web*. To appear early 2008.

Dasiopoulou, S., Saathoof, C., Mylonas, Ph., Avrithis, Y., Kompatsiaris, Y., & Staab, S. (2007). Introducing Context and Reasoning in Visual Content Analysis: An Ontology-based Framework. In Paola Hobson, Ioannis Kompatsiaris (Editors), *Semantic Multimedia and Ontologies: Theories and Applications*. Springer-Verlag.

Description Logics for Information Integration. (2002). *Computational Logic: Logic Programming and Beyond*. LNCS; Vol. 2408, pp 41 – 60. London, UK: Springer-Verlag.

Dhamankar, R., Lee, Y., Doan, A., Halevy, A., & Domingos, P. (2004). iMAP: Discovering complex semantic matches between database schemas. *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 383 – 394.

Di Noia, T., Di Sciascio, E., Donini, F. M., & Mongiello, M. (2003). A system for principled matchmaking in an electronic marketplace. *Proceedings of the 12th international conference on World Wide Web*, WWW 2003, 321–330.

Do, H. H., & Rahm, E. (2002). COMA - a system for flexible combination of schema matching approaches. In *Proceedings of Very Large Databases* VLDB 2002, 610–621.

Dou, D., McDermott, D., & Qi, P. (2005). Ontology translation on the Semantic Web. *Journal on Data Semantics*, *3360*, 35–57.

Ehrig, M., & Staab, S. (2004). QOM: Quick ontology mapping. *Third International Semantic Web Conference*, ISWC 2004, LNCS 3298, 683–697.

Euzenat, J., & Shvaiko, P. (2007). *Ontology Matching*. Springer-Verlag.

Euzenat, J., & Valtchev, P. (2004). Similarity-based ontology alignment in OWL-lite. *Proceedings of European Conference on Artificial Intelligence* ECAI 2004, 333–337.

Fensel, D., Kifer, M., de Bruijn, J., Domingue, J. (2005). *Web service modeling ontology (wsmo) submission*. w3c member submission.

Garg, S., Goswami, A., Huylebroeck, J., Jaganathan, S., & Mullan, P. (2005). MediaCaddy - Semantic Web Based On-Demand Content Navigation System for Entertainment. *Proceedings of the 4th International Semantic Web Conference*, ISWC 2005. LNCS 3729, 858 – 871.

Giunchiglia, F., & Shvaiko, P. (2003). Semantic matching. *Knowledge Engineering Review Journal*, *18*(3), 265–280. doi:10.1017/S0269888904000074

Giunchiglia, F., Shvaiko, P., & Yatskevich, M. (2004). S-Match: an algorithm and an implementation of semantic matching. In *Proceedings of the First European Semantic Web Symposium*, ESWS2004, LNCS 3053, 61–75.

Guarino, N., Masolo, C., & Vetere, G. (1999, May). OntoSeek: Content-Based Access to the Web. *IEEE Intelligent Systems*, *14*(3), 70–80. doi:10.1109/5254.769887

Haarslev, V., Moller, R., & Wessel, M. (1999-2007). *RACER: Semantic middleware for industrial projects based on RDF/OWL, a W3C Standard*.

Halevy, A. (2001). Answering Queries Using Views: A Survey. *The VLDB Journal*, *10*(Issue 4). doi:10.1007/s007780100054

Halevy, A., Rajaraman, A., & Ordille, J. (2006). Data Integration: The Teenage Years. 10-year best paper award. *VLDB.*

Hasselwanter, T., Kotinurmi, P., Moran, M., Vitvar, T., & Zaremba, M. (2006). WSMX: a Semantic Service Oriented Middleware for B2B Integration. In *Proceedings of the 4th International Conference on Service Oriented Computing*, Springer-Verlag LNCS series, Chicago, USA.

Heinecke, J., & Cozannet, A. (2003). Ontology-Driven Information Retrieval. a proposal for multilingual user requests. *Workshop on Ontological Knowledge and Linguistic Coding at the 25th annual meeting of the German Linguistics*, Feb. 26-28, 2003.

Hibbard, J. (1997). Knowledge management-knowing what we know. *Information Week* (October 20).

Hull R., (1997), Managing Semantic Heterogeneity in Databases: A Theoretical Perspective, Tutorial at *PODS 1997*.

Hull, R., & Su, J. (2005). Tools for composite Web services: A short Overview. *SIGMOD Record*, *34*, 2. doi:10.1145/1083784.1083807

Lehtola, A., Heinecke, J., & Bounsaythip, C. (2003). Intelligent Human Language Query Processing in mkbeem. Workshop on Ontologies and Multilinguality in User Interface, in the *Proceedings of Human Computer Interface International*, HCII 2003, 4, 750-754.

McGuinness, D. L., Fikes, R., Rice, J., & Wilder, S. (2000). An environment for merging and testing large ontologies. Proceedings of the *Seventh International Conference on Principles of Knowledge Representation and Reasoning* (KR2000), Breckenridge, Colorado, 483–493.

Mcllraith, S. A., Son, T. C., & Zeng, H. (2001). Semantic Web services. *IEEE Intelligent Systems . Special Issue on the Semantic Web*, *16*, 46–53.

Mena, E., Kashyap, V., Sheth, A., & Illarramendi, A. (1996). Observer: An approach for query processing in global information systems based on interoperability between pre-existing ontologies. *Proceedings of the First International Conference on Cooperative Information Systems* CoopIS'96, 14–25.

Nixon, L., Mochol, M., Léger, A., Paulus, F., Rocuet, L., Bonifacio, M., et al. (2004). *Prototypical Business Use Cases*. (Technical report Deliverable 1.1.2), Knowledge Web IST-NoE.

Norheim, D., & Fjellheim, R. (2006). Knowledge management in the petroleum industry. *Proceedings of the 3rd European Semantic Web Conference*, ESWC 2006 Industry Forum.

Papazoglou M., Traverso P., Dustdar S. and Leymann F. (2006). *Service-oriented computing research roadmap*. Technical report.

Petrash, G. (1996). Managing knowledge assets for value. *Proceedings of the Knowledge-Based Leadership Conference*, Linkage Inc., Boston, MA, October 1996.

Petridis, K., Bloehdorn, S., Saathoff, C., Simou, N., Dasiopoulou, S., Tzouvaras, V., et al. (2006). Knowledge Representation and Semantic Annotation of Multimedia Content. *IEEE Proceedings on Vision Image and Signal Processing, Special issue on Knowledge-Based Digital Media Processing*, Vol. 153, No. 3, pp. 255-262, June 2006.

Petrini, J., & Risch, T. (2004). Processing queries over RDF views of wrapped relational databases. In *Proceedings of the 1st International workshop on Wrapper Techniques for Legacy Systems*, WRAP 2004, Delft, Holland, 2004.

Pistore, M., Roberti, P., & Traverso, P. (2005). Process-Level Composition of Executable Web Services: "On-thefly" Versus "Once-for-all" Composition. The Semantic Web: Research and Applications. *Proceedings of ESWC 2005*, Heraklion, Crete, Greece. LNCS 3532, Springer Verlag.

Preguica, N., Shapiro, M., & Matheson, C. 2003). Semantics-based reconciliation for collaborative and mobile environments. In *Proccedings of the Eleventh International Conference on Cooperative Information Systems*, CoopIS 2003, LNCS 2888, 38-55.

Rahm, E., & Bernstein, P. (2001, Dec.). A survey of approaches to automatic schema matching. *Very Large Databases Journal*, *10*(4), 334–350. doi:10.1007/s007780100057

Roman, D., Keller, U., Lausen, H., de Bruijn, J., Lara, R., & Stollberg, M. (2005). Web Service Modeling Ontology . *Applied Ontology Journal*, *1*(1), 77–106.

Rosenbloom ST, Miller RA, Johnson KB. (2006). Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. *Journal of the American Medical Informatics Association.*

Semantic Web Case Studies and Best Practices for eBusiness (SWCASE). At *ISWC2005*, online version

Shvaiko, P., & Euzenat, J. (2005). A survey of schema-based matching approaches. [JoDS]. *Journal on Data Semantics*, *4*, 146–171. doi:10.1007/11603412_5

Shvaiko, P., Giunchiglia, F., Pinheiro da Silva, P., & McGuinness, D. L. (2005). Web explanations for semantic heterogeneity discovery. In *Proceedings of the Second European Semantic Web Conference*, ESWC 2005, 303-317.

Shvaiko, P., Léger, A., Paulus, F., Rocuet, L., Nixon, L., Mochol, M., et al. (2004). *Knowledge Processing Requirements Analysis*. Technical report Deliverable D 1.1.3, Knowledge Web IST-NoE.

Spackman, K. A., Dionne, R., Mays, E., & Weis, J. Role grouping as an Extension to the Description Logic of Ontylog Motivated by Concept Modeling in SNOMED. *Proceedings of the AMIA Annual Symposium 2002*, San Antonio, Texas, p. 712-6, November, 9-13, 2002.

Stamou, G., & Kollias, S. (2005), *Multimedia Content and the Semantic Web: Standards, Methods and Tools*. Wiley.

Stamou, G., van Ossenbruggen, J., Pan, J., & Schreiber, G. (2006). Multimedia annotations on the Semantic Web. *IEEE MultiMedia*, *13*(1), 86–90. doi:10.1109/MMUL.2006.15

Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J. Z., & Horrocks, I. (2005). A Fuzzy Description Logic for Multimedia Knowledge Representation. In *Proc. of the International Workshop on Multimedia and the Semantic Web*, pp 12-19, ESWC 2005, Heraklion, Grece, June 2005.

Traverso, P., & Pistore, M. (2004). Automated composition of Semantic Web services into executable processes. In *Proceedings of the Third International Semantic Web Conference,* ISWC 2004, 380–394, 2004.

Velegrakis, Y., Miller, R. J., & Mylopoulos, J. (2005). Representing and querying data transformations. *Proceedings of the 21st International Conference on Data Engineering* ICDE 2005, 81-92.

Vitvar, T. Zaremba M., Moran M., Haller A., Kotinurmi P. (2007b). Semantic SOA to Promote Integration of Heterogeneous B2B Services. *The 4th IEEE Conference on Enterprise Computing, E-Commerce and E-Services* (EEE07), IEEE Computer Society, July, 2007, Tokyo, Japan.

Vitvar, T., Zaremba, M., & Moran, M. (2007a). Dynamic Service Discovery through Meta-Interactions with Service Providers. In *Proceedings of the 4th European Semantic Web Conference* (ESWC2007), Springer-Verlag LNCS series, Innsbruck, Austria.

Wiederhold, G. (1992). Mediators in the architecture of future information systems. *IEEE Computer*, *25*(3), 38–49.

Wiig, K. (1997). Knowledge management: where did it come from and where will it go? *Journal of Expert Systems with Applications*, *13*(1), 1–14. doi:10.1016/S0957-4174(97)00018-3

Zhong, J., Zhu, H., Li, J., & Yu, Y. (2002). Conceptual graph matching for semantic search. In *Proceedings of the 10th International Conference on Computational Science*, 2393 (2002), 92-106

Zyl, J., & Corbett, D. (2000). A framework for Comparing the use of a Linguistic Ontology in an Application. *Workshop Applications of Ontologies and Problem-solving Methods*, ECAI'2000, Berlin Germany, August, 2000.

ADDITIONAL READINGS

W3C http://www.w3.org/2001/sw/

Annual Semantic Web applications challenge: http://challenge.semanticweb.org

Journal of Web semantics (Elsevier)

The Semantic Web: research and Applications LNCS series: LNCS 2342 (ISWC 2002), LNCS 2870 (ISWC 2003), LNCS 3053 (ESWS 2004), LNCS 3298 (ISWC 2004), LNCS 3532 (ESWS 2005), LNCS 4011 (ESWC 2006), LNCS4273 (ISWC 2006), LNCS 4519 (ESWC 2007), LNCS 4825 (ISWC 2007).

ENDNOTES

1 http://www.web2con.com
2 http://www.ontoweb.org
3 http://Knowledge Web.semanticweb.org
4 http://rewerse.net
5 http://cordis.europa.eu/fp7/home_en.html
6 Knowledge Web Deliverable D 1.1.4v3 http://knowledgeweb.semanticweb.org/semanticportal/deliverables/D1.1.4v3.pdf
7 http://knowledgeweb.semanticweb.org/o2i/
8 http://protege.stanford.edu/index.html
9 http://oaei.ontologymatching.org/2007/
10 http://www.neon-project.org

11 http://www.ontologymatching.org/
12 http://www.acemedia.org
13 http://www.nlm.nih.gov/research/umls/umlsmain.html
14 http://us.expasy.org/sprot/
15 http://obo.sourceforge.net/main.html
16 http://www.ncbi.nlm.nih.gov/Genbank/index.html
17 http://www.ncbi.nlm.nih.gov/sites/entrez
18 http://www.connectingforhealth.nhs.uk/
19 http://www.edisante.org/
20 http://www.etiam.com/
21 http://cen.iso.org/ and http://www.tc251wgiv.nhs.uk/
22 http://www.hl7.org/
23 http://www.d-m-p.org/docs/EnglishVersionDMP.pdf
24 http://www.hl7.org/Special/Committees/structure/index.cfm#Mission
25 http://www.cocoon-health.com
26 http://www.srdc.metu.edu.tr/webpage/projects/artemis/index.html
27 http://www.simdat.org
28 Some of the well-known definitions of KM include: (Wiig 1997) " Knowledge management is the systematic, explicit, and deliberate building, renewal and application of knowledge to maximize an enterprise's knowledge related effectiveness and returns from its knowledge assets"; (Hibbard 1997) "Knowledge management is the process of capturing a company's collective expertise wherever it resides in databases, on paper, or in people's heads and distributing it to wherever it can help produce the biggest payoff"; (Pettrash 1996) "KM is getting the right knowledge to the right people at the right time so they can make the best decision".
29 http://www.cenorm.be/cenorm/index.htm
30 http://www.knowledgeboard.com
31 Semantic Web, Use Cases and Challenges at EADS, http://www.eswc2006.org Industry Forum.
32 See for example in the Petroleum industry (Nordheim D. et al., 2005)
33 See for example Use of Ontology for production of access systems on Legislation, Jurisprudence and Comments (Delahousse J. et al., 2006) http://www.eswc2006.org/industry.html
34 E.g. see the EU Integrated project "DIP Data, Information, and Process Integration with Semantic Web Services", http://dip.semanticweb.org/
35 agentcities RTD project http://www.agentcities.org/EURTD/
36 Semantic Business Automation, SAP, Germany http://www.eswc2006.org Industry Forum
37 http://sws-challenge.org/2006/submission/deri-submission-discovery-phase3/ http://sws-challenge.org/2006/submission/deri-submisson-mediation v.1/ http://sws-challenge.org/2006/submission/deri-submisson-mediation v.2/
38 http://events.deri.at/sercomp2007/
39 Semantically Enabled Knowledge Technologies http://www.sekt-project.com/
40 http://www.acemedia.org
41 http://www.w3.org/2005/Incubator/mmsem/
42 http://www.kpace-noe.net
43 http://www.x-media-project.org
44 http://www.boemie.org
45 http://www.mesh-ip.eu
46 http://www.rottentomatoes.com
47 http://www.imdb.com/
48 Annual Semantic Web applications challenge: http://challenge.semanticweb.org
49 Semantic Technology Conference http://www.semantic-conference.com/ ; European Semantic Technology Conference http://www.estc2007.com/
50 SemWeb@5: Current status and Future Promise of the Semantic Web, James Hendler, Ora Lassila, STC 2006, 7 March 2006, San José, USA

[51] Oracle Database 10g using RDF natively supported by the 10g Enterprise Edition
[52] http://www.w3.org/2001/sw/sweo/public/UseCases/
[53] http://www.w3.org/2001/sw/hcls/
[54] Semantic Web Advanced Development for Europe http://www.w3.org/2001/sw/Europe/
[55] Forester, Erica Driver, October 2007
[56] http://www.w3.org/2005/rules/wg/wiki/RuleML
[57] http://www.oasis-open.org/committees/semantic-ex/faq.php, http://www.wsmx.org http://sourceforge.net/projects/wsmx
[58] http://www.w3.org/TR/sawsdl/
[59] http://lsdis.cs.uga.edu/projects/meteor-s/
[60] http://www.w3.org/Submission/OWL-S/
[61] http://www-ags.dfki.uni-sb.de/~klusch/owls-mx/index.html
[62] http://www.w3.org/Submission/SWSF/
[63] ESTC 2007 Keynote speech from Susie Stephens (Oracle)
[64] Semantic Wave 2006, Part-1 Mills Davis
[65] Web 2.0 and its related phenomena becomes increasingly interesting for businesses. In January 2007 a research programme conducted by the Economist Intelligence Unit and sponsored by FAST gauged the relevance of Web 2.0 to large corporations throughout the world and across a wide range of industries. The research, which consisted of an online survey plus follow-up interviews with senior executives at large public corporations, found that Web 2.0 now has significant implications for big business across a wide range industry sectors. By 2006, and even earlier at some companies, the world's multinationals began to see many Web 2.0 technologies as corporate tools. In fact, according to the survey, 31% of companies think that use of the Web as a platform for sharing and collaboration will affect all parts of their business (Economist Intelligence Unit (2007): Serious business. Web 2.0. goes corporate. A report from the EIU sponsored by FAST. Also to mention two majors initiatives: MySpace with 300 million users (Dec 2007) http://www.myspace.com and Facebook with 60 millions users (Nov 2007) http://www.facebook.com

This work was previously published in Semantic Web for Business: Cases and Applications, edited by R. García, pp. 1-29, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 4.3
A Strategic Framework for Integrating Web 2.0 into the Marketing Mix

Samantha C. Bryant
Philip Morris, USA

ABSTRACT

Marketing strategy set by the marketing mix has remained fundamentally the same through years of other business disciplines being significantly disrupted by emerging technologies. Emerging Web 2.0 technologies such as wikis, blogs, YouTube, and virtual worlds are not only affecting how companies tactically approach marketing, but also their marketing strategies. This chapter will explore the impact of Web 2.0 technologies on marketing and brand management and how companies can leverage these technologies to strengthen relationships between their brands and consumers through a Web 2.0 marketing mix. This new Web 2.0 marketing mix supplements the traditional four-p marketing mix (price, product, promotion, and placement) with a new "p" lens: participation. The focus of this analysis is on B2C marketing of products and services only.

DOI: 10.4018/978-1-60566-122-3.ch003

INTRODUCTION

Marketing has gone through a number of evolutions and technology has revolutionized a number of disciplines. New generations of consumers are consuming media in a different fashion than before. Gone are the days of the 30-second Super Bowl advertisements and here are the days of Facebook, Flickr, and MySpace. Collaborating and participating on the Internet is the preferred entertainment.

Web 2.0 technologies such as wikis, blogs, YouTube, and *Second Life* are changing the behavior of consumers like never before.

- Empowering them with knowledge from a myriad of sources
- Enabling them to self-organize around brands and share their passion (or dissatisfaction) for a brand
- Enabling them to act as marketers of brands

Marketing strategy set by the marketing mix has remained fundamentally the same through years of other business disciplines being significantly disrupted by emerging technologies. However, Web

2.0 is not only affecting how companies tactically approach marketing, but also their marketing strategies.

This chapter will explore the impact of Web 2.0 technologies on marketing and brand management and how companies can leverage these technologies to strengthen relationships between their brands and consumers through a Web 2.0 marketing mix. The focus of this analysis is on business-to-consumer (B2C) marketing of products and services only.

BACKGROUND: WEB 2.0 TECHNOLOGIES AND MARKETING

Web 2.0 technologies, also known as social software technologies, are a second generation of Web-based communities and services that facilitate collaboration and sharing between users. Web 2.0 does not refer to an update to any technical specifications of the Web but to changes in the way it is being used. These technologies are built on an architecture of participation ("Web 2.0," 2007).

Little has been published about Web 2.0 technologies, particularly about their effect on marketing. In 2004, *High Intensity Marketing* explored the effects of a new emerging stream of networked technologies on marketing. However, these technologies were the mere beginning of what was to explode a few years later as Web 2.0. (Mootee, 2004). *Wikinomics: How Mass Collaboration Changes Everything* discussed Web 2.0 technologies, with examples of companies' use to interact with their consumers and improve their product offerings (Tapscott & Williams, 2006). The American Marketing Association (AMA) is one of the largest professional organizations for marketers and is trusted to provide relevant marketing information to experienced marketers. One channel through which it educates its members is conferences. The AMA has just begun to acknowledge how Web 2.0 is disrupting traditional marketing theory. In 2008, the AMA hosted its first conferences on Web 2.0 (social media) and marketing. The conferences focused on how companies need to recognize the impact social media has on their brands and how they can benefit from using social media as new marketing tools (American Marketing Association, 2007a).

Not surprisingly, the most written about this relationship between Web 2.0 technologies and marketing has been via Web 2.0 technologies themselves, particularly blogs. Blogs, also known as weblogs, are shared online journals or diaries where people can post entries via the Web. LiveJournal, Blogger, and WordPress are examples of online blog services where users can post their thoughts with an emphasis on user interaction within the community. LiveJournal is one of the most popular, with currently over 14 million journals and communities. Numerous individuals blog daily about what they are witnessing and hearing about in the marketing world. These same individuals then harness the power of Web 2.0 by linking to each other's blogs and commenting on the author's thoughts. What results is insightful speculation about trends emerging as a result of Web 2.0's impact on marketing and brand management.

THE TRADITIONAL MARKETING MIX (FOUR *P*S)

With such revolutionary technologies disrupting consumers' lives, many businesses wonder what the implication is to the marketing of their products and brands.

The marketing mix, invented in the 1950s, is the mix of controllable marketing variables that a company uses to pursue the desired level of sales in the target market. The most common model of these factors is the four-factor classification called the four *P*s. Optimization of the marketing mix is achieved by assigning the amount of the marketing budget to be spent on each element of the marketing mix so as to maximize the total contribution to the firm (American Marketing Association, 2007b).

The four *P*s of the traditional marketing mix are product, pricing, promotion, and placement.

- **Product:** The product aspects of marketing deal with the specifications of the actual goods or services, and how it relates to the user's needs and wants. Generally, this also includes supporting elements such as warranties, guarantees, and support.
- **Pricing:** Pricing refers to the process of setting a price for a product, including discounts. The price need not be monetary; it can simply be what is exchanged for the product or services, such as time, energy, psychology, or attention.
- **Promotion:** Promotion includes advertising, sales promotion, public relations, and personal selling, and refers to the various methods of promoting the product, brand, or company.
- **Placement:** Placement or distribution refers to how the product gets to the customer, for example, point-of-sale placement or retailing. This *p* is also the place, referring to the channel by which a product or service is sold (e.g., online vs. retail), to which geographic region or industry, and to which segment (young adults, families, professionals; American Marketing Association, 2007b).

While effective for setting marketing strategy since its creation, the traditional marketing mix lacks relevancy when it comes to Web 2.0. A new term coined Marketing 2.0 is being used to describe the impact Web 2.0 has had on the discipline of marketing. Companies are finding many uses of Web 2.0 technologies to successfully connect with their consumers. These are discussed in the next section.

ISSUES AND EXAMPLES: THE EFFECT OF WEB 2.0 TECHNOLOGIES ON MARKETING AND BRAND MANAGEMENT

Numerous companies are integrating Web 2.0 technologies into their portfolio of marketing channels. Let us explore some Web 2.0 technologies and how they are affecting marketing strategy.

Second Life (Virtual World)

A virtual world is a computer-based simulated environment intended for its users to inhabit and interact, usually represented in the form of graphical representations of avatars. *Second Life* is an Internet-based virtual world developed by Linden Research, Inc. that is a massive multiplayer online game. *Second Life* enables its users to interact with each other through motional avatars, providing an advanced level of a social network service. Residents can explore, meet other residents, socialize, participate in individual and group activities, and create and trade property and services with one another.

Second Life has received much media attention regarding its benefits to the business world, most significantly from a cover story in *Business Week* in April 2006 that brought the virtual world to the attention of the masses, including a number of business leaders. The unique avatar population in the virtual world *Second Life* topped 7 million in 2007, with about 4 million distinct individuals participating in the online world (Rose, 2007).

Second Life has been recognized to have so much marketing potential that new companies have been established to assist companies with establishing their presence. For example, Millions of Us, an agency specializing in virtual worlds, designs and measures marketing programs for clients across a wide spectrum of platforms, especially *Second Life* (*Millions of Us*, 2007). Another company, The Electric Sheep Company, is the largest company in the world dedicated to designing experiences and add-on software for 3-D virtual worlds and has implemented major projects in both *Second Life* and There (another virtual world) technologies (Carter, 2007).

In addition to Millions of Us and The Electric Sheep, REPERES has formed as the first market research institute in *Second Life*. REPERES performs quantitative research by surveying avatars and qualitative research via private interviews with avatars. REPERES assists companies in the development of their products and offers in *Second Life* using the following:

- A panel of avatars that are representative of the overall population of *Second Life* in terms of nationality, gender, and age; this panel is called upon to address issues faced by brands seeking to establish themselves or develop an offer on *Second Life*
- An understanding of behaviors, innovations, and trends in *Second Life*
- A space for tests and cocreation projects to be tried and evaluated (*REPERES Second Life*, 2007)

In mid-2006, Scion sensed an opportunity to engage the technology and design-oriented communities of *Second Life*. Millions of Us and Scion collaborated to create Scion City, a *Second Life* island that housed the first virtual-world car dealership, representing the first time a major auto manufacturer created a presence in a virtual world. *Second Life* residents had the opportunity to not only purchase all three Scion models in the dealership, but also customize them to make them their own. This is consistent with Scion's approach to allow buyers to customize numerous aspects on real-world cars as well. This launch was such a success that Scion and Millions of Us have continued to work together to extend Scion's presence in *Second Life* through the following:

- The expansion of Scion City into a full-fledged urban environment where residents are able to develop their own homes and businesses around the Scion dealership
- A simultaneous real and virtual launch of Scion's 2008 line at the Chicago Auto Show, meaning the new vehicle was launched in both *Second Life* and the real world at the same time
- Free expert-led customization classes in *Second Life* for consumers to learn how to personalize their virtual Scions; consumers can then showcase their designs in a Scion gallery in *Second Life*

The Scion xB launch received substantial media attention, and the buzz only grew with each subsequent release and event. Scion City has organically developed its own culture and loyal base of residents; one even created a MySpace page chronicling the evolution of this community, which shows the convergence and power of Web 2.0 technologies working together (*Millions of Us*, 2007).

While Scion was the first major automobile manufacturer with a significant presence in *Second Life*, numerous car companies have followed suit, including Pontiac, Mercedes, BMW, Nissan, and Toyota. Mercedes-Benz operates a car dealership that sells virtual cars and gives away branded racing suits to avatars. BMW even allows avatars the opportunity to test drive their vehicles.

At the 2007 Food Marketing Institute Supermarket Convention and Educational Exposition, Kraft Foods, Inc. unveiled more than 70 new products. Kraft chose to showcase these new offerings at a virtual supermarket in *Second Life*. Online, consumers and convention attendees could interact with Kraft's latest products and take part in online forums with Kraft experts. By having new products introduced online, consumers could see, "touch," and learn about the product before it was able to hit physical supermarket shelves (Kraft Foods, 2007).

In addition to Scion and Kraft, numerous other brands have built a presence in *Second Life*, including Adidas, Dell, Reebok, Sony BMG, Vodaphone, Sun, Sears, AOL, and Circuit City.

Reebok opened a virtual store that is an extension of its real-world RBK custom campaign. The store sells plain white sneakers by size and then features coloring machines for avatars to customize their shoes. Reebok grounds also contain a basketball court for avatars to play on.

Adidas owns an island that features branded video clips and billboards. Avatars who purchase Adidas shoes at the store may then bounce on a trampoline next to the store, demonstrating Adidas' shoes' bounciness.

Wired magazine reports that at least 50 major companies have a *Second Life* presence (Rose, 2007). YouTube is a video-sharing Web site where users can upload, view, and share video clips. So many brands and products have moved in that a YouTube video of brands in *Second Life* has been developed (Hayes, 2007).

Widgets

A widget is a small application that can be embedded on different Web pages. Widgets represent a new indirect marketing channel that enters consumers' lives via other Web 2.0 technologies and effectively promotes a brand, product, or service and generates awareness. A widget can be placed by users onto their personalized home pages, blogs, or other social networking pages. The widget is not intrusive advertising as a consumer must actively choose to add the widget.

The content on widgets can include blogs, live discussions, bookmarks to other Web sites, webcasts, video, games, and more. Companies are already capitalizing on widgets to market their brands, products, and services. On its traditional Web site, CBS lets consumers select widgets to embed in their social networking profiles or blogs to allow them (and everyone else who views their pages) to see advertising about the shows. For example, CBS launched a constantly updating rich-media widget to promote a new series. Consumers can watch full episodes directly on the widget, get short mini clips of some of the stars, send the widget to a friend, link out to download ring tones, and more. The widget is designed to stand out on any site it is placed on ("CBS Mobile Launches Widget to Promote New Animated Series," 2007).

Sony promoted the film *Zathura* via widgets vs. a traditional online option like banner ads. Widgets enabled Sony to provide an application related to the movie and more importantly allowed users to interact with it. The widget was offered via Freewebs, a Web site that enables consumers to easily create their own Web sites at no cost. The 11 million Web site owners of Freewebs were able to embed the widget within their Web sites, and 11,000 Web sites took this up within 6 weeks. The widgets were viewed 600,000 times, and long after the original movie was out of theaters, thousands of widgets were still delivering content (Jaokar, 2006).

To enable PC users to experience a game offered on a different gaming system (Xbox), interactive agency AKQA created a weather widget to promote the *Microsoft Flight Simulator X* game for the Xbox. The widget allows users to virtually fly and find out the weather at any airport through a live feed from the National Weather Service. In the first 2 months, users downloaded the widget more than 150,000 times, spending an average of 23 minutes with the flight simulator, the agency says (Steel, 2006).

Numerous other companies are using widgets to market their products and services. Reebok created a widget that allows users to display customized pairs of RBK shoes for others to critique. Radio stations are offering widgets that stream a station's broadcast live. Airlines including American Airlines and Air France offer a ticket purchasing widget that allows a consumer to purchase a ticket through the widget without having to visit the airlines' Web sites. (Guiragossion, 2007).

Having a user add a widget to a personalized site that is visited by individuals who share similar interests is an effective way to share your marketing message with those who would find it most relevant. By adding the widget to his or her page, the user is showing endorsement for the brand, making others trust the brand.

Flickr (Tagging)

Tagging is the assignment of descriptive contextual tags to data, such as Web links stored with memorable words for easy future access. An example is Flickr, an online photo-management and -sharing application that allows users to collaboratively organize their photos by tagging them with descriptors that are searchable.

On Flickr, users can post pictures and tag them with words that describe the photo so others can search and view the photos. One quick search on Flickr reveals 88,923 pictures tagged with Coke, 4,650 with Crayola, 3,472 for Vitamin Water, and 32,367 for Heineken. One user posted a picture entitled “Just been to the Heineken Experience,” described as a photo taken of the photographer’s friend after an exciting trip to the Heineken Experience tourist attraction housed in the former Heineken brewery on the Stadhouderskade in Amsterdam. The administrator for a group entitled Got Heineken? contacted the user and invited them to join the Heine group. The Heine group invites members to post pictures of themselves holding a nice ice-cold Heineken or simply a photograph of anything related to Heineken. There is no official indication that Heineken is sponsoring it. Currently, the group has 363 photos posted by 212 members.

Eurostar, the United Kingdom’s high-speed train, recognized that moving from Waterloo to St. Pancras International Station was a significant event in the brand’s life. Eurostar set up an account with Flickr, EurostarForTomorrow, to promote the move to a new station. Eurostar posted a number of photos of the new station, and collected a number of photos other users have taken and saved them as favorites. This was an effective way to generate interest about the Eurostar brand and promote a significant event in the brand’s history. It also recognized Eurostar consumers who had taken photos of Eurostar and shared them with others on Flickr (Terret, 2007).

Social Networks

Social networks are another Web 2.0 technology disrupting traditional marketing strategy. Social networks are Web sites designed to allow multiple users to publish content. Users are able to connect with those sharing similar interests and to exchange public or private messages. Facebook and MySpace are popular examples. Facebook is a social utility that connects users with other people by enabling them to publish profiles with text, photos, and videos; to review friends’ profiles; and to join a network. MySpace allows users to create a community and share photos, journals, and interests with a growing network of mutual friends. Some marketers are making their own social networking sites such as Coke’s mycokerewards.com and USA Network’s Characters Welcome Web site.

For those brands that have not created their own social networking site, existing social networks enable consumers to create brands for themselves and assist others with creating their brands. At MySpace, brands have member profiles and make friends with other MySpace members. At Facebook, members join Facebook brand groups, just like they join Facebook fraternity or hobby groups, and can even display brand logos on their personal profiles.

Social branding may prove to be the ultimate product placement strategy. Companies can create a page for their brand or product. For example, Big Sky Brands, maker of Jones Soda, created its own page. In addition to company information and a list of retailers carrying the Big Sky Brand’s products, its MySpace page contains a blog with discussions about the new Jones Activated Energy Boosters, the latest extension to the brand family.

The blog also talks about promotions, such as the Hot Topic and Jones Soda Carbonated Candy Summer Promo. Currently, Big Sky Brands has 495 friends who, in the comments section of the page, engage in discussions about how much they love the brand and which products are their favorites. The site also links to Big Sky Brands Music MySpace page, another site dedicated to empowering their consumers with Big Sky Brands' enthusiasm through the medium of music (Big Sky Brands, 2007).

In addition to this official Big Sky Brands Web site, a number of unofficial MySpace pages have been developed to support Jones Soda. The majority of these page owners have become MySpace friends of Big Sky Brands. In contrast, Swiffer WetJet also has a number of MySpace pages, but none are sponsored by Procter & Gamble.

Virtual worlds, widgets, tagging, and social networks are Web 2.0 technologies offering opportunities for companies to integrate them into their marketing mix. The next section provides a simple framework to help leverage Web 2.0 to supplement and enhance their marketing efforts.

SOLUTIONS AND RECOMMENDATIONS: A REVISED MARKETING MIX TO MAXIMIZE THE BENEFITS OF WEB 2.0 TECHNOLOGIES

A characteristic of Web 2.0 is an architecture of participation that encourages users to add value to the application as they use it. Tapscott and Williams (2006) argue in their book *Wikinomics: How Mass Collaboration Changes Everything* that the economy of Web 2.0 is based on mass collaboration that makes use of the Internet. Companies can leverage the collective power of their consumers by leveraging Web 2.0 technologies to enable the consumers to participate in the marketing of the brand.

Idris Mootee (2004) in his book *High Intensity Marketing* examines the role of strategic marketing in the network economy, which is relationship driven, network centered, technology enabled, and information intensive. Mootee developed an analytical model supplement to the traditional marketing *p*s called "The New 4P's." Mootee's new four *p*s are participation, peer-to-peer communities, predictive modeling, and personalization. Participation focuses on allowing consumers to choose their products, providing valuable insights to companies about consumers' needs and wants. Mootee cites Dell, Procter & Gamble, and Levi's as examples. Mootee also briefly touches on consumers playing a role in defining and owning brands, such as Burton snowboards.

Participation is more relevant than ever with the Web 2.0 revolution, but it is more than a supplemental *p* to the marketing mix. Participation should now encompass all of the four *p*s as a lens through which the *p*s should be approached. The previous section explored a number of examples of B2C companies successfully leveraging Web 2.0 technologies to market their products. Applying a participation lens to the marketing mix can lead similar companies to similar success.

To be more successful at marketing using Web 2.0 technologies, the traditional four-*p* marketing mix discussed earlier (price, product, promotion, and placement) should be approached with a new *p* lens: participation. This is the Web 2.0 marketing mix.

In this new marketing mix, each of the four elements is approached by enabling consumers to participate in it. Let us explore how consumers are currently participating in each *p* of the marketing mix and the steps a company can take to capitalize on that interaction and encourage more.

Web 2.0 technologies provide a medium for consumers to directly participate with companies and their brands. Consumers can write blogs about their favorite brands, define themselves and create communities in social networks like Facebook and MySpace by associating themselves with brands, develop their own commercial advertisement for brands and post them on YouTube, and tag pictures they have uploaded of their brands on Flickr.

In addition, Web 2.0 technologies enable companies to more readily identify brand evangelists: those consumers that are devoted to a brand or product and preach their devotion to the world.

Companies can make it even easier for consumers to participate in their marketing by evaluating each *p* in their marketing mix and fostering involvement by their consumers via Web 2.0 technologies. Let us specifically examine each marketing mix element.

Product

The product aspects of the marketing mix are the specifications of the actual goods or services, and how they relate to the user's needs and wants. The scope of a product generally includes supporting elements such as warranties, guarantees, and support.

In *Wikinomics: How Mass Collaboration Changes Everything*, Tapscott and Williams (2006) explore a new generation of "prosumers" who treat the world as a place for creation, not consumption. These prosumers can be engaged to participate in all aspects of a product, from identifying consumers' unmet needs to the development of a new product to supporting the product once in a consumer's hands. Companies need to encourage consumers to participate in their products and recognize the contributions that are made. There is an allure of prestige and sense of social belonging that develops within prosumer communities (Tapscott & Williams).

Web 2.0 provides immediate feedback about products. Companies do not need a Web site to harness the power of the numerous discussions of products on Web sites beyond the control of the company. Blogs are a great tool to identify latent consumer needs and wants. Communities are free focus groups of very raw, unscripted feelings not tainted by groupthink, or the act of reasoning or decision making by a group that often occurs in traditional focus groups. Companies can also learn more about the total product life cycle beyond just the physical product, and identify additional uses for it. Companies can gain valuable customer insight and interaction via Web 2.0 in a quicker and cheaper fashion than traditional market research.

Flickr can be an avenue for trend spotting. Trend spotting is a relatively new consumer research methodology that seeks to anticipate what consumers will desire in the future and to keep existing products relevant. For example, on Flickr, a company can search for what consumers are carrying in their purses. Companies can ask consumers to participate in identifying product trends by sending out a call to action to post photos of what is in their purses. Maybe Capital One with their "What's in your wallet?" campaign could ask consumers to post Flickr photos of what is in their wallets, including a Capital One credit card. While a company may not know a lot about the consumer, Web 2.0 enables it to reach into consumers' lives and learn more about their behaviors (Brighton, 2005).

Lego fostered an early prosumer community. Lego's Mindstorms enables users to build working robots out of programmable bricks. Users reverse-engineered the products and shared feedback with Lego. Lego developed a Web site for users to share their discoveries and inventions with other enthusiasts and the company. Users can even virtually develop their own models and then order the bricks to physically build it. This enables Lego's consumers to become a decentralized virtual design team, far larger than the number of the in-house designers (Tapscott & Williams, 2006).

Procter & Gamble is also successfully leveraging the powers of Web 2.0 to enhance its research and development efforts for new products. The research and development team had a success rate of less than 20%, below industry standards of 30%. Via the Web, Procter & Gamble turned to the outside world for new and better ideas, and now more than 35% of the ideas come through the Web, resulting in success for 80% of Procter & Gamble's new product launches.

One example was the discovery of a way to put edible ink pictures on potato chips. The solution came from an Italian professor at the University of Bologna who had invented an ink-jet method for printing edible images. This technology helped the company get the new Pringles Prints potato chips out in a single year, about half the normal time for such a process (Stephens, 2007).

Cadbury brought back the discontinued Wispa chocolate bar after a campaign on Web sites like Facebook, MySpace, and YouTube demanded its return. The chocolate company says that it is frequently contacted by consumers asking for old favorites to be reintroduced, but said the numbers that had joined the Internet campaign to relaunch Wispa were unprecedented (Wallop, 2007).

In the support stage of the product life cycle, brands have an opportunity to engage with consumers during a crisis using Web 2.0. Consumers are pretty responsive to companies and brands who engage, especially when there is a problem. Speaking at Nielsen Business Media's Next Big Idea Conference, EVP of strategic services at Nielsen Online Pete Blackshaw focused his remarks on brand intelligence as he explained how companies can "defensively" market their products by turning negative trends in the marketplace to their advantage. Blackshaw noted the surge in the blog traffic surrounding consumer crises, such as the Mattel toy lead paint and poisonous pet food recalls from brands including Alpo and Mighty Dog, are opportunities for brands to have a touch point with consumers. Companies, said Blackshaw, should learn to "manage around the spikes, listen, react," and move money out of mass media and into online channels, asking consumers to participate in the support of a product, even during a recall (Kiley, 2007).

Pricing

The pricing *p* of the marketing mix refers to the process of setting a price for a product, including discounts. The price need not be monetary; it can simply be anything exchanged for the product or services, including time, energy, psychology, or attention.

Companies can leverage Web 2.0 to learn more about value that consumers perceive about brands and what consumers perceive they are paying for a product beyond just the monetary value. For example, online Web services such as eBay and craigslist let the consumer (or market) determine the price for products.

Consumers are able to discuss prices online via blogs or social networking sites for plane tickets, HDTVs, furniture, cars, and so forth. They can even let other consumers know where to find the best discounts.

A mashup is a Web page or application that integrates complementary elements from two or more sources. The most popular mashups include Google Maps as a source to identify certain things on a map. For example, one mashup shows secret fishing holes in the United States via a Google map. Mashups help locate the best price, such as CheapGas.

Beyond monetary value, companies can also get see what consumers are forfeiting other than money by choosing their brands or products. For example, a number of loyal Starbucks consumers are concerned about the impact they are making to the environment by enjoying Starbucks coffee in a new cardboard cup everyday with a cardboard sleeve. These consumers found a Web site that sold inexpensive reusable cloth holders to protect hands from the hot coffee cup and were elated to share it with one another. Starbucks could learn about this concern of its consumers and address it, possibly by offering this reusable cloth sleeve product in stores or developing one of their own.

Promotion

Promotion is the element of the marketing mix where consumers can most readily participate and add value to the marketing of a brand. This includes advertising, sales promotion, public relations, and personal selling, and refers to the various methods of promoting the product, brand, or company.

Consumers can develop an advertisement for a product or brand and then publish that ad for free on YouTube, tagged with keywords to inform others of the ad.

Companies may also use Web 2.0 to promote their brands with an advertisement. Smirnoff developed a viral marketing video for the launch of their new Raw Tea product. The video was placed on YouTube and at this time had over 4 million views (Iamigor, 2006). The Web 2.0 technology enabled consumers to access the video on YouTube and then participate in disseminating the information by informing their friends about the video. The launch of Raw Tea also included a Web site with the videos and a sharing capability. By placing the advertisement on a Web 2.0 site, Smirnoff made it easier for their consumers to participate by rating and sharing the video on a site they already frequent.

A wiki is any collaborative Web site that users can easily modify via the Web, typically without restriction. A wiki allows anyone using a Web browser to edit, delete, or modify content that has been placed on the site, including the work of other authors. One popular wiki is Wikipedia, a free encyclopedia that anyone can edit. Wikipedia is updated every second by thousands of active contributors, making it an up-to-date reference source vs. a printed encyclopedia that is updated monthly or yearly.

Wikis capture the knowledge of the collective whole. Consumers can define your product or brand on Wikipedia. Consumers define what a brand stands for or describe a product, including its intangible intrinsic value. Companies should check for accuracy but not stifle the participation by their consumers to define the brand.

A key to enabling consumers to participate in promotion is to make it easy to search and find areas where other consumers talk about your brand, product, or service, or create your own site where consumers can talk about what you are interested in learning more about. Procter & Gamble developed a social networking site for women called Contessa. The intention of the social network is not about selling products, but for P&G to learn more about its women consumers and learning about their needs and habits (Ives, 2007).

Placement

Placement or distribution refers to how the product gets to the customer, for example, point-of-sale placement or retailing. This fourth *p* of the marketing mix has also sometimes been called *place*, referring to the channel by which a product or service is sold (e.g., online vs. retail), to which geographic region or industry, and to which segment.

Companies can distribute their products via Web 2.0. American Apparel operates a clothing store in *Second Life* that sells virtual clothing for avatars. American Apparel has run promotions in *Second Life* where after purchasing a clothing item in *Second Life*, a consumer receives a coupon for a discount on the same or similar item at an American Apparel store in the real world (Jana, 2006).

Coke established the Virtual Thirst Pavilion in *Second Life*. It sponsored a contest to develop a Virtual Thirst vending machine. The winning vending machine will be rolled out throughout *Second Life*, making this real-world distribution channel as ubiquitous in *Second Life* as it is in the real world. Coke did not intend for consumers to merely replicate an existing real-world vending machine but to create a portable device for Second Life's in-world digital society that unleashes a refreshing and attention-grabbing experience, on demand. Our goal is to enable individual creativity in pursuit of a "vending" machine that can exist only in your wildest imagination. Virtual worlds make it possible for such innovations to occur, and we selected Second Life as the most conducive to this experiment,

says Michael Donnelly, director of global interactive marketing (Coca-Cola, 2007). This *Second Life* contest establishes a new distribution channel for the brand in a virtual world, but also taps into consumers' creativity for ideas for real-world distribution mechanisms.

On Flickr, in addition to tagging photos, a user may add notes that are visible when someone viewing the photo scrolls over a particular aspect of it. For new product introductions, such as a new pair of Nike shoes, a photo of the shoe may be uploaded to Flickr with associated tags (Nike, new, shoe, blue, limited edition, New York store) and notes that can link to a Web site where the shoe may be purchased.

Companies can even create a mashup using Google Earth to locate products that are not widely distributed such as specific shoe sizes. Online retailers such as Amazon.com and Travelocity are building widgets that can drive traffic to their sites for consumers to make purchases.

CHALLENGES

The consumer defines the brand; a brand is not what a company defines it as, but what a consumer says it is. Via Web 2.0, consumers have numerous avenues to add their interpretation of brands, for better or worse. Brands need to relinquish control to get influence. Two barriers to adoption are regulation of content and security.

One of the reasons Web 2.0 technologies like YouTube and Flickr are successful is because they are authentic: "The lack of corporate polish adds to the feeling that there are real people behind the idea" (Moore, 2006).

What does it say about a brand to have a presence or be communicated via these technologies? Before embracing participation in Web 2.0, a company must determine if a brand is compatible with the spirit of Web 2.0. Web 2.0 is more about spirit, concepts, and principles than definition. It is imperative that a brand be in accordance with that spirit before launching an initiative. If a brand is incompatible with this experience of openness and exchange, it is advised to create or use another brand or subbrand as a workaround for Web 2.0 initiatives to protect the integrity of the core brand (Smagg, 2007).

The best action for a company to take may be no action at all as long as it is recognized that Web 2.0 technologies are having an effect on marketing. In summer 2007, *Wired* magazine exposed *Second Life* for not achieving the commercial potential that was initially expected. A trip to Starwood Hotels' Aloft Hotel in *Second Life* was described as creepy due to the entire place being deserted, compared to the movie *The Shining*, where a tenant at a deserted hotel goes psycho due to the isolation—not exactly the reputation Starwood wanted for its new hotel chain brand. The NBA sought to capitalize on Web 2.0 for its marketing efforts by developing both an island in *Second Life* and a channel on YouTube. The YouTube channel saw over 14,000 subscribers with 23 million views, while the *Second Life* island had a mere 1,200 visitors. Those numbers are not as surprising when it is revealed that the traffic in *Second Life* is slightly more than 100,000 Americans per week (Rose, 2007).

In late 2007, IBM broadcast a television advertisement featuring two employees of an unknown company talking about avatars from a virtual world. The dialogue is as follows:

"This is my avatar. It's all the latest rage. I can do business, I even own my own island! It's innovation!" "But...can you make money?" "Um... virtual or real money?" "Real money. The point of innovation is to make real money." "Oh. My avatar doesn't know how to do that." (Vielle, 2007)

While the advertisement is mocking the value of Web 2.0 technologies, in a roundabout way, IBM is capitalizing on Web 2.0 to market its "Stop Talking. Start Doing." campaign. Yet, in mid-2007, IBM established a presence in *Second Life*, the virtual IBM Business Center staffed by real IBM sales representatives from around the world. In the press release for the launch of this virtual center, IBM boasts it has over 4,000 employees active in *Second Life*. The question remains: Is IBM not seeing the value from this investment in a Web 2.0 technology, or did it create the advertisement mocking avatars merely to generate interest for the IBM brand because avatars are popular at the moment? It also could have been buzz for its own presence in *Second Life* (IBM, 2007).

IMMEDIATE ACTION FOR COMPANIES

Companies should use Web 2.0 to assist with keeping a pulse on their brands' and products' involvement in Web 2.0 technologies. Most brands already have a Web 2.0 presence, and most likely, it is not officially endorsed by the brand.

Web 2.0 technologies make it very simple for a marketer to quickly and easily learn about what consumers are saying about a brand or product. RSS or real simple syndication is any of various XML (extensible markup language) file formats suitable for disseminating real-time information via subscription on the Internet. RSS has become a popular technology for bloggers and podcasters to distribute their content. NewsGator is a free Web-based RSS news reader that consolidates news and updates from the Web, blogs, premium content providers, and internal applications and systems and automatically delivers them to users.

RSS feeds can be set up to find and aggregate information about a brand or product from blogs, podcasts, and so forth, and be sent out via e-mail every morning to keep a marketer updated. Technorati is an RSS service that searches and organizes blogs and other forms of user-generated content (photos, videos, voting, etc.). Technorati is currently tracking over 110 million blogs and over 250 million pieces of tagged social media. The home page will immediately update you on "what's percolating in blogs now," but a quick search of your brand's name will share what is hot about your brand at the moment. It is an easy way to monitor the brand's image (*Technorati*, 2007).

Quick monthly visits to Web 2.0 Web sites can also keep marketers up to do date about their brands (and the competition). Visit Flickr and type in your brand name. Cheerios returns over 8,000 images, many showing how Cheerios plays a role in the everyday life of a child. Marketers will be happy to find this is consistent with the brand's image, values, and positioning.

Visit YouTube and type in your brand name. A McDonald's marketer may be interested to find that some commercials are tagged as creepy and racist and are not positively portraying the megabrand. However, a video containing the McDonald's menu song has had over 1.5 million views. The McDonald's menu song was a song listing all of McDonald's menu items from a promotion in 1989. Reviewing the thousands of comments about the video reveals that consumers are excited to view this video on YouTube and remember memorizing the song in 1989 when it was advertised during television commercials. This is definitely some valuable insight available for McDonald's.

Visit MySpace and type in your brand name or visit http://www.myspace.com/brandname. Mountain Dew does not have an official page, but a 21-year-old from Southern California owns http://www.myspace.com/mountaindew. JC from Newport Beach, California, owns Nike. In addition to learning about consumers' thoughts on your brand, this exploring can also result in insights about your loyal consumers who add your brand to their MySpace pages.

Visit Wikipedia and type in your brand name. Wikipedia's American Express entry has a wealth of historical information, including advertising. The entry is up to date with AmEx's most recent promotions, including its Member's Project. Anyone can contribute to a wiki, so periodically verifying information is correct and adding updates can benefit a brand ("Web 2.0," 2007).

With Web 2.0 technology, companies can quickly and inexpensively make things happen. You can have your advertising messages spread on the Web like a wildfire with social bookmarking sites, RSS, and other Web 2.0 methods, and without having to pay anything for it, have thousands of people coming to your Web site in a matter of days. What would have cost you millions of dollars in investment and a dedicated team of developers may now be accomplished with these Web 2.0 tools by a couple of guys in a garage in just a few days (Beaudoin, 2007).

Web 2.0 technologies intermingle with each other as well, so a presence in one Web 2.0 technology can give you a presence in others. For example, MySpace allows its members to blog on their pages and add YouTube videos.

FUTURE TRENDS

The Web is not going anywhere and will only continue to evolve. More Web applications will be developed and compete for consumers' time and attention. As consumers become accustomed to communicating and collaborating on the Web, they may choose to take marketing your brands into their own hands. Companies should be delighted at this occurrence. It is best to identify and monitor these actions as they transpire, but not take actions to impede them. If a consumer is passionate enough to devote energy toward positively marketing your brand, you can be sure that once disgruntled they will also take action, only this time it could be detrimental.

CONCLUSION

Web 2.0 technologies will not replace traditional marketing such as direct mail or TV advertisements, but instead are new complementing marketing channels that many consumers will expect their brands to communicate through.

Companies that feel Web 2.0 technologies are a right fit for their brands should take action to market their brands using the Web 2.0 marketing mix as a guide. Enabling consumers to participate in each aspect of the marketing mix will help brands remain relevant in today's changing world.

REFERENCES

American Marketing Association. (2007a). *Beyond Marketing 2.0: Harnessing the power of social media for marketing campaign results.* Retrieved December 28, 2007, from http://www.marketingpower.com/aevent_event1221988.php

American Marketing Association. (2007b). *Dictionary of marketing terms.* Retrieved October 15,2007, from http://www.marketingpower.com/mg-dictionary-view1876.php

Beaudoin, J. (2007). *Web 2.0 for marketers ebook review on Web 2.0 portals.* Retrieved November 26, 2007, from http://www.web20portals.com/web-20-marketing/web-20-for-marketers

Big Sky Brands. (2007). *MySpace.* Retrieved November 26, 2007, from http://www.myspace.com/bigskybrands

Brighton, G. (2005). Flickr kills trendspotting. *PSFK.* Retrieved September 27, 2007, from http://psfk.com/2005/03/flickr_kills_co.html

Carter, B. (2007, October 4). Fictional characters get virtual lives, too. *The New York Times.* Retrieved November 12, 2007, from http://www.nytimes.com/2007/10/04/arts/television/04CSI.html?_r=1&ex=1192161600&en=6994ec8ad88ab2d8&ei=5070&oref=slogin CBS mobile launches widget to promote new animated series. (2007). *MuseStorm.* RetrievedDecember 28, 2007, from http://www.musestorm.com/site/jsp/spotlight/1042.jsp

Coca-Cola. (2007, April 16). *Coca-Cola launches competition to design online "virtual thirst" coke machine.* Retrieved October 14, 2007, from http://www.virtualthirst.com/virtualthirst-socialmediarelease.html

Guiragossion, L. (2007, October 29). A thousand ways to widget in the age of Web 2.0 marketing. *Web 2.0 Marketing.* Retrieved December 28, 2007, from http://web2pointzeromarketing.blogspot.com

Hayes, G. (2007). *A video comp of major brands in Second Life.* Retrieved December 28,2007, from http://www.youtube.com/watch?v=tEGHJuCbGdo

Iamigor. (2006, August 2). *Tea partay.* Retrieved October 15, 2007, from http://www.youtube.com/watch?v=PTU2He2BIc0

IBM. (2007, May 15). *Live IBM sales people to staff new virtual IBM business center.* Retrieved October 15, 2007, from http://www-03.ibm.com/press/us/en/pressrelease/21551.wss

Ives, B. (2007, December 1). More Web 2.0 stories, part two: Procter and Gamble embraces the wisdom of the Web for new product ideas. *The FASTForward Blog.* Retrieved December 28, 2007, from http://www.fastforwardblog.com/2007/12/01/more-web-20-stories-part-two-proctor-gamble-embraces-the-wisdom-of-the-web-for-new-product-ideas

Jana, R. (2006, June 27). American Apparel's virtual clothes. *Business Week.* Retrieved September 26, 2007, from http://www.businessweek.com/innovate/content/jun2006/id20060627_217800.htm

Jaokar, A. (2006, October 25). *Ajit Jaokar's mobile Web 2.0 blog: The widget widget Web.* Retrieved October 12, 2007, from http://www.web2journal.com/read/289798.htm

Kiley, D. (2007). Using bad news to a brand's advantage. *Business Week.* Retrieved November 14, 2007, from http://www.businessweek.com/the_thread/brandnewday

Kraft Foods. (2007). *Kraft foods goes digital to unveil more than 70 new products in first-ever virtual supermarket in Second Life.* Retrieved October 12, 2007, from http://www.kraft.com/mediacenter/country-press-releases/us/2007/us_pr_04192007.htm

Millions of Us. (2007). Retrieved October 12, 2007, from http://www.millionsofus.com

Moore, J. (2006). *Five reasons why…* Retrieved October 16, 2007, from http://brandautopsy.typepad.com/brandautopsy/2006/08/five_reasons_wh.html

Mootee, I. (2004). *High intensity marketing.* Canada: SA Press.

Rose, F. (2007). How Madison Avenue is wasting millions on a deserted Second Life. *Wired Magazine.* Retrieved November 12, 2007, from http://www.wired.com/techbiz/media/magazine/15-08/ff_sheep?currentpage=all

Second Life, R. E. P. E. R. E. S. (2007). Retrieved October 12, 2007, from http://www.reperes-secondlife.com

Smagg, C. (2007). *15 golden rules for Web 2.0.* Retrieved October 12, 2007, from http://visionary-marketing.wordpress.com/2007/07/03/web20

Steel, E. (2006). Web-page clocks and other “widgets” anchor new Internet strategy. *Wall Street Journal*. Retrieved November 12, 2007 from http://marcomm201.blogspot.com/2006/11/marketing-widgets-you-saw-it-here-first.html

Stephens, R. (2007). *P&G Web 2.0 success story.* Retrieved December 27, 2007, from http://www.rtodd.com/collaborage/2007/11/pg_web_20_success_story.html

Tapscott, D., & Williams, A. (2006). *Wikinomics: How mass collaboration changes everything.* New York: Portfolio.

Technorati. (2007). Retrieved December 28, 2007, from http://technorati.com

Terret, B. (2007). Brand in good Web 2.0 project shock. *Noisy Decent Graphics.* Retrieved December 28, 2007, from http://noisydecentgraphics.typepad.com/design/2007/11/brand-in-good-w.html

Vielle, T. (2007). IBM avatar commercial. *SL Universe.* Retrieved December 31, 2007, from http://sluniverse.com/php/vb/showthread.php?t=2797

Wallop, H. (2007). Cadbury plans Wispa revival. *Telegraph.* Retrieved November 28, 2007, from http://www.telegraph.co.uk/news/main.jhtml?xml=/news/2007/08/18/nwispa118.xml

Web 2.0. (2007). *Wikipedia.* Retrieved September 15, 2007, from http://en.wikipedia.org/wiki/Web_2.0, http://en.wikipedia.org/wiki/American_Express

This work was previously published in Social Software and Web 2.0 Technology Trends, edited by P. Deans, pp. 29-43, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 4.4

Applying Semantic Web Technologies to Car Repairs

Martin Bryan
CSW Group Ltd., UK

Jay Cousins
CSW Group Ltd., UK

ABSTRACT

Vehicle repair organizations, especially those involved in providing roadside assistance, have to be able to handle a wide range of vehicles produced by different manufacturers. Each manufacturer has its own vocabulary for describing components, faults, symptoms, etc, which is maintained in multiple languages. To search online resources to find repair information on vehicles anywhere within the European Single Market, the vocabularies used to describe different makes and models of vehicles need to be integrated. The European Commission MYCAREVENT research project brought together European vehicle manufacturers, vehicle repair organisations, diagnostic tool manufacturers and IT specialists, including Semantic Web technologists, to study how to link together the wide range of information sets they use to identify faults and repair vehicles. MYCAREVENT has shown that information sets can be integrated and accessed through a service portal by using an integrated vocabulary. The integrated vocabulary provides a 'shared language' for the project, a reference terminology to which the disparate terminologies of organisations participating in the project can be mapped. This lingua franca facilitates a single point of access to disparate sets of information.

CURRENT SITUATION

Repair scenarios for resolving a vehicle breakdown are varied, and can take place in a garage (repair by a qualified mechanic in a franchised or independent workshop) or by the roadside (repair by a qualified mechanic working for a Road Side Assistance (RSA) organisation, or a repair by a vehicle driver). For legal liability reasons, 'driver-assisted' repair scenarios only cover minor or temporary repairs of the type covered in owner's manuals, such as changing a vehicle wheel or a fuse.

In workshop scenarios, access to repair information may be provided through online access to repair information systems. Information may be provided publicly by a manufacturer for all users, or specifically to franchised dealers who are provided with

DOI: 10.4018/978-1-60566-066-0.ch002

access to information systems that are specific to the makes and models they retail.

Access to repair information in roadside scenarios is more complicated. A vehicle driver may not have access to the vehicle's owner's manual. In the context of a roadside repair by a mechanic working for an RSA, the mechanic might have access to repair information through a computer located in their van. RSA organisations, however, rely heavily on the detailed knowledge of their highly trained staff to diagnose faults without the aid of documentation. RSA mechanics aim to repair as many vehicles as possible at the roadside, but need to identify as early as possible if a car will need to be taken to a garage for repair. If the repair requires specialist equipment the RSA must be able to identify the nearest garage with suitable equipment that the car may be taken to for repair.

Fault diagnosis precedes vehicle repair in both repair scenarios. Details of the type of fault are ascertained at the point of contact with the customer, be this through direct conversation with the vehicle owner at a service centre, or by conversation through a call centre operator when a motorist initially reports a problem. When contact is made through a phone call it is important that call centre operators analyze the customer's situation in as much detail as possible. They have to be able to identify whether the problem is one that might be repairable at the roadside or whether a recovery vehicle is likely to be needed from the responses received to an ordered set of questions.

Customer contacts rarely lead to a detailed fault diagnosis because vehicle owners typically have insufficient knowledge of their vehicles to identify the cause of a problem. At best they can describe the symptoms produced by the fault and the conditions in which the symptoms manifest themselves (e.g. won't start when it is too cold). In many cases these descriptions can be used to identify the type of diagnostic tests that may have to be carried out before the cause of the problem can be identified.

PROBLEM STATEMENT

With the ever increasing use of electronics in vehicle components, identifying and correcting faults at the roadside or in an independent workshop is becoming a challenge. While the use of on-board diagnostic tools to report faults electronically via dashboard messages can assist mechanics, identifying the cause of a fault from such messages is not always a simple process. When faults are reported over the phone from remote locations sufficient diagnostic information may only be obtainable if the vehicle can be connected directly to the call centre information centre using tools such as personal digital assistants (PDAs) or mobile phones that can be connected to the vehicle's diagnostic ports.

A roadside assistance vehicle cannot contain the wiring schematics for all models of vehicles. Although, under European Union Block Exemption Regulation (European Commission, 2002), manufacturers provide access to all their repair information, repairers at the roadside are not always easily able to find the repair information that they need, particularly if this is related to a previously unreported fault, while physical and business constraints impose restrictions on the set of spare parts, tools, etc, that can be available in the workshop or repair van at any one time. Consequently, the following problem areas can be identified:

- Practical limitations exist on the level of information that can be provided in any repair context. There is variability in the amount and quality of information that is available to describe a fault and its associated symptoms and conditions in order to support fault diagnosis.

- Environmental variables such as geographical location, repair equipment, and spare part availability may combine to constrain the speed with which a repair can be affected, and determine the location at which the repair takes place.
- Logistics and supply chain management and facilitation can provide advance warning of required spare parts or repair equipment at the point of initial fault diagnosis, supporting decision processes such as the direction of a vehicle to an appropriate repair location or the triggering of inventory supply processes to pre-order required parts and arrange their delivery to the repair location.
- Maintenance of acceptable response times that meet customer expectations.

The MYCAREVENT project addresses these issues by facilitating the diagnosis of faults and the provision of repair information at the location where the fault is first described, be this in a workshop or at the roadside.

The MYCAREVENT project provides a single point of entry – a portal – through which a user can access services to support the description and diagnosis of a fault, and to search for and retrieve repair information from a variety of content providers. For this to be achievable, however, it must be possible to associate the terms used by the vehicle owner to describe the problem that has occurred with the terms used by the content provider to describe content or how to detect and solve the fault causing the problem. It should be noted that content can be of variable quality and scope – for example, repair information from a vehicle's manufacturer will typically apply to specific makes of vehicle, whereas information from third parties like technical specialists working for RSAs, or technical data for an automotive part or content from a third party information provider, may be more generic in application.

SOLUTION DESCRIPTION

The Mobility and Collaborative Work in European Vehicle Emergency Networks (MYCAREVENT) research project was sponsored by the IST (Information Society Technology) program of the European Commission. The 3-year project brought together leading manufacturers from the automotive sector, academic researchers and commercial IT suppliers to develop facilities for the provision of repair information to remote users in the automotive aftermarket. Remote access, for example a roadside repair, is enabled by the use of mobile services. Research focused on service development, process and organization management, e-business, communication networks and human-computer interaction.Work in MYCAREVENT has been organized in nine work packages:

WP 1: Project Management
WP 2: Use Case and Business Model
WP 3: Ontologies
WP 4: Mobile Communication
WP 5: Remote Services
WP 6: Service Portal
WP 7: Mobile Applications
WP 8: Training
WP 9: Demonstration and Dissemination

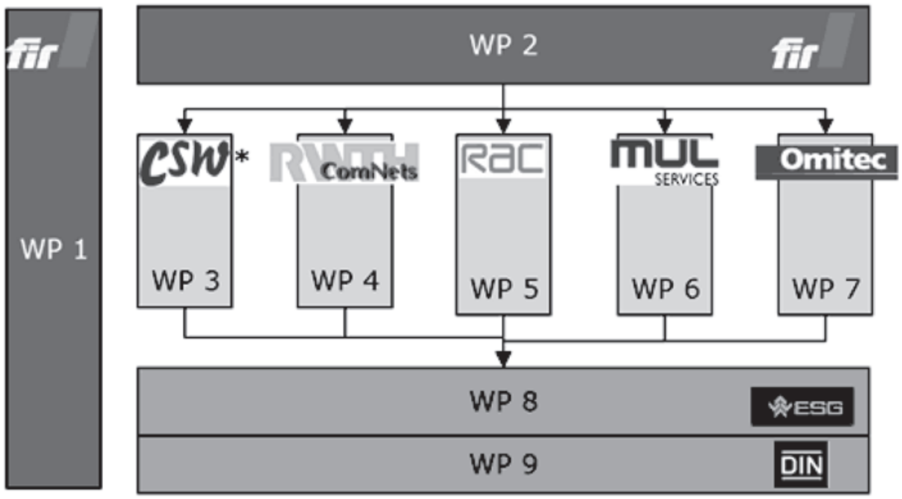
The relationships between these packages are illustrated in Figure 1.

WP1 supported project management, establishing and monitoring the administration and leadership of the MYCAREVENT project. It needs no further discussion here.

WP2 developed the fundamental business model and use cases that scope the project requirements and identify the customers, actors, processes, and the constraints on the legal and organisational environment within which the project solution operates.

The project scope can best be understood by looking at the project's three pilot scenarios, which demonstrate the functionality provided by the solutions and identify their targets:

Figure 1. MYCAREVENT work package relationships



- **Pilot I** was designed to demonstrate possible solutions for original equipment manufacturer (OEM) workshops and OEM roadside technicians who require remote access to the MYCAREVENT service portal to obtain instructions for specific repairs.
- **Pilot II** was designed to demonstrate how the concept of access to car repair information via the MYCAREVENT service portal could be extended to help mechanics working in independent workshops and roadside assistance services to identify faults.
- **Pilot III** was designed to demonstrate the concept of "Driver Self Help" in those scenarios where the driver can carry out simple repairs using advice provided by the MYCAREVENT service portal.

The ontology work package (WP3) defines the information structures that enable cross-system interoperability and the integration of content from disparate sources and heterogeneous databases.

The mobile communications and devices work package (WP4) provides a secure and reliable communication service between users (roadside assistants, drivers, and mechanics) and the service portal. These services are intended to enable the exchange of fault codes and repair information from remote locations, such as those required to carry out roadside repairs.

The remote services work package (WP5) allows a driver to search for self- help information, such as that provided in the owner's handbook, using standard Web browser software that resides on their smart-phone or PDA.

The service portal work package (WP6) defines the core project portal, the gateway for accessing repair information.

The mobile applications work package (WP7) allows the MYCAREVENT Service Portal to be used to deliver automated diagnostic tests to trained mechanics. This requires the application of additional access security and other middleware services within the portal interface.

The role of WP3 Ontology and WP6 Service Portal work packages is explained further in the following sub-sections.

OBJECTIVES

The MYCAREVENT Service Portal acts as a *gateway* to technical information on automotive diagnosis, repair and maintenance that is available from automotive manufacturers and independent organisations supporting the aftermarket.

To ensure high user acceptance, the MYCAREVENT work packages use innovative *state-of-the-art technologies* to find the 'right' information for user needs. To make this possible the service portal includes the following subsystems:

- *Core e-business infrastructure* for the flexible implementation of workflow and business processes.
- *Service data backbone* providing secure links to services as well as manufacturer and third party information repositories.
- An ontology-based *advanced query service (AQS)* for guided navigation through different data resources and terminologies.
- *Expert system hub* combining the capabilities of distributed (specialised) expert system nodes.
- *Authoring tools* for specific types of technical information, such as the interactive circuit diagrams (IACD) used to identify faults in electronic systems.

The remainder of this section explains how the ontology-based advance query service applies Semantic Web technologies to identify solutions to repair problems.

OVERVIEW

The MYCAREVENT Ontology work package was responsible for the development of the models, data structures and terminology sets used to support the work carried out by the service portal. The work package drew on the expertise of data modelling specialists, implementers and content providers (including OEM and RSA organisations) to build an agreed set of 'information artefacts' to be used across all MYCAREVENT services.The Ontology work package developed:

- A *Generic and integrated information reference model (GIIRM)* (MYCAREVENT, 2005), providing a high-level conceptual model of the MYCAREVENT mobile service world.
- A set of W3C XML Schemas derived from the GIIRM, which are used for the representation of data in messages, metadata and interfaces.
- Terminology for populating the GIIRM, enabling repair information, symptoms and faults to be described in a generalized way.
- A W3C Web ontology language (OWL) (McGuinness, 2004) ontology, derived from the GIIRM and the terminology, in which data sources can be registered for access by MYCAREVENT applications.

Details

Since the publication of Tim Berners-Lee's futuristic paper on The Semantic Web in *Scientific American* in May 2001 (Berners-Lee, 2001) the concepts that form the backbone of a system that can add semantics to Web resources has begun to form. As was pointed out in that paper:

For the Semantic Web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning.

The goal of the MYCAREVENT ontology work was to link together collections of information created by different vehicle manufacturers, component suppliers and repair organizations in such a way that we can use the collected information to conduct automated reasoning wherever possible.

The start point for the work package was the development of a formal model that could record the relationships between the information components used to identify and repair faults. This top-level model was designed to be generalized enough to apply to any type of repairable product.

Figure 2 shows a diagrammatic representation of the *MYCAREVENT Generic and Integrated Information Reference Model (GIIRM)* which was developed to manage the inter-relationship between information message components exchanged between information suppliers and the service portal. The diagram is expressed in the Object Management Group's Unified Modeling Language (UML) (Object Management Group, 2007).

The information required to populate the classes defined in this model are supplied by information providers in the form of Information Bundles that conform to the ISO/IEC 14662 Open-EDI Reference Model (ISO/IEC 14662, 2004). In this standard Information Bundles are defined as:

The formal description of the semantics of the recorded information to be exchanged by parties in the scenario of a business transaction. The Information Bundle models the semantic aspects of the business information. Information bundles are constructed using Semantic Components.

A unit of information unambiguously defined in the context of the business goal of the business transaction. A Semantic Component may be atomic or composed of other Semantic Components.

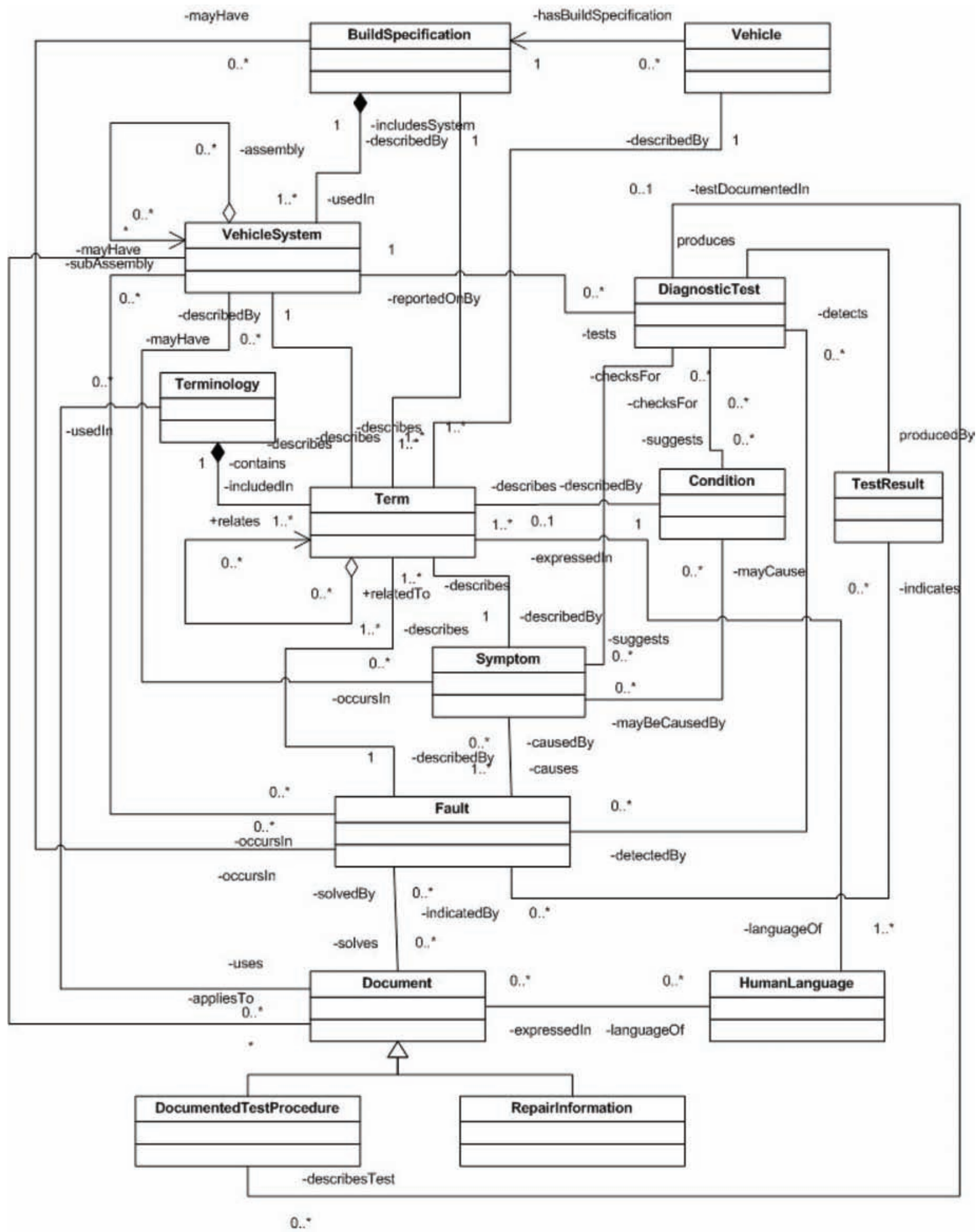
The model allows, therefore, for simple (i.e. 'atomic') and composite (i.e. 'non-atomic') attribute values, represented in the model using the concept of the 'representation class' which can be either an atomic datatype or a non-atomic composite data type as defined by the 'Naming and design principles' established in Part 5 of the ISO Metadata Registries (MDR) standard (ISO/IEC 11179-5, 2005). The GIIRM has foundations in abstract concepts and existing standardisation work. This design philosophy and layer of abstraction provides a generic model independent of any detail specific to implementation. It is, therefore, a platform-independent model.

To simplify the process of identifying relevant information sources, the model includes the concept of a Term. Terms can be used to describe a vehicle instance, its build specification, a system or subsystem used in the build specification, a condition under which a problem was detected, a symptom of the problem or a detected fault. Terms can be grouped into Terminologies that can be applied by different manufacturers within their documentation and information delivery systems.

Users of the MYCAREVENT Service Portal may or may not be aware of the terms used by manufacturers within their documentation. Users need to be able to enter terms that they are familiar with for describing problems, etc. The MYCAREVENT Advanced Query Service (AQS) needs to be aware of the relationships between the terms applied by a particular user community and the terms applied by a particular manufacturer. Terminology, and the mapping of terms in one terminology to terms in another terminology, is central to the AQS as it enables the querying of disparate information sources. The parameters of a search query can be established by a user using their preferred terminology, which may be manufacturer-specific or generalized using a MYCAREVENT specific term set. This search context is then passed to the AQS, which translates the MYCAREVENT terminology into the terminology used to describe the content that will be searched to retrieve information. This level of indirection allows disparate corpora of information to be searched using their own terminology. From an integration perspective, the AQS enables content integration using metadata describing repair information content created by OEMs or third parties, or through direct interface to a vehicle information source system.

The AQS has been developed using open-source technology to enable the latest develop-

Figure 2. MYCAREVENT generic and integrated information reference model



ments in Semantic Web technology to be adopted. It uses the Jena RDF triple store ("Jena", n.d.) to record OWL classes and individual occurrences of these classes. The Jena 2 Database Interface allows W3C Resource Description Framework (RDF) (Klyne, 2004) triples to be stored in MySQL, Oracle, PostgreSQL or Microsoft SQL databases, on both Linux and WindowsXP platforms.

An important feature of Jena 2 is support for different kinds of inference over RDF-based models (for RDFS, OWL, etc). Inference models are constructed by applying *reasoners* to models (Dickinson I, 2005). The statements deduced by the reasoner from the model can appear in the inferred model alongside the statements from the model itself. RDF Schema (RDFS) (Brickley, 2004) reasoning is directly available within Jena: for OWL an external reasoner needs to be linked to the Jena engine through a Reasoner Registry. The Pellet reasoner (Clark & Parsia, 2007) is used within MYCAREVENT to ensure that all inferred relationships are identified prior to searching. Jena includes an OWL Syntax Checker that can be used to check that OWL files are correctly formed.

Jena includes an implementation of the SPARQL query language (Prud'hommeaux, 2007) called ARQ. SPARQL has been developed as part of the World Wide Web Consortium (W3C) Semantic Web activity to provide a transportable technique for RDF data access that serves a similar purpose to the structured query language (SQL) used to access information held in a range of relational databases. The MYCAREVENT AQS generates SPARQL queries, based on the objects in the GIIRM, which are used to identify the concepts being referred to by terms entered by users. Because the query service is based on RDF it can query the contents of any OWL data property used to record information about an individual class member within the MYCAREVENT ontology, or any language-specific RDF label associated with a class or individual, irrespective of whether or not it is a term that has been specifically declared within a terminology.

SEARCHING AND RETRIEVING REPAIR INFORMATION

The key to the success of the MYCAREVENT portal is to allow users to ask questions using terms that they are familiar with and to use these questions to generate alternative versions of the question that OEM and other information provision systems associated with the portal can answer. The workflow steps used to establish a search query are described in the following sections.

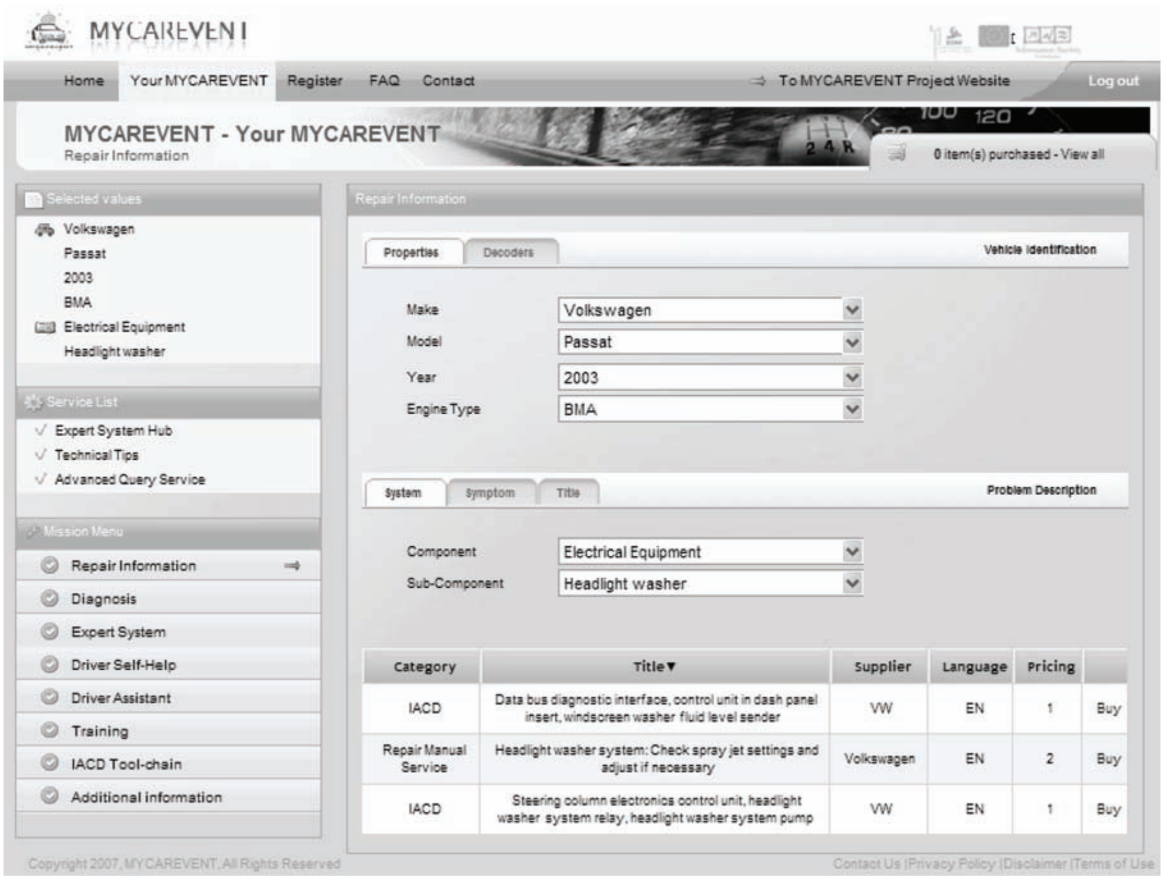
MYCAREVENT queries are implemented in a controlled, context sensitive, manner to provide guidance to users as they enter information into the service portal. Figure 3 shows the information components used to identify the type of vehicle to be repaired within the portal.

Users complete each field in turn. As they do so the options available to them in the next field are restricted using the ontology. So, for example, as soon as the user identifies the make of vehicle to be repaired, the set of options that can be used to complete the model field is reduced to the set of models appropriate for the entered make. Completing the model field restricts the range of years that can be entered in the Year field, selecting a year restricts the set of Series that can be selected, and so on.

For countries such as the UK where a vehicle registration number decoder is accessible it is possible to enter the vehicle registration number (VRN) into a field displayed under the Decoder tab and have the entered number return information that has been recorded about the year of manufacture, engine type, fuel type, etc, by the vehicle registration authority. Alternatively the manufacturer's vehicle identification number (VIN) can be used to automatically identify system components. Where a decoding service is not available each field in the vehicle description has to be completed in turn.

When as much information on the vehicle as is available has been recorded the user can be shown a list of available services, which can range

Figure 3. MYCAREVENT Repair Information Form



Category	Title▼	Supplier	Language	Pricing	
IACD	Data bus diagnostic interface, control unit in dash panel insert, windscreen washer fluid level sender	VW	EN	1	Buy
Repair Manual Service	Headlight washer system: Check spray jet settings and adjust if necessary	Volkswagen	EN	2	Buy
IACD	Steering column electronics control unit, headlight washer system relay, headlight washer system pump	VW	EN	1	Buy

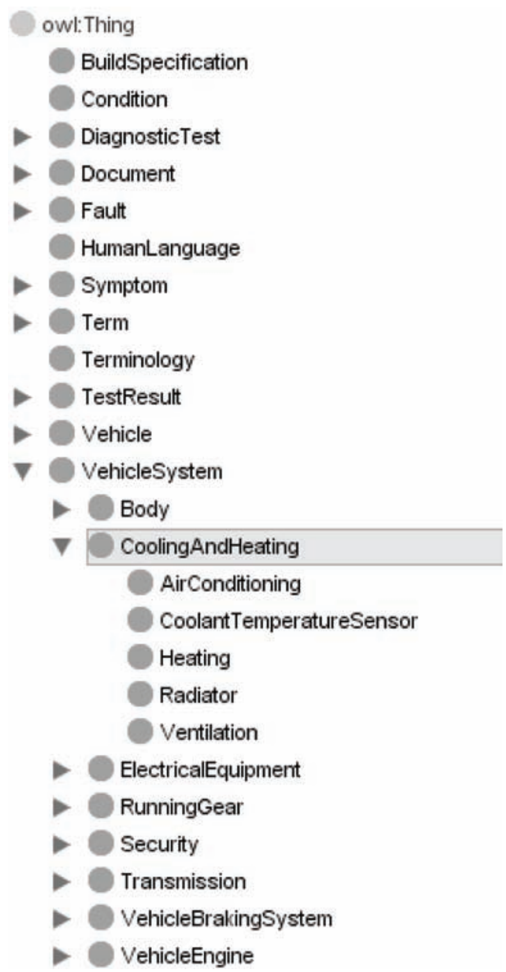
in complexity from a set of online technical tips through the use of the advanced querying system to identify relevant information resources to the use of an expert system to diagnose problems.

By reducing the set of selectable options at each stage to that recorded in the ontology we not only refine the search criteria but also reduce the likelihood that subsequent queries will be rejected. Only those services that are relevant to the type of vehicle to be repaired and the service that is being employed to repair it are offered to users so that, for example, users are not prompted to carry out tests using diagnostic equipment that is not fitted to the vehicle or accessible to the repairer, and are prompted to carry out all tests that must be performed to identify a specific fault.

THE ROLE OF THE MYCAREVENT ONTOLOGY

The ontology used by the AQS consists of a series of specializations of the concepts in the GIIRM. These concepts reflect the core classes and properties of the information required to describe and assert associations between faults, symptoms, conditions, vehicles, vehicle systems, diagnostic tests, terminology, repair information documents and human language. These classes are further sub-classed or specialised to refine concepts from the abstract and generic model level to a lower and more ‘concrete’ or ‘real world’ level that aligns with the business requirements. To illustrate, if business analysis shows that a particular information provider has circuit diagram, repair procedure,

Figure 4. Specialization of generic model classes

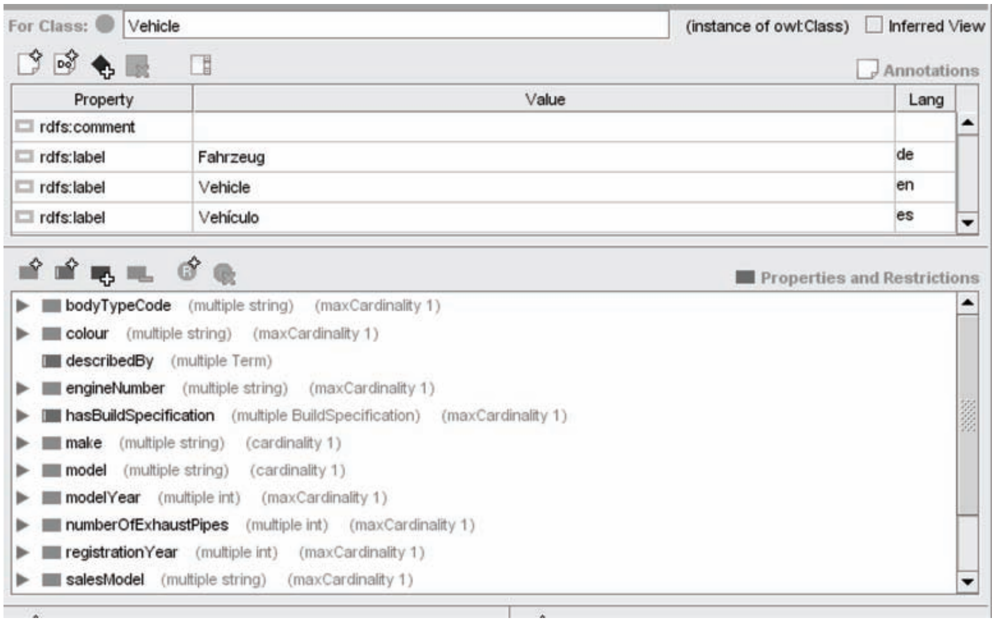


and owner manual types of repair information document, then the terminology used to describe document types is enhanced to include these new kinds of document, and within the ontology new document sub-classes are created to reflect these document types. The development process keeps the terminology and ontology in alignment with each other, and with the GIIRM which defines the base model and so ensures interoperability across the portal and database(s) accessed by it.

Figure 4 shows how the Vehicle System class is specialized into two levels of sub-class. Each class has associated with it multilingual labels, and a set of properties. Figure 5 shows the properties associated with the Vehicle class as they are displayed using the Protégé ontology editor used to maintain the MYCAREVENT ontology.

As a value is assigned to each of these properties within the MYCAREVENT System Identification form, the set of related properties that can be found by querying the triple store is reduced. Only those entries that are used in matched terms need to be displayed to users when they are required to select options for a currently empty field. The order in which responses are requested can be optimized to ensure that the minimum set of options is provided at each response point.

Figure 5. Properties of a vehicle



In MYCAREVENT users never see the underlying ontology while completing the basic Repair Information form shown in Figure 3. They do not need to browse the class tree, or know which properties they are dealing with. The ontology is simply used as a Web service by applications that need to request data from users, so that they can restrict the set of choices offered to users to that appropriate for the currently identified processing context, which is shown in the top right-hand window on the display. The role of the ontology is, therefore, to reduce information overload for users.

In later stages of the process, when accessing data using the Symptoms and Title tabs, users are prompted to enter keywords that are searched against lists of symptoms or the titles of documents. In these scenarios matches can only be made in those situations where the title contains relevant wording, or where the symptoms entered match symptoms recorded in one of the terminologies.

ALTERNATIVES

On-line access to information for vehicles is available from a variety of sources, including individual OEMs and third party automotive data suppliers. Subscription models vary, but typically documents are only available to paying subscribers and may require the installation of specialist software or training in the use of a specialized information retrieval system. For an individual user requiring access to information for a variety of vehicles, the availability of a single portal through which multiple sources of information can be accessed and queried is an attractive option. Rather than paying multiple subscriptions and having the burden of maintaining integration with multiple information access points, the portal provides a central point of access from which a user can search for and retrieve information.

Each manufacturer and third party information provider maintains their own vocabulary of terms for describing components, faults, symptoms, etc, in multiple human languages. If a trained mechanic knows which terms a manufacturer has applied to the fault they have identified then it is possible to search the manufacturer's database using a free text search. But if a mechanic is not familiar with the terms used by a specific manufacturer, there are unlikely to be any synonyms for them provided by information suppliers.

The issue of understanding the meaning of a particular terminology is not confined to metadata describing different types of vehicle systems, but applies to data recorded as codified values as well. On-board and plug-in diagnostic devices report faults using a set of codes. Some codes are internationally agreed, and will identify the same fault in a number of vehicles. Other codes are manufacturer specific. As with terminology, descriptions and names for fault codes may be maintained in a variety of human languages. For manufacturer specific codes the same code may be used to identify different problems in different makes of cars. Without access to a decoder that turns the codes into meaningful descriptions of faults for a specific vehicle build specification, knowing the code does not necessarily help the repairer.

Manufacturers of diagnostic tools need to ensure that their decoders are always up-to-date, providing the correct interpretation of codes for the latest models. Users of such tools need to be able to update their tools regularly to ensure the correct analysis of faults. While many OEMs already provide an up-to-date tool set for their vehicles through their BER portals, the extent of this is not consistent across manufacturers, and so a central point for updating information online provides one approach to addressing any gap which may arise between the release of new vehicle types and the availability of tools to diagnose their problems, other than those provided by the vehicle manufacturer.

By providing a central point for searching for repair information, issues such as the need to ensure this content is up-to-date, and legal liability for any claim arising from errors, need to be taken into account. For access across the single European market, irrespective of language and affiliation, however, centralized diagnostic services of the type that can be supplied using the MYCAREVENT service portal provide a viable alternative once agreement can be obtained from OEM and other content providers for making the information available online through the service portal.

COST AND BENEFITS

In 2004 the European Commission reported, in their European Competitiveness Report (European Commission, 2004a) that the automotive industry, as one of Europe's major industries, contributes about 6% to total European manufacturing employment and 7% to total manufacturing output. Total value added produced in the motor vehicles industry in the EU-15 in 2002 was roughly the same as in the US, some €114 billion.

With 209 million passenger cars in use in 2002 the European Union (EU-25) is by far the largest single market for cars in the world. It accounts for roughly 38% of all cars on major international markets. On average, four out of ten EU inhabitants own a car. According to the ACEA, the European Automobile Manufacturers Association, 15 million new passenger cars were registered in the EU and EFTA in 2006.

According to the EU report, it is expected that 90% of all future innovation in vehicle manufacturing will be driven by IT. This affects both the electronics dominated spheres of multimedia entertainment and navigation systems and the traditional mechanical components such as the chassis, body, engine or brakes. For instance, the percentage of electronics in the chassis is expected to increase from 12% to 40% in the next decade. Similar developments are expected for safety features, e.g. pedestrian protection, traction control, backward driving cameras, night-view display on the windscreen, sensor controlled brakes or fuel economy regulation. Product differentiation will be increasingly driven by electronics: for example, performance tuned variants of the same engine will differentiate suppliers. The value of electronic components in vehicles could rise from its current 20% today to 40% by 2015.

Since October 2002 motor vehicle distribution and servicing agreements within the EU have come under the new Block Exemption Regulation (BER). Under the new regulations repairers cannot be required to use original spare parts. Only if repair costs arise which are covered by the vehicle manufacturer, for example warranty work, free servicing and vehicle recall work, can the vehicle manufacturer insist on the use of original spare parts. Other than that, matching quality spare parts of other manufacturers or of independent suppliers can be used.

The automotive aftercare market had a turnover of around €84 billion per annum at the end of the 20th century; automotive replacement parts account for around half of this figure, some 45% of which is supplied by independent aftermarket (IAM) suppliers (European Commission, 2004b). The 210 million motorists in the EU spend on average €400 each per year and approximately €5,000 during the average vehicle lifetime on repair and maintenance.

Major service providers in the automotive industry are franchised dealers (120,000 dealers employing 1.5 million people in 1999) and independent repair shops (160,000 garages employing about 600,000 people). In addition, 18,000 roadside service vehicles fulfil 14 million missions a year.

By providing a single access point through which details of electronically available information can be searched, using generic Semantic Web technologies rather than manufacturer-specific solutions, the MYCAREVENT Service Portal simplifies and speeds up the task of finding information on how to repair vehicles with specific problems. With over a million potential customers it provides a cost-effective solution to information distribution.

By using ontologies to establish relationships between the terms used by vehicle owners and repairers to describe faults and the terms used by manufacturers to classify and describe faults, the MYCAREVENT Advanced Query Service can provide a more flexible solution to finding information, resulting in a higher likelihood of mechanics being able to find the information they need in a timely manner.

RISK ASSESSMENT

Each manufacturer produces thousands of models, each of which can have many build specifications. Model details and build specifications have to be defined prior to manufacture, but cannot be used within the service portal until the product is released. Unless the release date is known in advance, data relating to vehicle models, build specifications, system components, etc, cannot be added to the ontology at the time they are captured by the manufacturer, but need to be made available at the point when relevant documentation is released.

Manufacturers are naturally reluctant to maintain two sets of information, which could get out of step with one another. It must be possible to automatically convert information in local systems into the form that can be used by the AQS. As an alternative it should be possible to turn a query to the AQS into a query to the manufacturer's product database.

Another area of risk is in the level at which data is described, and the equivalence of terminology at different levels. A constraint on the MYCAREVENT ontology is that it currently only recognizes two levels in the system component hierarchy, system and sub-system. If a manufacturer uses a multi-level system hierarchy this needs to be flattened into a two level hierarchy for reference within the service portal. This means that entries at lower levels in the hierarchy have to become members of the appropriate higher level sub-system, thus restricting the level of refinement that can be applied to queries. This restriction is necessary because otherwise it would not be easy to convert AQS queries to queries that could be applied to manufacturer-developed services that can only handle two levels of querying.

Where diagnostic information is a requirement for identifying build specifications and associated repair information, obtaining the necessary information without access to OEM-provided diagnostic equipment can be a problem, especially in roadside breakdown scenarios. Unless the repairer can send appropriate information to the portal it will not be possible to retrieve relevant repair instructions. For this reason, other MYCAREVENT work packages have concentrated on how to get information from on-board diagnostic devices to a portal, diagnostic tool or expert system capable of identifying the cause of the problem.

Not all concepts are applicable to all makes, or to all models made by a specific manufacturer. Where a feature is specific to a particular manufacturer it is not to be expected that other content providers to the portal (be they manufacturers or third party information providers) will have equivalent terms in their terminology. If a user requests information on this subject for another make of vehicle the system will not be able to match the term. In such cases a number of strategies can be adopted to find appropriate terms, including:

- Identifying the term in the terminology of another manufacturer and informing the user that this term is manufacturer-specific
- Identifying the sub-system with which the term is associated by the originating manufacturer and offering a set of terms associated with the same sub-system that are used by the manufacturer of the vehicle being repaired
- Identifying other terms that include one or more of the words in the entered term, which may or may not identify higher-level concepts
- Identify other terms that include the largest identifiable substring of the entered term in compound nouns such as those used in German.

Expanding terminologies to cover all European languages, particularly agglutinative languages such as Finnish and Hungarian, where there are many compound words that could be derived from a term, will make identifying potential matches much harder. For such languages it will be vital to be able to define relationships between alternative references to a term within terminologies.

Identifying which terms have significance in which documents is another problem area. Unless the sub-systems that documents refer to are unambiguously recorded in either the data or metadata associated with a document, and the faults that can be solved using a document are recorded in the ontology or manufacturer's information base, refining queries down to the level of identifying documents that are specific to a particular problem with a given sub-system will not be possible. The best that can be achieved is to identify the set of documents that refer to a particular sub-system and allow users to determine from metadata describing the type of document, etc, whether it may be suitable for solving the problem.

Relying on user selection of suitable documents introduces another risk. Manufacturers want to be paid for preparing and supplying data. Users only want to pay for information that they know will solve their problem more efficiently than alternative solutions. If the cost of information is too high users will not risk purchasing something that may not solve the problem. If the cost of information is too low manufacturers, or third party documenters, supplying the information will not be able to recover the cost of preparing the information for distribution. Because of legal liability concerns, and the requirements by the BER, manufacturers are reluctant to supply information units which do not contain all the legally required warnings, safety notices, etc, that can apply to the repair scenario.

OEMs want to supply units of information that are known to contain all relevant details for the sub-system(s) that are connected with the fault.

The rate at which documents change is also a concern to information suppliers. It is not possible to maintain an up-to-date repository that includes all repair information generated by all vehicle manufacturers, even if an efficient enough content management system was available to store and access them. The best a service portal can expect is to receive metadata about which documents are available for which sub-systems, and the roles those documents serve. If the metadata supplied with each document fails to identify the type of faults that the document can help to correct, it will not be possible to associate faults with the documents that can be used to repair them. These risks were identified and confirmed by the OEMs involved in the project, and in some cases would prevent them from being able to integrate their content with MYCAREVENT.

The key strength of using an ontology-based approach to service portal management is that it reduces information overload on users, who otherwise would find it difficult to find their way through the maze of specifications and information types supplied by different manufacturers. By minimizing the set of options provided at each stage in the process, the MYCAREVENT advanced query service makes it possible to identify information resources provided by a range of manufacturers through a single reference point.

Until manufacturers are able to provide information as to which faults can be solved using which documents a weakness of the service portal is that it will necessarily rely on users making the final choice between a range of information resources that cover a particular component. Where diagnostic tests are available their results can be used to narrow down the range of possibilities. When diagnostics are implemented an associated problem is that of identifying the relationships between symptoms, the conditions they can occur under and the diagnostic test results that can be used to identify specific faults. At present this information is generally not available from manufacturers. Until it is systematically recorded the efficient identification of faults within the portal will be difficult.

The recording of the symptoms reported when a particular fault has occurred is, however, also an opportunity for the service portal. By recording the symptoms reported by users, the conditions under which they occur and the fault that was eventually identified as the cause of the problem within the service portal it should become possible, over time, to generate statistics that can be used to predict the likelihood of a particular fault being the cause of an exhibited symptom.

The size of the European automotive industry is another major risk. If all vehicle suppliers adopted the system, and it covered all cars in current production, the potential user community could be as many as 2,000,000 people. Several portals would be required to cope with such a load. To keep the systems synchronised it would be necessary to adopt a time-controlled update system, with updates being scheduled for early in the morning when system use is low. A separate system would have to be assigned the task of receiving information from manufacturers and accumulating them ready to carry out a single daily update of online portals. The downside of these process integration issues would be that any changes made to documentation, build specifications, etc, would not be available on the day they were recorded by the manufacturer. This risk can be managed using Trading Partner Agreements and Service Level Agreements between the portal and the information providers, following standard practices for managing business relationships.

Expanding the proposed system to cover all vehicles would also require significant expansion of system functionality, because commercial vehicles have a much wider range of build specifications. One vehicle manufacturer reported to have 93,000 build specifications for trucks. Part of the reason for this is that there are more distinguishing features, such as type of steering, number of axles, body type, couplings, etc, used to define the build of a commercial vehicle. For such vehicles it becomes important to use the unique vehicle identification number (VIN) rather than its vehicle registration number (VRN) to obtain accurate details of the build specification.

FUTURE RESEARCH DIRECTIONS

As Tim Berners-Lee pointed out in his seminal paper on the Semantic Web (Berners-Lee, 2001):

Traditional knowledge-representation systems typically have been centralized, requiring everyone to share exactly the same definition of common concepts such as "parent" or "vehicle". But central control is stifling, and increasing the size and scope of such a system rapidly becomes unmanageable.

Two important technologies for developing the Semantic Web are already in place: eXtensible Markup Language (XML) and the Resource Description Framework (RDF). XML lets everyone create their own tags—hidden labels such as <author> or <title> that annotate Web pages or sections of text on a page. ... Meaning is expressed by RDF, which encodes it in sets of triples, each triple being rather like the subject, verb and object of an elementary sentence. These triples can be written using XML tags.

An ontology is a document or file that formally defines the relations among terms. ... We can express a large number of relations among entities by assigning properties to classes and allowing sub-classes to inherit such properties. ... Ontologies can enhance the functioning of the Web in many ways. They can be used in a simple fashion to improve the accuracy of Web searches—the search program can look for only those pages that refer to a precise concept instead of all the ones using ambiguous keywords. More advanced applications will use ontologies to relate the information on a page to the associated knowledge structures and inference rules.

The real power of the Semantic Web will be realized when people create many programs that collect Web content from diverse sources, process the information and exchange the results with other programs. ... The Semantic Web, in naming every concept simply by a URI, lets anyone express new concepts that they invent with minimal effort. Its unifying logical language will enable these concepts to be progressively linked into a universal Web.

The need has increased for shared semantics and a Web of data and information derived from it. One major driver has been e-science. For example, life sciences research demands the integration of diverse and heterogeneous data sets that originate from distinct communities of scientists in separate subfields. . . . The need to understand systems across ranges of scale and distribution is evident everywhere in science and presents a pressing requirement for data and information integration.

The need to integrate data across a range of distributed systems is by no means restricted to the scientific community. It is a fundamental characteristic of any e-business scenario that needs to be linked to back-office systems or to systems, such as those used for payment management, run by other companies. OWL allows the UML-based modelling techniques that are fundamental to the design and maintenance of back-office systems to be swiftly integrated with the XML-based messaging approach that has been widely adopted for inter-system communication.

The trend towards globalisation that characterises today's business environment is established and set to continue. Increasingly, demands are placed upon organisations to integrate information from diverse sources and to deliver new products and value propositions in narrower timescales. To meet these demands IT organisations need to evolve towards loosely coupled systems where services can be assembled to support the execution of business processes in a flexible way. A service oriented architecture (SOA) is not the only answer, though – for an SOA to be effective, a common view on to the data of the organisation needs to be available, so that data can be provided when and where needed to the processes consuming that data.

RDF-based data integration has a lot to offer because it provides a way to access information held in disparate systems without imposing a new structure on source data. If metadata is available, or can be generated, a metadata-based approach provides a framework structuring, processing and querying information sources.

The use of OWL and Semantic Web technologies moves us beyond simple metadata to structures where additional rules that determine the logical meaning of the data can be layered on top of existing data by the application of an ontology. Ontologies allows rules to be specified which can be reasoned over using methodologies such as description logics, allowing inferred models of data to be constructed. Not only is the explicit meaning of the data recorded, but the implicit meaning of the data can also be inferred and exposed by applying such rules to data.

The use of ontology-based approaches is another step along the path from the computer to the conceptual world used by humans. Now programming has progressed from binary assembler languages to 4th generation object-oriented paradigms, ontologies allow knowledge to be encoded as data structures in a way that reflects the understanding and semantics of the domain and of human users. Data can be modelled in a manner that is more intuitive and conceptually closer to the way humans think. Ontologies allow humans to use their own terminology to model the domain in a way that reflects how they understand it and speak about it. Furthermore, they can now encode knowledge and logic about the data structure, moving it out of application logic.

In this chapter we have said nothing about how OWL's limited set of description logic (DL) rules can be used to constrain the values assigned to ontology properties or to infer membership of a class from the presence or absence of property values. MYCAREVENT has not identified any points at which rules more complex than those needed to constrain cardinality or to ensure that all object properties are members of a given class or set of classes need to be applied to repair scenarios. But in many business scenarios more complex rules, including access control rules and permissions management, will be needed to ensure that business constraints can be met. The presence of an alternative, expert-system based approach to rule definition and application within MYCAREVENT has meant that the service portal team has not fully investigated the role that inferencing rules might play in the development of e-business applications, though a number of possibilities have been identified, including ones related to digital rights management and skill-based access control to information resources.

Work began in 2006 on a W3C Rule Interchange Format (RIF), an attempt to support and interoperate across a variety of rule-based formats. RIF (see www.w3.org/2005/rules for details) will eventually address the plethora of rule-based formalisms: Horn-clause logics, higher-order logics, production systems, and so on. Initially, however, the Phase 1 rule semantics will be essentially Horn Logic, a well-studied sublanguage of first-order logic which is the basis of logic programming. Among the deliverables scheduled from the RIF Working Group for the end of 2007 is:

A W3C Recommendation on using this rule interchange format in combination with OWL. This document is needed to help show implementers and *advanced users how these technologies overlap and the advantages and limitations around using them together. This document must clearly state which features of OWL can be mapped to (or otherwise interoperate with) Phase 1 rules and which cannot, and software using this mapping must be demonstrated during interoperability testing. The document may also discuss rule language extensions to cover the excluded OWL features.*

A second phase, scheduled to be completed in 2009, will extend rule processing to provide full first-order logic, negation, scoped negation-as-failure and locally closed worlds.

Until RIF tools are readily available, OWL users will have to make use of proposals such as that for a Semantic Web Rule Language (SWRL) (Horrocks, 2004) that extends OWL's built-in set of simple rule axioms to include Horn-like rules. While MYCAREVENT has not currently identified any rules it needs to deploy within the AQS which cannot be implemented using predefined SPARQL queries, it is anticipated that there will be other applications based on the GIIRM for which more complex queries of the type provided by SWRL may be needed. It will be interesting to see, as RIF develops, whether the additional functionality offered by adopting Horn-clause or higher-order logics provides a simpler solution to the type of reasoning currently being performed by the expert system currently used to identify the causes of problems within MYCAREVENT.

CONCLUSION

The automotive market has become one of the most important and complex industries in the EU, due to the rapid development and change in electronics, electrics, software and hardware. Economically, it is a major contributor to the EC economy, accounting for circa 6% of total European manufacturing employment and 7% of total manufacturing output.

Due to the EU Block Exemption Regulation, service providers have the right to access different kinds of repair information, training material and tools.

The MYCAREVENT project gathered partners from across Europe to establish a model of excellence leveraging innovative applications and state-of-the-art technologies, to offer a way for making the market more transparent, competitive and lucrative. It developed and implemented new applications and services which could be seamlessly and securely accessed by mobile devices deploying Semantic Web technologies. These tools allow us to provide manufacturer-specific car repair information that matches problems identified by Off/On-Board-Diagnostic systems.

Breakdown information is presented in different languages. Mobile workers in different countries can interact with service portals of independent service suppliers as well as those of car manufacturers. Using the MYCAREVENT Service Portal it becomes possible to provide a single point of access to information for any make of car, so ensuring that any car manufactured in Europe can be repaired in any European workshop or by any European roadside assistance organisation, irrespective of the preferred language of the owner or the mechanic.

REFERENCES

Berners-Lee, T. (2001). The Semantic Web. *Scientific American*, (May): 2001.

Brickley, D., & Guha, R. (2004). *RDF Vocabulary Description Language 1.0: RDF Schema*, W3C.

Clark & Parsia LLC. (2007). *Pellet: The Open Source OWL DL Reasoner*.

Deliverable, M. Y. C. A. R. E. V. E. N. T. D3.2 (2005) *Generic and Integrated Information Reference Model*. L.3.2_Generic_and_Integrated_Information_Reference_Model_DT_v01.00.pdf

Dickinson, I. (2005). *HOWTO use Jena with an external DIG reasoner.*

European Commission. (2002). *EC Regulation 1400/2002; Application of Article 81(3) of the Treaty to categories of vertical agreements and concerted practices in the motor vehicle sector.*

European Commission. (2004). The European Automotive Industry: Competitiveness, Challenges, and Future Strategies, *European Competitiveness Report 2004*,

European Commission. (2004). *Proposal for a Directive of the European Parliament and of the Council amending Directive 98/71/EC on the Legal Protection of Designs*. Extended Impact Assessment.

Horrocks, I., et al. (2004). *SWRL: A Semantic Web Rule Language Combining OWL and RuleML*.

ISO/IEC 11179-5:2005. *Information technology -- Metadata registries (MDR)* -- *Part 5: Naming and identification principles.*

ISO/IEC 14662:2004. *Information technology -- Open-edi reference model*.

Jena – A Semantic Web Framework for Java.

Klyne, G., & Carroll, J. (2004). *Resource Description Framework (RDF): Concepts and Abstract Syntax*, W3C.

McGuinness, D., & van Harmelen, F. (2004). *OWL Web Ontology Language Overview*, W3C.

Object Management Group. (2007) *Unified Modeling Language (UML)*.

Prud'hommeaux, E., & Seaborne, A. (2007) *SPARQL Query Language for RDF*, W3C.

Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The Semantic Web Revisited. *IEEE Intelligent Systems*, *21*, 96–101. doi:10.1109/MIS.2006.62

This work was previously published in Semantic Web for Business: Cases and Applications, edited by R. García, pp. 31-49, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 4.5

The Web Strategy Development in the Automotive Sector[1]

Massimo Memmola
Catholic University, Italy

Alessandra Tzannis
Catholic University, Italy

ABSTRACT

Especially in recent years, a transformation is ongoing: the Web, besides being a means of information sharing (internal-external), becomes a powerful tool for saving costs, reducing the distribution structure, initiating distance transactions, and ever more, becomes a mechanism of integration with the external environment and a catalyst of experiences for all stakeholder. Starting from the identification of the key elements, potentialities, and of the impact of the Internet on firms' performance, competitiveness, effectiveness, and efficiency, this chapter is focused on the changes in the automotive sector due to the integration between business strategy and Web strategy. Therefore, starting from the consideration of a clear identification and subsequent sharing need of strategic goals, a research work will be presented exploring, on the basis of an interpretative model, the Internet potential in the automotive sector in order to achieve the identification of an optimal path definition and development of Web strategy. This objective will be developed through a desk analysis focused on the strategic positioning of the current businesses in the automotive sector (i.e., complexity evaluation of the presence on the Internet, strategic architecture, quality, and effectiveness of this presence).

DOI: 10.4018/978-1-60566-024-0.ch009

INTRODUCTION

I need a new car!

Centy, as I affectionately call my old car, is gasping its last breaths. The "poor thing" has really had it, but then, it is really old; 10 years have passed from that happy day it made its entry in our family.

However, I am not so sure I like the idea of searching for a worthy successor. I am not an automobile fan, and my knowledge of them is rather limited. I am, however, sensitive to environmental problems, and I try my best to reduce the impact of my own person on the ecosystem as much as I can, considering its balance has already been seriously damaged.

What bothers me most is the idea of the long search ahead of me. I do not have any particularly difficult requests, but I want to consider such an important purchase very carefully, avoiding hurried decisions. I need a sensible family vehicle, with a good balance between performance, consumption, and above all cost. I need to acquire information, but just the idea of going from one dealer to another puts me in a bad mood. I can not stand having to waste time on this, and above all, having to pollute for a car I do not even own yet!

Also, interaction with the dealers is more often than not rather uncomfortable. At times—especially the more prestigious brand dealers—they treat you with a presumptuousness as if you were a nuisance, a matter that needs to be settled as quickly as possible. Other times, for the more sporty brands, they give you so many figures and acronyms, that they make you feel like "you did not do your homework" as in your school days. Lastly, with some there is feeling of a lack of transparency, something like "tell 'm everything but not really the truth."

Of course I do not want to demonize all car dealers. These are only impressions, and on top of that, only personal ones. But I do believe that, at least once, we have all felt the uneasiness I described above, either completely or in part.

I am holding the yellow pages in my hands, ready to copy my city's car dealers addresses on a piece of paper, but then a question just comes to mind, popping up almost unconsciously: "Why not use the Internet?"

I do most of my work nowadays through the Web: I keep in touch with my friends scattered around the world, I keep my bank account, book my holidays, buy music, and lately I have also taken care of my physical well-being using the telemedicine services offered by my city's health department. But then I say to myself: "But for a car it is different!" But is it really? Why would the Internet not have brought about the same changes in life style, in the way of thinking, of buying, in the automotive sector as it did in other areas?

I start my Web search by typing the word "car" into the search engine. Promptly a long list appears with sites specialized in online car sales. I am starting to question my beliefs. Some sites only offer general information, deferring the sale to a moment of real interaction. There are some sites though that are true virtual car dealerships, and allow you to get through the whole buying procedure on their site, with even a home delivery service of the newly purchased car. So after all I would not spend too much time, as I did with Centy.

However, I do not feel ready for an online purchase. I have identified the model that seems right for me, but I need some extra information. Through the faithful search engine, I reach the Web site of the car manufacturer. I am welcomed into a very sophisticated ambient. I click on "Product Range" and I find myself in a virtual car dealer showroom.

The models are well-presented, with clear 3D images. The technical features provided are exhaustive for each model. I discover the "car configurator," a very interesting tool that allows you to configure your car by choosing the color, the interior, the optional features, and, once the final price has been determined, to have access to another service that allows one to request personalized financial options.

It is also possible to print out a customized brochure showing the specific car model with the chosen color, interior, and optional extras. I find out that the site offers me the possibility to not only get to know my local dealer, but also book a test-drive at my pleasure.

And there is more! The site offers a whole series of services that I would never have imagined that go well beyond the usual general information about the manufacturer. I realize that the contents vary according to the phase of the purchasing process: when the customer is in the process of choosing a car, the customer can benefit from tools that compare the models and optional extras, and that allow people to create the car that is closest to their expectations. Afterwards, some gadgets that make the car look closer and more real, can be downloaded (pictures and videos). Once the purchase has been completed, the customer can monitor the delivery process to see when it will arrive at the customer's door and after that, book the assistance and maintenance services, activating an SMS reminder service.

Often the Web becomes an occasion to create a community of customers of the same brand. Various services have been created to this end, from the normal forum to the more sophisticated blog, that allow an almost one-to-one interaction with the manufacturer, offering the possibility to interact in various ways and degrees with the managers, and to know the reasons behind some project choices, or to be informed about new models before they are launched on the market.

I also find an online magazine that I can virtually read and that, very coincidentally, not only contains car related articles, but "talks" about environment, style, and sports. I realize I am understanding, really entering the brand's life style. Once I own the car, through the Web site, I can request a loyalty card that, apart from working as a credit card on the main international circuits, allows me to obtain discounts not only for the car's assistance (i.e., maintenance, repairs etc.), but also in many shops in line with the brand's style and features.

At this point I am ready to leave the virtual world to get back to the real one. I can calmly think about my visit to the dealer where I booked my test drive. I have got all the information I need and I have made my choice. I could even make the purchase online. I have seen that some sites offer this possibility, but entering the dealer's showroom will allow me to satisfy the emotional value of the purchase, that is, to hold tight the steering wheel of the long-desired vehicle, breathe in the "smell of new," touch the dashboard. The Internet is not able to satisfy these senses, at least not yet!

The lines above are obviously fantasy, but they are not science fiction, as they describe a situation that is very likely in the light of the information and service contents that we have found on the Web analyzing the companies' Web sites of the automotive sector during our research work. We had to use our creativity simply because we imagined an ideal Web site that should hold all the solutions, information, and services that in reality the car manufacturing companies activate only partially in their portals.

It has rightly been observed that the Web represents "a new space, a new territory, made of computers, connections, software and above all, information…inhabited by individuals, companies, and by organizations; a place made of games, commerce and exchange" (Porter, 2001).

Comparing the results of our research with similar ones in other branches, such as banking (Frigerio, in this book), tourism (Yaobin, Zhaohua & Bin, 2007), retail business (Duke, Chul, Sang-II & Soung, 2006), and the health sector (Baraldi & Memmola, 2007), it seems that the automotive sector is, generally speaking, running far behind the others. The feeling that we get is that the companies belonging to this branch, even though they have been developing a "public presence" on the Web for quite a time, have hardly exploited its potential, limiting themselves in various cases to a low profile Web strategy, structured as a source of information rather than as a service. The Internet has therefore played a role up to now that is not dissimilar from any other media that allows a company to develop a one-way communication with the customers, and more generally, with the company's stakeholders.

This delay and lack of sensitivity toward "evoluted usage forms" of the Internet's potentialities can be partly justified by the strong emotional value linked to the purchase of a car, leading the customer to the dealer's showroom to touch the product that the customer is going to buy, to verify its qualities, thereby trusting tactile senses to guide and reassure. And it must be said that past attempts to develop forms of e-commerce in the automotive field, although they requested the purchase to be finalized at the dealer's, did not have much success.

Over the past years however, this trend has been changing dramatically. According to a research carried out by McKinsey (2003), during the next 10 years we will see the third revolution in the automotive sector, after the creation of the Ford mass production factory and the lean production of the Toyota Production System. Customers of the automotive branch, according to the research, will be expecting "hightest performance" for the same price, generating strong pressure on cost reduction and innovation capacity in the automotive companies. These factors will, without a doubt, lead to a whole new set-up of these companies' supply chain.

In fact, various companies are starting to grasp the idea that the Web can become an essential component in the so-called "low cost car strategy" to which many companies (e.g., Renault, Tata, General Motors, Fiat, Volkswagen, and Toyota) are decidedly heading. Through the Web the distribution network, which currently represents an important part of a car's final price, can be removed or reduced to a minimum.

Not only! The Web becomes the privileged instrument of interaction with the so-called "iPod generation," who through the company's portal or a specific product site, can contribute to the various steps of the product's project or image definition (e.g., accessories, internal and external graphic features, etc.). In this sense Fiat's experience with its Web site www.500wantsyou.com is surely significant.

Finally the Web can play a vital role in supporting customer relationship management (CRM) policies, or in the broadening of the product system with a series of complementary services linked to the automotive world (e.g., assistance, insurance, loans, etc.).

This chapter presents the results of a complex research aiming to define in what measure and with which modes the companies of the automotive sector are using the potentialities of the Internet to create value for their main stakeholders.

POTENTIALS OF THE INTERNET FOR THE AUTOMOTIVE SECTOR

Why does the Internet continue to receive so much attention? Why are the Internet and its potentialities always being discussed? What makes the Internet a better technology than so many others that preceded it? Does the Internet really change the way "to do business"? Does the Internet create new business models or does it change the existing ones?

Quite a few years after the boom of the New Economy and the subsequent enlivenment of the managerial literature at the turn of the century, perhaps the question will receive more sensible answers and above all, be sustained by a longer experience. The Internet surely is a surprising technology through which it is possible to communicate, interact, sell or distribute products and services, and above all, create a powerful tool to bind customers through communities in the form of forums, chat groups, blogs, user groups, and so forth.

What mainly distinguishes the Internet from previous technologies is that it sums up all their features and all their potentials in one low-cost standard, that is, through the Web people can watch TV, listen to the radio, and talk over the telephone. Therefore it may make more sense to talk about Media strategy rather than Web strategy: to underline the necessity, from a company's viewpoint, to govern the various interactive channels, on and off the Web, in a harmonious and coordinated way.

The Internet, therefore, changes or can change the way to do business. It has rightly been observed that the brick and mortar companies, that is, which existed before or which were not founded in function of the Internet, have to develop their own business model in order to seize opportunities, taking into account the limits (but also the strong points) of the current situation.

The most important thing is not to succumb to the charms of the Internet, which is only a tool after all, a dependent variable of the model of analysis of the company's governance. It is therefore necessary to align the choices that have been made within this scope with those which have determined the business strategy, selecting the Internet's best potentialities that better fit the purpose, thereby carefully studying realization time and modes.

The Internet as a Media and a Geographical Distance Reduction Tool

The Web facilitates exchange based relations, overcoming the obstacles of time and space. More than 10 years since the diffusion of the Internet, this may seem very obvious. According to the authors, this is the main reason behind its success though. The Internet is certainly not the only media available: the alternatives are the press, television, and radio. As said above, the Internet is extraordinary because it summarizes them all; at the same time developing interaction and the possibility to realize a communication process that is not just one way.

As in other areas, also in the automotive sector this potential allows re-examination of the development logic of the supply chain (e-procurement), the distribution processes (relations with the dealers), and of the relationship with and assistance to the customer. Therefore, it does not matter if the dealer is located far from the manufacturer; through the company's Intranet[2] the dealer will be able to see if a particular model is available or, alternatively, how long the production will take. A small size, niche manufacturer will be able to offer cars for sale to potential customers all over the world, even if the manufacturer has no dealerships at all.

The Internet as a Means to Rationalize Time

The Internet enables efficient time use, as a result of its power to reduce or dilate it. The first distinctive feature appears when looking for information about a particular car, the Web offers the possibility to find it quite quickly. In the same way the Internet allows a time dilation realization if services, streamlining the transaction, are implemented (e.g., booking of test-drive, of assistance services, etc.).

Web Externalities at the Base of the Internet

Technologies or products present Web externalities if their value is proportional to the amount of users, that is, the more the users, the higher the value (Katz & Shapiro, 1986). In order to grasp the meaning of these words, think of the usefulness of a telephone that can be used to talk to only one person. The value of such a phone would most certainly be lower than that of a phone which has a potential for worldwide communication. Clearly, the more people connected to the telephone system, the higher its value will be for its users. The Internet represents, in this sense, clear proof of this: the higher the number of people connected, the higher is the value of the Net.

The authors have had a clear demonstration of this potential, participating at a discussion forum about a new small Italian car manufacturing company. The continuously increasing number of participants and their interactions is progressively creating a pressure group aimed at demanding the satisfaction of the customers' wishes on the manufacturer's part.

The Internet as a Distribution Channel

The Internet, in general, operates as a distribution channel for all products with high information content that can be digitized (e.g., software, music, videos, images, plane tickets, various services related to banking, insurances, and even healthcare).

In the automotive sector, the Internet could be an interesting substitute for the existing distribution channels. Through the Web, low-cost cars could be sold, making it necessary to reduce the cost of the value creation channel as much as possible. And there is more! The Web could act as the main distribution channel for all those companies that have difficulties with traditional approaches, for example, small size or niche manufacturers or companies of a collateral sector such as car tuning.

The Internet as a Tool to Reduce Asymmetrical Information

Information is asymmetrical when, in a transaction, one party holds information that is relevant to the same transaction that the counterparty does not have. A classic example, printed in all economics books, is that of the transaction of a used car: here the asymmetry lies in the fact that the buyer does not only ignore the real condition of the car, but often does not even have the means to know if the asking price is congruous with the market value, or to evaluate the technical features. The Web offers such information, for example, the prices asked by different dealers for similar models, or consumers' associations that certify the quality of the product one intends to buy.

The reduction of these asymmetries can work both ways. Through the Web a car manufacturer can get to know about the expectations, tastes, and opinions of its customers. A very good example is Fiat's blog *quellichebravo* and it's Web site www.500wantsyou.com, gathering information essential for the development of new products.

The Internet as a Transaction Cost Reduction Tool

In the automotive sector too, the Internet could reduce the costs of the transaction process, as a direct consequence of the above-analyzed elements. Often in fact, companies have to do research work to find suppliers that offer the raw materials needed for the requested quality standards; the customers have to acquire information on the products' features, on the prices, and, in general, on the alternatives offered by the market; the dealers have to acquire information on the customer's financial situation, or interact with the headquarters to evaluate availability and delivery terms of each model. In the above activities, surely the majority of the information necessary to the transaction's completion can be easily obtained on the Web.

The Internet as a Tool to Support Corporate Governance and Personnel Training Courses

Finally, the Internet's technology is revealed as an irreplaceable tool sustaining the following activities:

- The company's analysis, control, and governance (business intelligence systems, ERP systems that are ever more oriented towards Web-based applications).
- Distance training of operators, with huge cost savings and significant benefits regarding the service level.
- Management of the company's information service and integration (through Intranet and management networks) of the company's divisions, organizational units, production sites, and dealerships.

OBJECTIVES AND METHODOLOGY OF THE RESEARCH

Objectives of the Research

This research work, carried out through a general survey of the Web sites of the main companies of the automotive sector, aimed at verifying how and to what extent these companies are currently exulting the potentials offered by the Internet to create value for the major stakeholders. In particular, the aim was to collect data that would allow us to evaluate:

- The strategic positioning—obviously on the Internet—of the companies of the automotive sector.
- What results have been achieved, or rather how much of the potential offered by the technology, is being used by the companies to define their own public presence on the Web.
- If and to what degree the strategic initiatives are successful in terms of visibility and registered traffic.

The Methodology of the Research

R. M. Grant (1998) defines the main objective of a company's strategy as the aim of "guiding the management's decisions to excellent results through the search of a competitive advantage; it is simultaneously a means of communication as of co-ordination inside companies."

In a correlated way, the Web strategy could be defined as that systemic set of decisions and actions aimed at determining in what measure and through which modes the company's positioning on the Internet can constitute an advantage in the pursuit of the company's strategy.

It is evident that business and Web strategies need to undergo an integrated process of definition, development, and implementation. On the other hand, if through the business strategy the company's management can understand what to use the web for, through the Web strategy it should be able to understand how to use it. The colonization of the virtual space offered by the Web should then be realized through a progressive alignment and a systematical evaluation (Memmola, 2007) of:

- The company's main strategic orientations.
- The degree of acceptance and awareness of the Internet's potentials on the part of the company's internal and external stakeholders (i.e., customers, personnel, suppliers, dealers, etc.).
- The possible impact on the organizational structure and internal processing.
- The company's general usage logic of other information and communication technologies sustaining its processing.
- The costs and benefits linked to the service or information contents activated one by one on the Web.
- The performance measurement mechanisms enabling the evaluation of the project's success.

However, the creation of a strategy (business strategy or Web strategy) consistent with the internal and external environment in which the company operates is not sufficient. Porter (2001) observes that it is essential to also "act upon" the strategic positioning, that is, on the ability to do business in a different way than the competitors, in order to create value for the customers, that is, not imitable. Which means using the Internet and ICT solutions in order to apply an univocal characterization of the product system to the image, the quality, and to complementary services rather than to the distribution logistics in such a way that it will be impossible, or at least very difficult, for the competitors to copy.

It is subsequently necessary to define the strategic profile of the Internet presence, in order to be able to continue with the positioning of the single company compared to the direct competitors, and to do the necessary categorizations and possible benchmarking actions.

According to Vittori (2004) and Buttignon (2001), the evaluation of a company's approach to the Web is normally based on the technological dimension (i.e., show-case or static site, interactive site), the recipients of the communication process (i.e., B2B, B2C, suppliers, customers, etc.), or on the goal(s) of the site (i.e., e-commerce, customer service, branding, database creation, etc.).

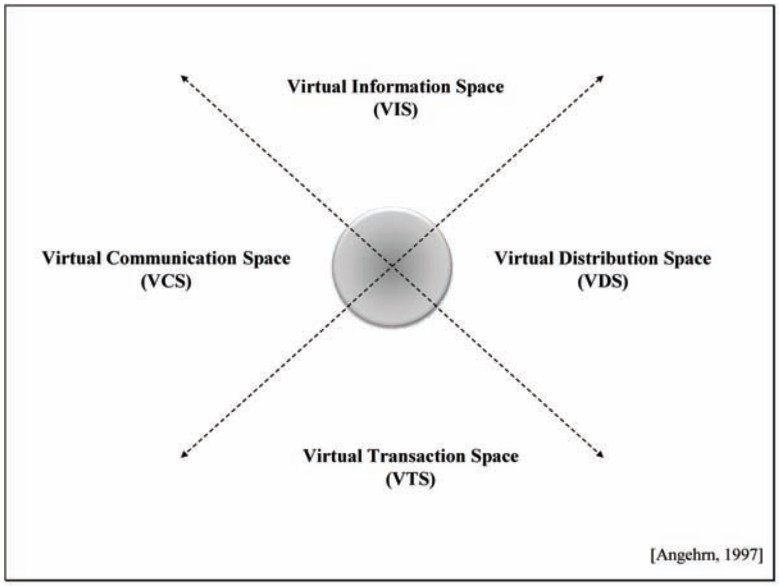
Such orientations are not consistent with the objectives and the set-up that we have put into practice in our research work. In order to proceed with the evaluation of the strategic positioning of the companies of the automotive sector, we have used the model created by Angehrn (1997) that goes under the acronym of ICDT, which stands for *information, communication, distribution, transaction*. This approach explains in a simple, but effective way, the sort of strategic approach used by companies for the definition of their own Web site, evaluating how and to what degree they are using the potentials offered by the Internet.

The model's name (Figure 1) derives from the segmentation of the Internet's virtual space into four main spaces: the v*irtual information space* (VIS), the *virtual communication space* (VCS), the *virtual distribution space* (VDS), and the *virtual transaction space* (VTS). This segmentation highlights the fact that "the Internet has extended the traditional market space by providing new spaces in which economic agents can interact by exchanging information, communicating, distributing different types of products and services and initiating formal business transaction" (Angehrn, 1997). In particular:

- VIS: The space that companies use to introduce themselves to customers or other stakeholders; in essence, it is a "one way," low-cost communication channel that the companies of the automotive sector use to present their organizational structure, mission, commitment to the environment, and linked services (e.g., assistance, maintenance) or complementary products (e.g., secure driving courses etc.).
- VCS: Offers the possibility to create a space on the Web to build relationships and exchange ideas and opinions; in this case, the aim of automotive companies is that of customer retention through the creation of virtual communities such as a forum, a chat, or a blog where customers and fans may exchange information or express their views on products, brand, services, and so forth.
- VDS: The Internet offers a new, efficient, and inexpensive distribution channel, suitable for a variety of services and products. All companies offering goods or services that can be digitized (e.g., e-books, videos, music, etc.), but also companies wanting to offer services to support their traditional products (e.g., online assistance, training, etc.) can benefit. Obviously, the automotive companies belong to the latter category and they can, for instance, offer the customer the possibility to download their car instructions manual, to receive reminder services for car maintenance, and so forth.
- VTS: The space allows one to realize remote transactions with different interlocutors (e.g., customers, suppliers, personnel, etc.); this Internet scope has traditionally been identified as e-commerce, related to the online sale of goods or a series of contents which in any case "assists" the transaction process (e.g., booking of test-drives, maintenance and repair services, job offers, etc.).

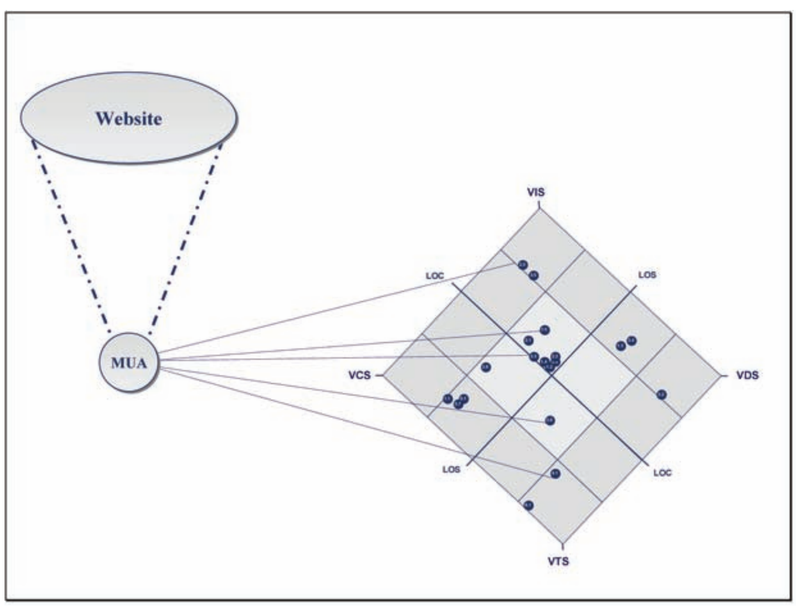
With Angehrn's model, a "map" of the Internet presence can be drawn up, aiming at offering a synthetic and prompt view of the Web strategy pursued by the company (Figure 2).

Figure 1. The ICDT model (Angehrn, 1997). © 1997. Albert A. Angehrn. Used with permission



To this end the Web site is ideally divided into a series of minimum units of analysis (MUAs). A MUA represents an area inside the site characterized by the homogeneousness of the contents (informative or service) representing a precise "occupation" area of the Net's virtual space. As such, it does not necessarily coincide with the single Web page, but can be spread over more pages and can share the same page with other MUAs. In the VIS area the MUAs which can

Figure 2. Creation of the map of the Internet presence



be activated are mainly informative ones, for example, those informing about the company's history, mission, product range and, so forth. In the VDS area instead, a MUA does normally have the task to give a content service, such as a reminder service for the car maintenance. Each single MUA will then be placed within its virtual correspondent area (i.e., VIS, VDS, VCS, VTS). The more external the MUA, the higher the degree of technological sophistication (LOS) and the level of content customization (LOC) measured along the orthogonal axes in the figure.

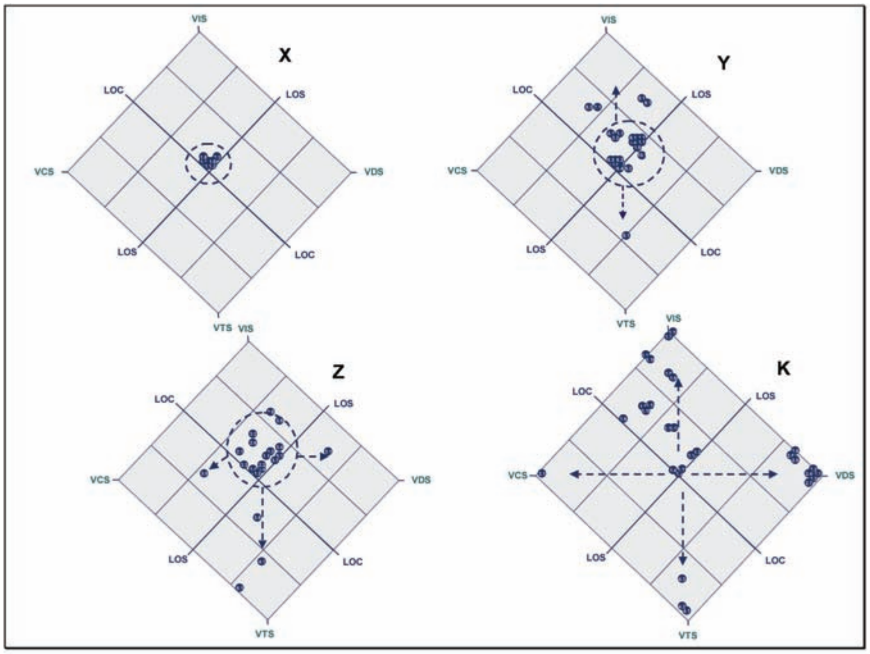
The Internet presence map allows us to evaluate and compare how and to what degree the automotive companies are using the Web to pursue their business strategy. It is therefore possible to work out a taxonomy of the strategic approaches to the Internet that we have divided into four main scenarios (Figure 3):

- The *low profile strategy*: In this scenario (represented by typology X in Figure 3) the company, although present on the Net, is (almost) unable to exploit its opportunities. Probably, the company's top management does not consider the Internet's technology as a means through which the company can pursue its business strategies, or the company has not clearly defined what the Web should be used for, or even, it may simply represent the company's answer to their stakeholders' low degree of acceptance and awareness of the Internet.
- The *medium profile strategy*: The scenario (typology Y in Figure 3) defines an operative context in which the company seems only interested in exploiting a few opportunities that the Web is able to offer. Very likely, in this situation the company's top management is still not fully aware of the opportunities that the Web can offer, or in any case it does not want to make substantial investments in this technology, due to the fact that the stakeholders do not value the Internet very highly.
- The *high profile strategy*: In this case (typology Z in Figure 3) the Web strategy starts to actively support the pursuit of the institutional strategy. The company, although it started exploiting the Web as a mere low-cost communication tool, is now starting to significantly broaden and qualify its presence on the Web, "aggregating" other types of virtual space (e.g., VTS, VDS, VCS).
- The *very high profile strategy*: In a context as the one represented by Company K in Figure 3, the top management considers the Web strategy an essential part of the business strategy and the stakeholders accept and are aware of the opportunities offered by the Internet. The occupation of the virtual space offered by the Internet is homogeneous and characterized by high levels of technological sophistication and of content customization; the company also "lives" through, and on the Web. An essential part of the relationship with the customers (or the other stakeholders) is lived through this means of interaction.

The "colonization" of these virtual spaces can obviously be developed according to different approaches and modes. Subsequently, the evaluation of the strategic positioning on the Internet of the automotive companies has been carried out through the consideration of four analytical viewpoints:

- The **complexity of the Internet presence**, or the intensity with which a company uses the Internet to sustain its institutional activities. An intensive and extensive usage of this technology will very likely be realized through a Web site that has got "plenty of" content able to give information about the company (VIS), furnish goods/services

Figure 3. The taxonomy of the strategical approaches to the Internet



(VDS), create virtual places of communication (VCS) and carry out transactions (VTS).

- The **strategic architecture of the Internet presence**, or the more or less balanced modes through which the company has continued with the colonization of the different virtual spaces made available by the Internet, privileging some stakeholder typologies (e.g., private customers, suppliers, dealers, rental companies, business customers, etc.) and, perhaps, neglecting other typologies.
- The **quality of the Internet presence**, measured by the LOS and by the LOC of the contents proposed by the Web site.
- The **effectiveness of the Internet presence**, in terms of visibility (based on the main search engines) and success (visitors' count) of the company's Web site.

The Complexity of the Internet Presence

The aim of this analytical perspective is to measure the consistence of the Internet presence in the automotive sector, not so much in terms of the Web site's number of pages, but content richness serving the purposes of:

- Providing information about the VIS area.
- Distributing services in the VDS area.
- Creating spaces for virtual communication in the VCS area.
- Realizing transactions in the VTS area.

The higher the number of contents offered in the various areas, the higher the degree of complexity that can be attributed to the site, which is measured through two indicators:

- The amount of MUAs counted in the site.
- The content usability index (CUI).

Table 1. Methodology for analysing the Internet presence

Step 1: Internet presence complexity	• examine Web site contents and classify them in MUAs • assess the existence of other structural features (e.g., foreign language version, native language search engine, Web site map, Extranet)
Step 2: Internet presence strategic architecture	• position retrieved MUAs in corresponding virtual space areas (VIS, VDS, VCS, VTS) • verify Web site use of Internet potential
Step 3: Internet presence quality	• assign the LOS of each MUA • assign the LOC of each MUA
Step 4: Internet presence effectiveness	• verify Web site visibility by means of the most common search engines • measure average Web site traffic

The concept of MUA has been explained above, whereas the CUI allows one to measure, in complex terms, the existence of a series of "structural characteristics" that have been activated within the site to improve content usability. It is infact logical to suppose that when the amount of content increases, for a Web site user looking for information, the degree of complexity increases accordingly; the presence of an internal search engine or a site map can be very helpful to this end. We also wanted to consider the site's usability by a particular category of stakeholders, that is, people with poor eyesight, verifying the possibility of easy modification of character size, background color, or the existence of a voice describing the contents. The CUI is thus the outcome of the ratio between the number of structural characteristics deemed noteworthy according to the methodology of the research, that is:

- Site map.
- Search engine.
- Site accessibility by poor sighted people (character size and background modification, etc.).
- Reserved area/Extranet.

An Extranet area can be defined as a company's nonpublic presence on the Internet. Whereas the use of an Intranet is totally private, within the company and by personnel only, an Extranet may give information and contents outside the company's scope in a reserved way. In fact, customers can use it to share information or services (the so-called customer clubs). Therefore, although the Extranet is not a tool aimed at improving the legibility of the site's contents, it actually increases the degree of complexity of the Internet presence.

The Strategic Architecture of the Internet Presence

Building a Web presence allows a company to use virtual space, which areas (i.e., VIS, VDS, VCS, VTS) can be used in various ways and to various extents, depending on the company's Web strategy.

This analytical perspective allows one to highlight the Web strategy followed by the company in the "colonization" of such space in terms of:

- Number of occupied virtual spaces, in order to evaluate the size of the presence built on the Internet.
- Exploitation degree of the Net's potential in global terms and of each typology of virtual space (VIS, VDS, VTS, VCS).
- Stakeholder typologies that the company is privileging (or, possibly, neglecting).

The indicators used to evaluate these aspects are respectively:

- The degree of "colonization," resulting from the simple enumeration of the Internet's virtual space areas occupied by the company (VIS, VCS, VDS, VTS).

- The *coverage index*, or the degree of coverage of the virtual space considered as a whole and of each single area.
- The *coverage index* in respect to the different stakeholder categories, that is, the degree of coverage of the demand for information and services expressed by the site's users (i.e., customers, dealers, suppliers, etc.), who for whatever reason are interested in the company.

Defining the degree of colonization of the Internet's virtual space is a simple but effective way to measure the evolutionary phase of a company's Web strategy. However, since an area can also be "colonized" through only one MUA, it is clear that this indicator furnishes a very approximate indication that needs to be integrated by the study of the mean coverage and the single area coverage index. If the degree of colonization indicates the company's presence in one of the Internet's virtual areas, through the coverage index it is possible to assess how much the company is benefitting, through its site, from the opportunities offered by the Net in that specific area.

In particular, the coverage index for each area (VIS, VDS, VCS, VTS) is fixed according to the ratio between:

- The number of MUAs found in the company's site for each area.
- The total number of MUAs found in the same area during the research.

The average coverage index is subsequently calculated through the mean average of the values calculated in each single area and supplies a synthetic, though approximate measure, of how much the company is exploiting the advantages offered by the Internet in general terms.

Finally, the coverage index regarding the different stakeholder categories is the outcome of the relation between:

- The number of MUAs found in the site that relate to a particular stakeholder category.
- The total number of MUAs that relate to the same stakeholder category found during the research.

The Quality of the Internet Presence

With regard to the qualitative dimension of the Web contents, the objective of the research work aims at using a measurement process that would be able to take into account the technical aspects as well as the traceability of the contents within the site, and that would respond as much as possible to people's informative or service needs.

Among the many alternatives, the drivers that have been adopted by the research methodology to evaluate the Internet presence are the LOS and the LOC, presented with their mean value through the MUAs found in the sites of the analyzed companies.

The evaluation of the level of sophistication (in the broadest sense of the word) is linked to the technology used. A progressive variable scale from 1 to 15 has been used to represent the various intermediate situations that can occur through:

- A MUA based on static pages realized with programs that were not specifically ideated for the Internet (e.g., MS Word, Powerpoint, etc.).
- A MUA based on real pages with the most modern tools of multimedial presentation (e.g., Flash View animations, audios, videos, images) and with a high degree of interactivity.

The level of customization is also based on the possibility to "cut" the offered site contents, according to the user's specific customization or service needs. In this case too, a progressive evaluation scale from 1 to 5 has been used. These values represent the extremes of a continuum of intermediate solutions that can turn up between:

- A MUA based on static pages, without any possibility of customizing the content, and without any additional services.
- A MUA based on pages that allow the creation of dynamic, customized contents with accessory services, for example, to build your own customized home page.

The Effectiveness of the Internet Presence

We wanted to complete the evaluation process of the Internet presence of the companies of the automotive sector through an analysis that would express the degree of "success" obtained by the Web sites. This intention implies, nevertheless, the solution of two methodological matters:

- The first concerns the need to translate in actual terms the very concept of "a Web site's success."
- The second is closely linked to the first and concerns the finding of variables able to give a measurement as objectively as possible.

There are basically two sorts of goals a company can have when setting up a Web site:

- Maximizing its visibility on the Web and therefore be reached by the highest number of potential users.
- Maximizing the use of the Web site by all the stakeholders.

The higher the degree of the above goal accomplishment, the more effective (and successful) the Internet presence will be, thereby also solving the second critical point, that is, the effectiveness level measurement of a company's Internet presence. The measurement has been obtained through the combined consideration of:

- A traceability index resulting from the major Italian and international search engines.
- A ranking of the Web site's registered traffic supplied by "Alexa.com," which draws up an international "hit parade" based on the registered traffic.

The first indicator tells how easy it is to "reach" the Web site through the major search engines (e.g., Yahoo.com, Google.com, Msn.com, Altavista.com, Virgilio.com, Arianna.com). As keywords, we have used, with an incremental approach, the company's name, the name of a product of the company, and lastly, a series of other information (i.e., the dealers and the promotions related to the company, a product's accessories line). The index is equal to the sum of the mean of the positive results (search engines with positive outcome compared to the total number of search engines) obtained for each level of measurement (i.e., company, product, and other information). A result is positive when the search engine has included the searched information in the first seven proposals (normally visualized on the first page of the search results).

Alexa.com is a site that makes a classification of all Web sites worldwide, giving them points according to the number of registered visitors. Alexa.com uses an evaluation model that takes into account both the number of accesses to the site, and the number of pages visited within the site. The number of accesses quantifies the number of visitors that have accessed the site on a particular day. The number of pages visited is calculated by the amount of internal URLs requested by the same visitor. Obviously the repeated request of the same URL, on the same day, is counted as a single page. The site with the highest combination of accesses and visited pages will be given a value equal to 1.

The evaluations furnished by Alexa.com relate to all the active sites on the Web. Therefore, a data normalization process had to be carried out in order to obtain a classification specifically related to the analyzed companies.

MAIN FINDINGS

The Research Reference

This project was developed to focus on the Italian customer point of view. We took into account all the car-maker operating in Italy at the time of the research[3] from the databases of the National Association of Italian Automobile Manufacturers (ANFIA) and National Union Foreign Car Dealers (UNRAE). The study took into consideration the Italian version of the Web site, or when this was not available, original language Web sites. This choice was made following two criteria:

- Automotive companies, except for a few cases, are global in nature. They develop a different approach in terms of products and services offered, reflecting the specific characteristics and the importance of the market, depending on each geographical area (state or continent). We decided, therefore, to evaluate their strategic positioning in a specific market area: Italy, in our case.
- The international sites of larger sized companies (singled out by domain name_company.com) are simply "entry systems" through which they are readdressed to the site of the country or language (singled out, for example, by domain name_company.it or name_company.uk).

Obviously, this decision may have an impact on the results of the research, which are in fact influenced by the strategic choice which the company decides to follow in that specific market. Nevertheless, research does not imply any change or adjustment to the framework which remains the same, regardless of the geographic area considered.

The research sample examines 56 automotive companies that roughly fall into two macro categories: generalist companies and niche companies. The first category includes those automotive companies with differentiated offer (in terms of price level and type/model of car and service features), multiple targets, and different strategic areas of market covered; while the second category includes automotive companies that have a strong brand and image, often part of a group, offer a limited range of luxury cars and focus on a segment, target type, or special use. These tend to present a higher level price range with many customized options (even if the variable is not discriminatory) (see Table 2).

For each dimension of analysis (e.g., complexity, strategic architecture, quality, effectiveness) a positive and negative benchmark was created, defining the average value of 5% of companies with the best and worst performances, respectively.

Analysis of the Internet Presence Complexity

The complexity of the Internet presence was evaluated through the combined consideration of the following indicators:

- Number of MUA activated within the Web site.
- CUI, calculated on presence or absence, within the site, of fixed structural characteristics such as: a) site map, b) internal search engine, c) accessibility to site for poor sighted people, and d) reserved area (Extranet).

The research identifies a total of 56 MUAs variously distributed in the different areas where it is possible to articulate the virtual space of Internet (VIS, VDS, VCS, VTS). Three distinct clusters (see Table 3) stand out from the Web sites of theAnalyzed companies:

Table 2. Macro categories of reference of the analyzed companies sample

Generalist Manufacurers	Niche Manufacturers
• Alfa Romeo • Audi • BMW • Cadillac • Chevrolet • Chrysler • Citroen • Dacia • Daihatsu • Dodge • Fiat • Ford • Honda • Hyundai • Jaguar • Jeep • Kia • Lancia • Land Rover • Lexus • Mazda • Mercedes • Mitsubishi • Nissan • Opel • Peugeot • Renault • Saab • Seat • Skoda • Ssangyong • Subaru • Suzuki • Tata • Toyota • Volkswagen • Volvo	• Abarth • Aston Martin • Bentley • Corvette • Daimler • Dr • Ferrari • Hummer • Lada • Lamborghini • Lotus • Maserati • Mayback • Mini • Pagani • Porsche • Rolls-Royce • Santana • Smart

- **Cluster A**: Companies, using the Web, with a high number of informative or service contents, with over 30 activated MUAs.
- **Cluster B**: Companies with an average profile approach and between 25 and 30 activated MUAs
- **Cluster C**: Companies with a low profile approach and less than 25 MUAs.

The simple enumeration of MUAs must not be misleading. We are not looking at performance criteria and stating that a company that activates a higher number of MUAs inside its site is better than one that activates just a few. A company's Internet strategy is nothing more than a "translation" of its business strategy on the Web. This explains how the activation of some MUAs, for example, those supplying informative content or provide a booking service for their company museum, is typical of companies that focus on tradition and historical authenticity of their brands as an added value element (e.g., Alfa Romeo, Ferrari, Lamborghini, Maserati, etc.)

From the above considerations, it results that the number of activated MUAs is not a satisfactory indicator of the complexity of the Internet

Table 3. Cluster of 56 analyzed companies

Companies with less than 25 MUA	Companies with between 25 and 30 MUA	Companies with over 30 MUA
Daimler	Honda	Citroen
Lada	Suzuki	Ford
Dr	Bmw	Jaguar
Abarth	Hyundai	Mazda
Pagani	Kia	Saab
Rolls-Royce	Seat	Volkswagen
Mayback	Chevrolet	Volvo
Santana	Ferrari	Lotus
Jeep	Smart	Porsche
Chrysler	Subaru	Fiat
Dodge	Audi	Skoda
Lamborghini	Cadillac	Lancia
Tata	Nissan	Alfa Romeo
Dacia	Land Rover	Mercedes
Hummer	Mini	Renault
Daihatsu	Mitsubishi	Maserati
Toyota	Opel	
Bentley	Peugeot	
Ssangyong		
Lexus		
Aston Martin		
Corvette		

presence, which gives way to the introduction of the second indicator, the CUI as previously defined.

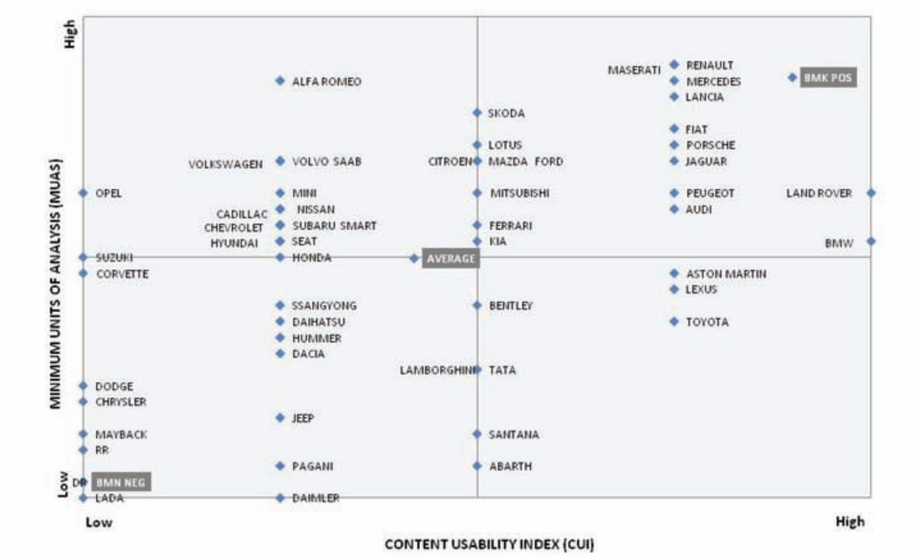
Overlapping the two parameters used to evaluate the complexity of the Internet presence, certain inconsistencies come to light: one would expect a high number of CUIs to result from a high number of MUAs, due to greater usability of the numerous contents offered. In fact, in several situations this hypothesis is denied by sites which, faced with a rather large number of activated MUAs, present a quite low CUI index (Alfa Romeo, Volkswagen, Volvo, Saab).

The good positioning of Maserati, Renault, Mercedes, and Lancia stands out. Alfa Romeo's situation is worth mentioning since it presents the greatest number of activated MUAs with a poor usability level due to the absence of basic navigational instruments (i.e., site map, search engine, accessibility for poor sighted people) on which the CUI index was developed. It is usually the general manufacturers who offer the greatest number of contents, compared with the niche manufacturers. Lotus is an exception, which even if it belongs to the second category, shows up in the higher part of the second quadrant.

Analysis of the Strategic Architecture and Internet Presence

This perspective of analysis recognizes the strategic profile of the Internet presence of companies

Figure 4. The complexity of the Internet presence



more in the automotive sector than in other areas of research. A company pursues a high profile Web strategy the moment it achieves a presence that is:

- "Widespread" in each of the areas in which the use of the Internet virtual space has been developed (degree of colonization).
- Quantitatively important in terms of number of MUAs activated compared to those potentially activeable (coverage index).

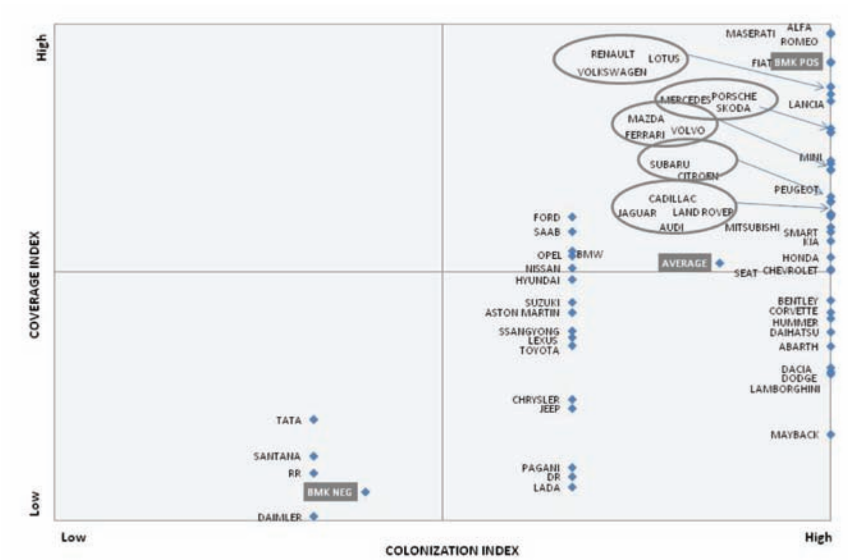
The results of the research demonstrate a generally elevated degree of colonization (average value 3.57) as a consequence of the fact that the automotive companies have predisposed Web sites with at least one MUA in each virtual space. Nevertheless, the potential of the Internet is still only used to a limited degree, that is, less than half of its potential. The average value of the coverage index is just 37%. The fact that many companies still only have a weak presence in the areas of greater value content such as VDS, VCS, and VTS (the coverage indices specific to these areas are generally very low) contributes to this result.

Nearly all of the analyzed companies, however, adopt an Internet strategy which concentrates on the information space (VIS), in which the contents are expressed in information regarding the company structure, the products and services offered by the company, the prices of the products or services, contacts, and so forth.

There are five distinct clusters of indexing (Figure 5):

- **Cluster A:** Companies with a Web strategy of medium/high profile that develop a widespread presence in all the areas and with an over average coverage index. Companies nearest to the positive benchmark like Alfa Romeo, Maserati, Fiat, Volkswagen, Renault, Lotus, but also Cadillac, Jaguar, Smart, Land Rover, and Audi, who show lower degrees of exploitation of the Internet opportunities, appear in this cluster.
- **Cluster B:** Companies with a medium profile Web strategy which develop a widespread presence in all the areas and with a coverage index value in line with or lower

Figure 5. The strategic architecture of Internet presence



than average, for example, Honda, Seat, Chevrolet, Corvette, Hummer, Abarth, and so forth.

- **Cluster C:** Companies with a Web strategy of average profile which develop a presence limited to three areas, but with an average coverage index value (e.g., Ford, Saab, Opel, Nissan, Aston Martin, and Toyota).
- **Cluster D:** Companies with an average/low Web strategy profile which develop a presence limited to three areas, but with a consistently lower than average coverage index value (e.g., Chrysler, Jeep, Pagani, DR, and Lada).
- **Cluster E:** Companies with a low profile Web strategy which develop a presence limited to only two areas and with a lower than average coverage index value (e.g., Tata, Santana, Rolls Royce, and Daimler Jaguar).

Another evaluation parameter of strategic architecture of the Internet presence concerns the evaluation of informative or service contents activated for each individual stakeholder category.

Figure 6 shows the performance of the average coverage index of the automotive sector with regard to the Italian market compared with the informative or service contents activated for each stakeholder; it is possible to develop the following considerations:

- It is confirmed the trend highlighted in the previous lines underlining that the companies in this sector are not yet able to fully exploit the potential of the Internet. On average, the index of coverage for each category of stakeholder varies its range from a low of 41% to a maximum of 54%.
- Certain categories of stakeholders (i.e., suppliers, business customers, rental companies, public companies, new employees) are more privileged than others (e.g., partners, employees, protected categories).
- The relatively low coverage for private customers shows how the potential of the Internet, compared with those who may be classified as the institutional recipients

Table 4. Stakeholder considered in the research

Stakeholder	Description
Providers	Companies providing semifinished goods (e.g.m alloy wheels, plastic wheel cover, piston, connecting-rod, etc) or providing services (financial companies, consulting, etc.) for the analyzed company.
Partner	Companies that, participating in the design or development of the product, may be considered a strategic partner of the analyzed company (e.g., Microsoft developing in collaboration with Fiat the Blue & Me service or Brembo, a world leader in braking, involved with a decisive role in the design of a new model).
New Employee	Those who are interested in starting a working relationship with the company analyzed (e.g., work with us...).
Employee	Those who have an employment relationship with the company analyzed.
Dealers	Companies that are holders of a concession of sales of the company analyzed (includes the car dealers and the so called “authorized” and “retailers”).
Car Rental Companies	The car rental societies.
Private Customers	Actual customers or prospects of the analyzed company.
Business Customers	This category is different from the previous one because it does not refers to private customers in a B2C logic, but to companies in a B2B logic. Expectations at a content or service level are obviously different from the previous stakeholder identified.
Persons With Disability	Those who are actual or prospect customers of the analyzed company, carrying a form of disability.
Public Administration	The public corporations (e.g., state, region, province, municipality) which may be interested in the territorial development policy implemented by the analyzed company.

of the Web site remains high and largely unexplored; this result is the consequence of a large number of MUAs related to this type of stakeholders (see Figure 7) a still relatively modest coverage especially in more “advanced” areas such as the VDS, the VCS, and the VTS.

Analysis of the Quality of the Internet Presence

The quality of the Internet presence is intended as the level of technological sophistication of the distributed contents and the possibility to enter in a personalized or assisted way. The following parameters of evaluation have been used to develop such a perspective analysis:

Figure 6. The stakeholder coverage index

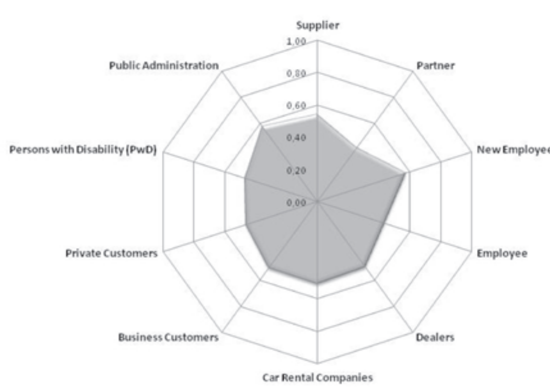
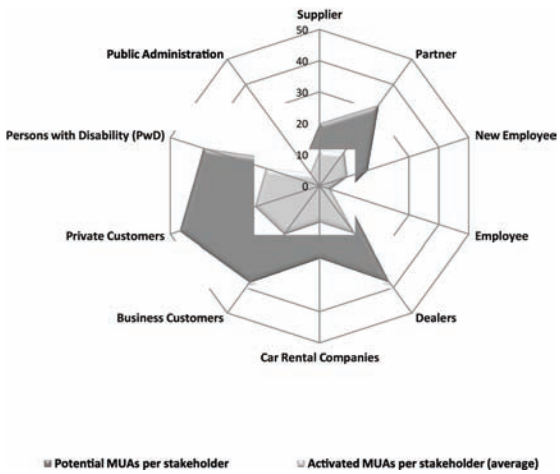


Figure 7. MUAs per stakeholder



- Level of technological sophistication (LOS)
- Level of customization (LOC)
- Measured on a scale from 1 to 5.

As a first approximation, Figure 7 shows the indexing of automotive companies regarding the qualitative dimension of their presence on the Internet. Although in this case the significant differences noted in other analysis perspectives were not evident. Three distinct clusters however arise:

- **Cluster A:** The first group basically shows the presence of companies like Mini, Ferrari, Maserati, and Alfa Romeo and assumes a "pioneering" position regarding the qualitative characterization of their presence on the Net. The more sophisticated the technology, the higher the possibility to allow customization of the Web site content; this is far superior compared to the average value of the two parameters.
- **Cluster B:** The second group in the position of "chasers" is particularly noteworthy considering it combines the majority of the companies present on the market with performances that are not far or around the average measured value.
- **Cluster C:** A final group of "latecomers" who present a considerable weakness in respect to this dimension of analysis, consisting of those companies that have values close to the negative benchmark, such as Aston Martin, Santana, Tata, Lada, and so forth.

A further evaluation takes into consideration the balance between LOS and speed of access to the contents in each individual Web site. Most of the sites (41%) present a "strong preference" for LOC rather than LOS (with higher values of the first index than the second one). So the companies tend to favor highly customized contents over technology, which ends up being a little sacrificed. Only 11 of the 56 analyzed companies, in fact, consider the level of sophistication more important than the level of content customization.

A final consideration concerns the alignment between business strategy and Internet strategy. Companies with greater evocative power, with a middle to high position in terms of market level or which target the young or a niche, tend to present sites that focus on aspects of emotional involvement, full of images or interactive contents. The aim, in this case, is to create a bond with the brand rather than the mere gratification of informative demands.

In this sense, it can be noted: how Mini stands out thanks to a technologically advanced site and to highly customized contents; how Fiat has a good position, thanks, above all, to new ideas introduced for the launching of the new Fiat 500; how Ferrari has one of the richest sites regarding service contents and high path diversification; and how Abarth presents an extremely distinguished image from a technological point of view, offering the possibility to customize the site structure, its aspect, colors, and music, almost achieving the type of tuning which is at the basis of this brand's philosophy.

Analysis of the Effectiveness of the Internet Presence

The measure of success of the strategic choices by companies in the automotive sector regarding Web positioning was achieved through the combined consideration of the following parameters:

- A tracing index of the site through a panel of search engines.
- A ranking, created by Alexa.com, which supplies an indication of the traffic registered by the Web site.

Generally speaking, the Web sites of automotive companies enjoy good visibility: 63% of them presents a high traceability index (more than 2), 23% presents average visibility (with an index between 1.60 and 2), and only 8 out of the 56 companies presented a decidedly low visibility.

Concerning registered average traffic, the results of the research offer two considerations:

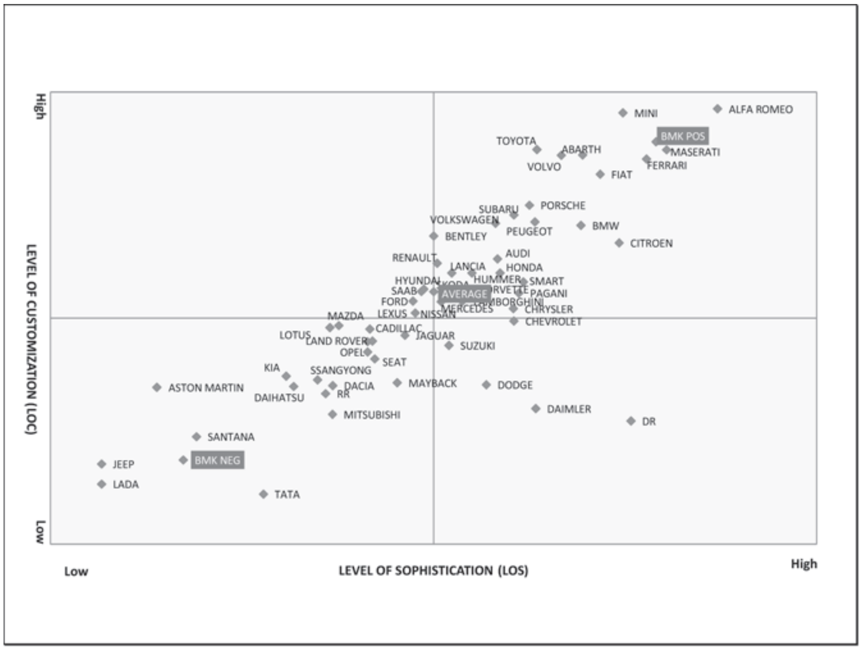
- A substantial aggregation of performance values corresponding to the positive benchmark; with the exception of a few companies (e.g., Fiat, Nissan, Volvo) great differences in terms of traffic registered do not exist in the surveyed Web sites.
- The consistent distance which exists between the average reference values and the companies which present a performance close to the negative benchmark (Figure 7). The companies in question (i.e., DR, Hummer, Ssangyong) are relatively new on the Italian market, or have an extremely focused offer. In our view, these could be the main reasons for such different results.

Focusing the attention on the companies present in the quadrant II, it is possible to note the existence of four clusters:

- **Cluster A:** Companies that have an effective presence on the Web with a high level of visibility and traffic registered. Companies like Fiat, Nissan, Volvo, Alfa Romeo, Ford, Citroen, Mercedes, and so forth, which are closest to the positive benchmark, are included in this first area.
- **Cluster B:** Companies that record high levels of traffic, but show a lower visibility compared with the search engines we used (e.g., BMW, Mazda, Audi, Suzuki, Subaru, Mini, etc.).
- **Cluster C:** Companies which present high visibility levels, but at the same time achieve a relatively worse performance in terms of recorded traffic (e.g., Volkswagen or Jaguar) who therefore present atypical behavior.
- **Cluster D:** "Follower" companies with visibility levels and recorded traffic lower than the average (e.g., Jeep, Tata, and Chrysler).

The previous arguments refer to the dynamics of competitive interaction among companies in the automotive sector specific to the Italian market. As in every other research area, also with regard to this perspective of analysis, the reference was

Figure 8. The quality of the Internet presence



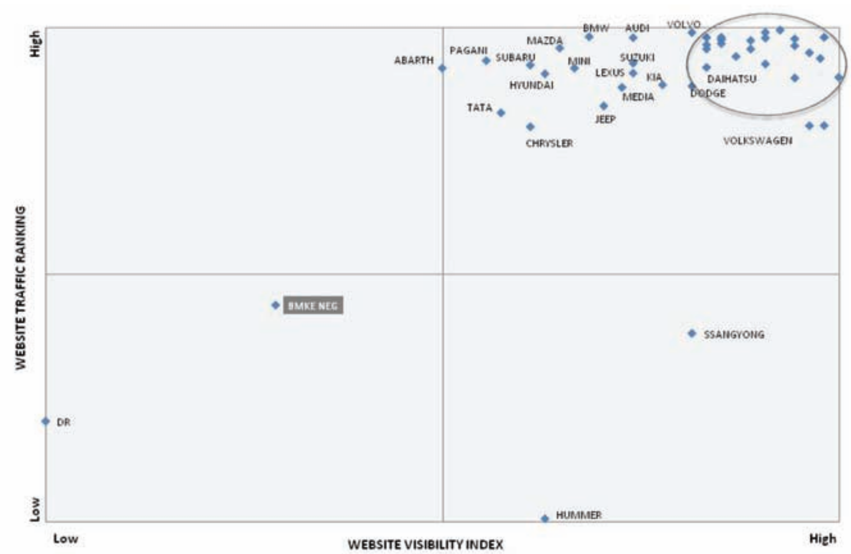
the Italian language Web sites of Italian automotive companies, whose contents, in certain cases, are not managed directly, but by a branch or the Italian importer. It is evident that the results obtained in terms of visibility, but above all recorded traffic inside the Web site, are influenced by the competitive positioning that the company achieves in the real world. In other words, the particularly high performance rate of Fiat, Alfa Romeo, and Lancia is also a consequence of the fact that the study deals with well-known national brands with high values of brand awareness. On the contrary, the results of Chrysler or Jeep can be fit into the logic of a minor diffusion of these brands, having almost a niche presence on the Italian market.

We have therefore tried to repeat the evaluation of this perspective of analysis using original language sites of the different companies (singled out by the extension .com or from the extension of the country of origin as, for example, Lotus [www.lotuscars.co.uk]). The results shown in Figure 8 are, evidently, more consistent compared to the impact that different automotive companies, or rather their brand names, have on the international scene; Toyota, Mercedes, Chevrolet, Ford, and Volkswagen are the players who achieve the highest performance levels. Italian brands undergo a decisive reduction, with the exception of Ferrari which, although it is a niche producer, has an international standing. Lancia's case is noteworthy; the brand name is present on the Italian market, but not particularly diffused on the international scene, which in this new evaluation presents performance levels that are a lot lower than the average.

LIMITATIONS AND FURTHER RESEARCH AREAS

In order to present a complete view of the phenomenon, the methodology used to analyze the presence of automotive companies on the Web inevitably experiences a few structural limits which may be clarified as follows:

Figure 9. The effectiveness of the Internet presence



- The Internet is an ongoing reality. The research project photographs the situation as it was at the time of the analysis of the Web sites, which took place in the months of March and April 2007.
- The aim of this research is to provide an objective evaluation of the presence on the Internet. Even though a precise scale of evaluation of the LOS and of the possibility to proceed to a customization of the contents offered by the Web site (LOC) had been defined, there are inevitable elements of subjectivity present in this phase which cannot be eliminated.
- The research examines the Italian user's point of view. We have already discussed the reasons for this choice which, however, do not prejudice the possibility of applying this methodology to any other geographical context. It was in our interest, furthermore, to evaluate the dynamics of the competitive positioning of the companies in the automotive sector in a specific geographical area.

The research presented in this chapter is the first phase of a complex project, still being carried out, which also includes the development of an in-depth analysis through a series of case histories which help to fully explain the possible definition paths of the Web strategy of an automotive company.

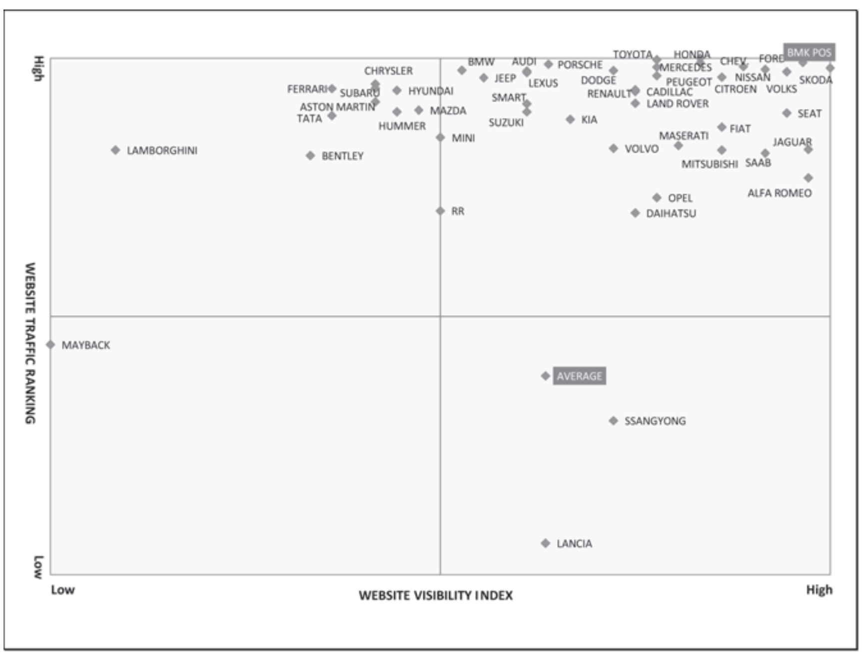
The intent is to move from an analysis which takes place "outside" of the company, which is resolved in the evaluation of the "finished product" of the Web strategy, through the site to an analysis "within" the company. The aim is to have a complete vision of the phenomenon object of the survey, studying the path of preparation, definition, and successive implementation of the Internet strategy, the critical aspects and the duration of development, the impact on the organization, and the expectations and requirements of the persons either directly or indirectly involved.

CONCLUSION

This research would not have had much sense 15 years ago. More to the point, it would have been

Figure 10. The effectiveness of the Internet presence (focus on 2nd quadrant)

Figure 11. The effectiveness of the Internet presence (international positioning)



impossible to carry out. The raw materials on which our study is based would have been missing, or, anyway, the object phenomenon of the survey would have had totally different levels and dimensions.

It is in this simple and obvious reflection, but absolutely corresponding to the truth, that the extraordinariness of Internet must be found. It is almost banal to say that Internet has overturned the logics and paradigms of our daily life. This same book has been written through the Internet! Its authors have interacted virtually over the Net, without ever having met physically.

Inevitably, Internet also changes the company logics and paradigms. Thanks to (or because of) this technology, the way to run a company has changed. The Web "opens" new opportunities. And again, not even these are new concepts. Very concisely it is possible to affirm that through a redesigning of the company proceedings, the Net paves the way for improved performance in terms of lower costs, higher quality, and greater capacity of innovation. The Internet is a remarkable instrument for interaction with the company stakeholders, while other media provide only one-way channels of communication. Through the Net it is possible to improve logistics and interaction. Parallel considerations can be made for other phases of the company's value chain, such as the distribution and the relationships with the car dealers in the case of automotive companies. Finally, the Web could become the preferred instrument for interaction with customers (not just the private ones, but also business-to-business), in order to fully understand their expectations and requirements, gain their loyalty, and even involve them in the design and launching of new products (as Fiat did with the site 500wantsyou.com concerning the new Fiat 500).

The feeling which emerges from the results of this research is that companies in the automotive sector are not fully exploiting the opportunities.

Frequently, a limited use of the Web is witnessed, focused on the less complex areas, such as the allocation of contents usually of the informative sort and low technological level. Although it cannot be said that automotive company sites are merely digital transpositions of the brochures or product catalogues, wider areas of improvement could be made, both in terms of services offered and greater personalization of the proposed navigation routes. In general terms, it can be stated that the focus is still basically on the product and on the image associated with the product.

A further consideration examines the approach to the Net, which does not always seem to occur strategically, but often with a logic tied to the improvisation or an image of the Internet as a mere communication tool. In several cases, the impression is that the Web site "frees" itself from the logics of strategic positioning that the company follows in the real world in order to follow merely commercial or promotional logics, on one hand, or to present the product and the company sometimes too quickly and superficially. Companies rarely demonstrate awareness that the investment in this technology must come through an appropriate alignment with the institutional strategy of the company (Minard, 2001), with a correct review of the main operative mechanisms and company governance.

If we were to retrace the results of our analysis to different steps of the purchasing process, it can be stated that the initial stages of seeking basic information and presentation of product alternatives are definitely well covered, with often complete contents and of a multimedia sort. The phase of the actual transaction, and above all, the one linked to the management practices of post sales services (e.g., assistance, maintenance, etc.) tends to be poorly managed, with a few exceptions. The "pioneeristic" practices allow the online management of accessory services such as financing, purchase or ordering, and possible payment for maintenance/repairs, or the activation of a reminder service which advises the client when to proceed with maintenance.

The contents allowing companies to gain customer loyalty both through playful components (e.g., screen savers, images, logos, wallpaper but also real videogames) and through mechanisms of online community (e.g., forums, chats, blogs and also personalized credit cards) are beginning to catch on. At times contents are increased without adequate attention to their usability, causing confusion and a frequent absence of a profiling logic of the user, of the user's interests and needs.

The road towards a mature use of the Web by the automobile companies is still a long one, at least judging from the results we have gleaned from our research. The question, or rather, the questions that must be asked at this point are: What is it worth? Is it worth investing in the Web? How can the Web, through either its more traditional ways or from its more evolved declinations (like the various blogs, Wikipedia, YouTube, and Facebook) help to accomplish the corporate strategy of a company?

These are important questions, we realize, that could generate a specific trend of research. As far as we are concerned, having studied the Internet phenomenon for several years, we have become more aware of the fact that it is not a question of how much to invest, but rather how to invest. The success and choice effectiveness of the strategic positioning on the Web are influenced by the company's brand, but they also depend heavily on the grade of originality of the informative contents or proposed services.

We would like to conclude with a sentence from one of our many "specialists" we met during the planning phases of our research work: "The point is that on the Internet we are all guests... and as guests, we have to bring something!" What is required, therefore, is to gain the attention, "bringing something which is not normally present in the home of our host, or rather, on the Net," that is, contents that are new and original. But, in order to avoid that, this becomes an exercise for its own sake, it is necessary "to understand what type of invitation we have received and how we intend to behave accordingly." We must know what the expectations of our stakeholders are and how we can achieve them on the Web. By doing this, the corporate strategy of our company will be realized.

REFERENCES

AA. VV. (2002). *Stanford-Makovsky Web credibility study 2002: Investigating what makes Web sites credible today.* Retrieved June 23, 2008, from www.webcredibility.org

Alemi, F. (2000, Fall). Management matters: Technology succeeds when management innovates. *Frontiers of Health Services Management*, *17*(1), 17–30.

Angehrn, A. (1997, August). Designing mature Internet business Strategies: The ICDT model. *European Management Journal*, *15*(4), 361–369. doi:10.1016/S0263-2373(97)00016-9

Angehrn, A., & Meyer, J. F. (1997, Summer). Developing mature internet strategies (cover story). *Information Systems Management*, *14*(3), 37–43. doi:10.1080/10580539708907058

Baraldi, S., & Memmola, M. (2006). How healthcare organisations actually use the internet's virtual space: A field study. *International Journal of Healthcare Technology and Management*, *7*(3/4), 187–207.

Buttignon, F. (2001). *Strategia e valore nella net-economy.* Milan, Italy: Il Sole 24 Ore.

Duke, H. C., & Chul, M. K., Sang-II, K., & Soung, H. K. (2006, December). Customer loyalty and disloyalty in Internet retail stores: Its antecedents and its effect on customer price sensitivity. *International Journal of Management*, *23*(4), 925–942.

Elango, B. (2000). Do you have an Internet strategy? *Information Strategy: The Executive's Journal*, *16*(3), 32–38.

Glaser, J. P. (2002). *The strategic application of information technology in health care* organizations (2nd ed.). San Francisco: Jossey-Bass.

Grant, R. M. (2005). *Contemporary strategy analysis 5/e*. Oxford, UK: Blackwell.

Kaplan, R. S., & Norton, D. P. (2004). *Strategy maps*. Boston: Harvard Business School Press.

Katz, M. L., & Shapiro, C. (1986). Technology adoption in the presence of network externalities. *The Journal of Political Economy*, *94*, 822–841. doi:10.1086/261409

Malcolm, C. (2001, February). Making a healthcare Web site a sound investment. *Healthcare Financial Management*, *2*(74), 74–79.

Malec, B., & Friday, A. (2001). *The Internet and the physician productivity: UK & USA perspectives*. Paper presented at the EHMA Congress, Granada.

Memmola, M. (2007). The development of a Web strategy in a healthcare organization: A case history. In L. Al-Hakim (Ed.), *Web mobile-based applications for healthcare management* (pp. 1-35). Hershey PA: Information Science Publishing.

Minard, B. (2002). CIO longevity and IT project leadership. *IT Health Care Strategist*, *4*(1), 3–7.

Porter, M. E. (2001, March). Strategy and the Internet. *Harvard Business Review*, *79*(3), 62–79.

Robert, M., & Racine, B. (2001). *E-strategy*. Milan, Italia: McGraw-Hill.

Shapiro, C., & Varian, H. R. (1999). *Information rules*. Boston: Harvard Business School Press.

Vittori, R. (2004). *Web strategy*. Milano, Italia: Franco Angeli editore.

Yaobin, L., Zhaohua, D., & Bin, W. (2007, June). Tourism and travel electronic commerce in China. *Electronic Markets*, *17*(2), 101–112. doi:10.1080/10196780701295974

ENDNOTES

1. This chapter is the result of joint research work done by all authors. Nevertheless, Massimo Memmola is the author of section I, II and III; all other authors have equally contributed to section IV, V and VI, the discussion and the conclusion.
2. The research was drawn up in February, 2007; The analysis of the Web site took place in March and April, 2007.
3. A net that can be accessed only by users of a particular company/organization.

This work was previously published in Business Web Strategy: Design, Alignment, and Application, edited by L. Al-Hakim; M. Memmola, pp. 170-197, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 4.6
Using Semantic Web Services in E-Banking Solutions

Laurent Cicurel
iSOCO, Spain

José Luis Bas Uribe
Bankinter, Spain

Sergio Bellido Gonzalez
Bankinter, Spain

Jesús Contreras
iSOCO, Spain

José-Manuel López-Cobo
iSOCO, Spain

Silvestre Losada
iSOCO, Spain

ABSTRACT

Offering public access to efficient transactional stock market functionalities is of interest to all banks and bank users. Traditional service oriented architecture (SOA) technology succeeds at providing reasonable, good Web-based brokerage solutions, but may lack extensibility possibilities. By introducing Semantic Web Services (SWS) as a way to integrate third party services from distributed service providers, we propose in this chapter an innovative way to offer online real-time solutions that are easy-to-use for customers. The combined use of ontologies and SWS allows different users to define their own portfolio management strategies regardless of the information provider. In deed the semantic layer is a powerful way to integrate the information of many providers in an easy way. With due regard for more development of security technological issues, research on SWS has shown that the deployment of the technology in commercial solutions is within sight.

DOI: 10.4018/978-1-60566-066-0.ch016

INTRODUCTION

When operating on the stock market, investors make their decisions on the basis of huge amount of information about the stock evolution, economic and politic news, third parties recommendation and other kind of sources. Thanks to the proliferation of the Internet banks the profile of an average investor is changing from a financial expert to common people making small investments on the online stock market. In addition to the business generated around the stock market operations, banks use their online stock market application to attract new and to reinforce the customer commitment.

Banks, as any other commercial organization, needs to optimize the deployment of new products and services to the market. The deployment time of new services or applications is an important issue in a highly competitive market, since it defines the future market share and revenues. Online banks are looking for technologies and architectural paradigms that would allow them to design, implement and deploy new services on a low cost basis and in a short time period. New services often imply integration of many already existing applications, some of them internal and others external to the organization.

This is the case of online stock brokerage solutions adopted by online banks. An online stock brokerage application proposes to the user to buy and sell its stock options via a computerized network. Banks are willing to offer an easy to use application including as much information and as many options as possible without incurring large development costs. We will show that the use of the Semantic Web technology, combined with a service-oriented architecture (SOA), greatly reduces the cost and effort of developing and maintaining an online stock brokerage solution.

A broker based on a semantic service oriented architecture has all the advantages of a service oriented architecture (e.g. modularity, reusability) combined with the advantages of Semantic Web technologies. Semantic Web technology main advantage is to give a clear semantic inside (and eventually outside) the enterprise which reduces the communication confusions (technical or human). This also leads to higher maintainability of the products and to a better automatisation of the system mechanisms. These advantages applied to SOA will be extended in the proposed solution of this chapter. Next section will first exposes the current situation of brokerage applications based on classical SOA.

CURRENT SITUATION: BROKERAGE APPLICATION BASED ON WEB SERVICES

Banking companies have invested heavily in the last few years to develop brokerage solutions based on a new dominant paradigm in the IT World: service oriented architecture (SOA). The concept of this paradigm is not new: propose a loosely coupled distributed system architecture where independent services provide functionality, so that the difficulty is divided which leads to reduce the development cost and improve the reusability. But the technologies to implement this paradigm are relative new. Web Services are one of the solutions that appeared a couple of years ago and that made the success of this paradigm. For this reason Web Services are often confused with the SOA paradigm.

In this section we first present in more detail the business case for the brokerage application that we propose. We will then explain why a service oriented architecture implemented using Web Services technologies is a suitable solution. The solution properties will be detailed and it will be shown that this kind of architecture is suitable for brokerage application. We then present what the benefits of such an architecture are from both, a technical and a business point of view.

Web-Based Brokerage Applications

Introduction

As a major interface between the financial world and the non-financial world, banks always try to improve their services related to the stock market. As the Internet represents one of the most interesting communication channels of recent years, banks are interested in using this channel to improve the quality of their service and thus increase their image and revenue. Such banks or bank departments have been called eBanks or online banks.We have identified three different strategies for online Banking:

- **Technological leader profile:** Banks that focus their strategy on technology and consider the Internet an opportunity to improve their markets. Also, the Internet specialized banks, usually recently founded banks (not subsidiaries) that have earn a significant market share, even though they do not offer their clients a wide range of products.
- **Follower banks profile:** Banks that first considered the Internet as a threat. When the market has matured, they changed their strategy from a defensive position to a competitive attitude towards those who were the first leaders in Internet banking. In some cases, subsidiary entities were created so as not to cannibalize their own market share.
- **Non "internetized" banks:** Banks that did not invest in the Internet because of their small size, their strategy or other reasons. However, they are a minority in terms of market share.

In these days, banks that already have Web-based brokerage application in place choose the technological leader strategy, while those banks that are only now considering developing their own applications follow the second strategy. Other mediums are also possible:

- **Branches:** Too expensive for Banks. Only for high end clients.
- **Phone banking service:** Expensive for banks. Only for selected clients
- **Mobile services (SMS, mobile phone applications):** Cheap for banks, usually free or at a small fee per service usage.

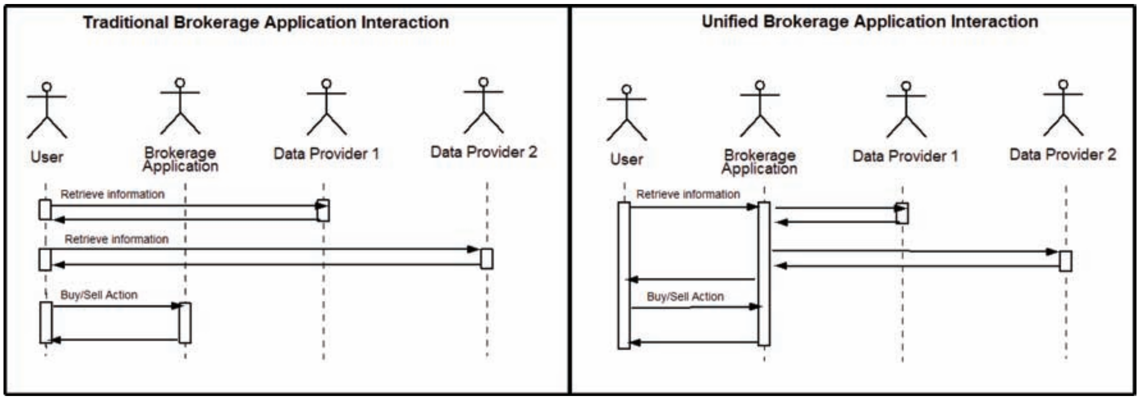
However, the Internet medium, as a cheap and universal way to perform banking operations is the highlighted solution of this business case.

In current brokerage solutions banks usually only offer the service of making the operations (buy, sell) but rarely integrate the search of relevant information to make the transaction decision. The delivery of this 'hard to integrate' functionality search, is, however, a useful to provide to the end user, who wants to buy and sell stock knowing the most relevant and current information. The user usually searches this information on external pages independently of the bank services. Some of these are free services which propose information of the stock market in real-time or with a minimum delay. Web pages such as Yahoo Finance (http://finance.yahoo.com/), Google Finance (http://finance.google.com), Reuters UK (http://uk.reuters.com/home), Xignite (http://preview.xignite.com/), Invertia (http://www.invertia.com) are good examples of financial information from different providers with different degrees of quality.

However, the need to search the information on several Web pages and then navigate to the brokerage application to execute the transaction is a waste of time and adds unwanted complexity to the end user. The idea is to build an online broker that merges and provides a unified or single point of access to information and operational services. In that way, the user will have an unified environment which integrates most of the tools required to fulfil his/her wishes of buying and selling stocks. A comparison of a traditional brokerage application interaction and the new interaction we proposed is showed in Figure 1. In the traditional interaction, the user must retrieve the information from each Data Provider (Yahoo Finance, etc.) independently and only then do the brokerage action (i.e. buy or sell). In the unified environment we propose, all the interactions are performed through the brokerage application and the brokerage application take care of showing the best data information regarding the user context (user profile, portfolio of the user, etc.).

In the next section we define the functional and non-functional requirements of our business

Figure 1. Brokerage application interaction comparison



case. The following sections present one relatively new but already commonly used way of using a service oriented architecture (SOA) to implement these requirements. This solution, however, has some problems exposed later on that we resolve exposing a new and innovative way using SOA combined with Semantic Web Technologies.

Functional Requirements

The functional requirements the brokerage application must support in order to fulfil the business case are summarised below:

- **Stock market consultation functionalities:** The application should be able to retrieve information about the stock market such as the price of a share, volume of a share, historical information, etc. Several sources can be used to obtain the information necessary to the supply the user with the information they require to make the trade.
- **Customer information consultation functionalities:** The application should be able to obtain easily the customer information such as his portfolio, buy and sell history and recent searches by the user.
- **Operational functionalities:** Invocation of operations on the stock market using the bank services: buy, sell.
- **Complex conditional queries:** Possibility to write a complex order in terms of conditions such as "if the stock value of Cisco is higher than X and its volume is lower than Y, ..." that may use different source of information. Logical combinations of the conditions should be possible.
- **Simple entry point of all services:** Complete integration of the conditional queries and operational functionalities within one simple entry point.

Non Functional Requirements

The non-functional requirements that the brokerage application must fulfil are:

- **Highly maintainable:** As the stock market is an entity subject to rapid change, it must be possible to maintain the application in an optimal way.
- **Usability:** The application is aimed at non-expert end users. The application should as usable as possible in order to present to the end user a friendly and easy-to-use interface.
- **Extensibility of the information source:** Possibility to easily add and change the providers of the banking information services.

And to extend and choose the categories of information that the user wants to see. For example, if the user is executing a buy or sell transaction the user may want to see different sets of information.

Solution Based on SOA/ Web Services

Introduction

During the last four decades, software design has been prone to many changes. After abstracting software code from the hardware infrastructure, computer scientists thought to write code in so called black-boxes and invented function oriented software design. The next big revolution was object oriented software design, in which data was intended to be packaged in objects where objects are metaphors of real world entities. Objects were then abstracted in components in order to manage the problem of the increasing number of objects. A component can be defined as a set of objects that has a coherent meaning as a standalone entity. What is the next level of abstraction? A composition of components will always be a component if we only focus on the data that these components contain. The composition must then be thought of a set of components that fulfil a given task. By doing so the packaging is no longer data-oriented but service-oriented, the set of components does no more contain information and methods to access to this information but is a black-box that offers one specific service. The services can then be composed in more abstract services and be part of the entire system, a service-oriented system.

Choosing service-oriented applications allows the clear separation of the users (commonly called 'consumers') from the service implementation (commonly called 'producers'). By having this distinction, the application can then be distributed on several platforms and possibly across networks. Each platform can have its own technology and can be located in any physical place. The software design has fundamentally changed in system design.

The Organization for the Advancement of Structured Information Standards (OASIS) defines the service oriented architecture as follows (MacKenzie et al, 2006):

A paradigm for organizing and utilizing distributed capabilities that may be under the control of different ownership domains. It provides a uniform means to offer, discover, interact with and use capabilities to produce desired effects consistent with measurable preconditions and expectations.

In this definition a service is designed as an entity with on one hand measurable preconditions and on the other hand measurable expectations. It can be reformulated as the input and output of a function in a typical programming paradigm, but in a service the input (precondition) and the output (expectation) are no data but state of the world or effect on the world.

We must clearly separate the architecture from the underlying technology that can be implemented. The Web Services made the fame of the SOA, but there exists other technologies which are totally suitable to be used in a Service-oriented Architecture such as: RPC, DCOM, CORBA, or WCF (Donani, 2006).

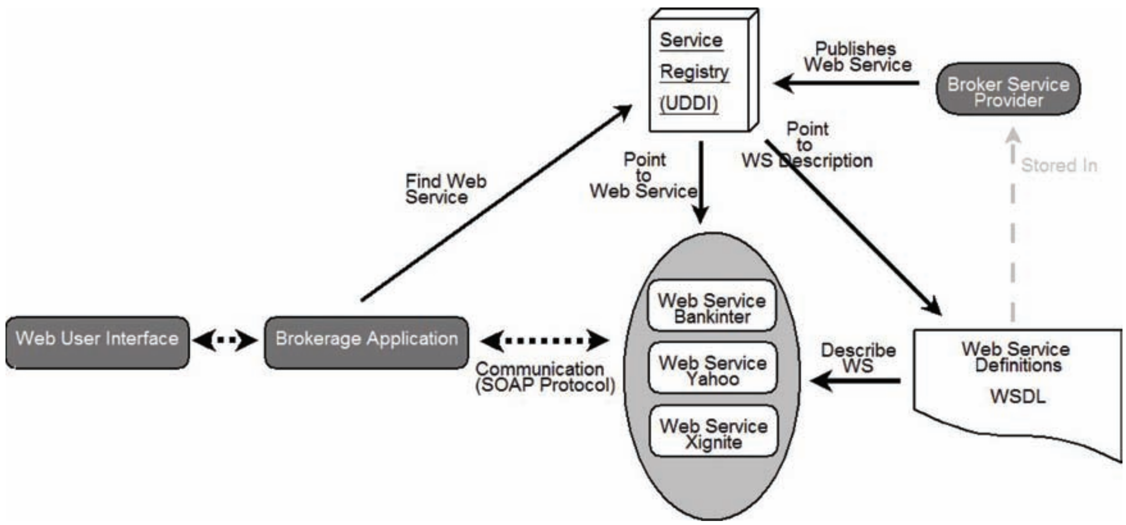
In the following subsection we present a standard SOA for a brokerage application using Web Services as the underlying technology.

Architecture

An overview of the architecture of a brokerage application based on a SOA/Web Service architecture (Booth et al., 2004) is described in Figure 2.

The Web Services are physically located in either the service provider such as Bankinter or Xignite that provides their own Web Services, or in a specific Web Service container in the case that the Web Service provider does not provide a Web

Figure 2. SOA Architecture of a brokerage application



Service Interface and some wrapping mechanisms are needed. This is the case with the above yahoo based Web Service in which analysis (wrapping) of the Web page is needed to extract the right information.

The Web Service Descriptions are stored inside a Broker Service Provider which publishes all the descriptions and is in charge in managing the publication inside a Service Registry.

Through this registry, Web Services can be found and communication between the application and the Web Service is done using the SOAP.

This quite simple and elegant architecture has a lot of advantages that are explained in the following subsections on a technical and business perspective.

Benefits of SOA on a Technical Perspective

SOA is finding increased adoption across more and more business domains. This evolution can be explained by a set of technical benefits. The two most important ones are the following:

- **Reuse:** By decentralizing the systems in self contained atomic Web Services, a SOA allows the redundancy inside the system to be reduced since the Web-services can be used more than once for different purposes. This allows also delivering new functionalities in shorter time.
- **Loosely-coupled:** In a loosely-coupled system, each entity (Web Services) makes its requirements explicit and makes few assumptions about other entities. This permits not being aware where we locate the Web Services and thus increases the IT efficiency (critical Web Services can have an adapted hardware framework), improve the Quality of Service and reduce the costs. Another advantage of a loosely-coupled system is to have standard interfaces. This way there is no need for technical people to know about the whole details of the system and allows a strategically organisational separation of skills inside the project team.

Benefits of SOA on a Business Perspective

As many times the decision of choosing an architecture is done on a business level, it is important for an architecture type to have good benefits on a business perspective. The main SOA advantages are:

- **Business effectiveness:** If one word could be chosen to describe a SOA, it would be the word "agility". By using loosely-coupled services, the responsiveness to market is highly increased. Each process of the system is better controlled and allowed a deployment of resources based on the business needs.
- **Cost efficiency:** Service-oriented architectures Enables reduction of the development costs by separating the skills and efforts in specific development areas. The separation in services allows putting the resources in the technical areas that correspond best to their skills and thus reduce the costs of training or hiring new resources. The maintenance costs of such systems are also highly reduced because of well separated services and technology and location independence. Last cost reduction is the price/performance optimization based on the freedom to select the adequate platform.
- **Risk reduction:** Dividing problems in smaller parts always has the advantage to reduce the risk of a project, thus SOA is based on this division, we can say that the risk of projects based on SOA are inherently reduced. Another point is the risk reduction of the system deployment which can be done incrementally.

Problem Statement

We have presented the benefits of using a SOA based on Web Services to develop brokerage applications. However, there are a number of points in which the standard SOA is unable to respond:

- **Web services heterogeneity:** In the Web Service technologies, the information is described at a syntactical level. For example, in WSDL XML Schema technology is used to describe what the interchanged objects are. However, this kind of technology only allows describing the type of the objects: string, date, integer etc. The semantic of what the objects are is missing. This adds inside the project development a lot of potential problems:

Misunderstanding of what is interchanged (if the documentation is badly done, errors can be quickly done)

Integration problems due to different type definition (for example, one service could describe an address with different fields while the service consumer use an unique string)

- **Poor visibility:** As defined by the standardization group OASIS inside the OASIS SOA Reference Model, the visibility *refers to the capacity for those with needs and those with capabilities to be able to see each other* (MacKenzie, 2006. In standard SOA, the visibility is mainly provided by means of a registry which lists the available services. By having only a syntactic description of the Web Service, the visibility is highly reduced and more efforts in terms of search and analysis are needed by any entity that wants to consume a Web Service.
- **Manual work:** In standard SOA system, an important effort of Web Services integration is needed in order to develop the entire system. The orchestration (composition) of the necessary Web Services corresponds to one the highest effort time spent inside the project due to the mostly manual work that these efforts imply. Middleware is often used in order to solve the orchestration. Mediation (data conversion) is on this topic the major problem. The previously cited visibility problem also implies

a lot of effort time because of the manual effort spent during the localization of each Web Service.

All these points constitute the problems of actual brokerage application based on pure SOA. However, as discussed earlier SOA has a lot of advantages. This encourages us to re-use this solution and improve it. The proposed solution of the next section aims at solving these problems of the actual solution by adding semantic technologies to this SOA paradigm.

SOLUTION: BROKER BASED ON SEMANTIC WEB SERVICES

As seen in the previous section, SOA technology provides a number of powerful concepts that when applied to brokerage applications allow us to construct flexible and easily extensible systems. However, problems have been identified with this kind of technical approach; those problems are responsible for an important part of the cost of such applications. The identified problems were *vocabulary heterogeneity*, *poor visibility* of the services and *manual work* needed in the development and maintenance phases. If we succeed developing a homogeneous vocabulary for Web Services, then we would increase the visibility and reduce the manual work. Three solutions were proposed in (Verma et al., 2007):

- **Pre-agree on all terms (operation name, parameters):** This implies a high oral communication between the development team and a lot of documentation writing. Pre-agreeing on the terms with no formally technical structure implies a high risk for any company due to the risk of losing common knowledge or getting integration problems.
- **Comment all aspects of a service:** In this solution, the comments are added inside the IT components. Each service contains the description of what the service proposes to do. Operation names and the semantic of the parameters are described in natural language.
- **Semantic descriptions:** This solution envisages a formal description of the services, called annotation. The services are not described using natural language but with the proper mechanism of the chosen technology.

This last solution, called Semantic Web Service (SWS), is the one that we propose in this section because it represents the most advanced and thus suitable of these three solutions. The annotation is done by using so called semantic technologies, in which the components of the system (here Web services) are formally described by using semantic resources (usually ontologies). This technology is the base of the vision of Tim Berners-Lee who put the base of a Web where the computer will be able to optimally understand and compute the information (Berners-Lee, 2001). We will explain how the use of ontologies, which is the base of most semantic system, adds visibility to the components by homogenizing the vocabulary used. We will also point out how this enhancement leads to a reduction of manual work and thus a reduction of cost.

Powerful Functionalities

The aim of annotating Web Services is not only to add clarity in the Web Service definitions but also to allow the Web Service to be read by machines. This machine-readability makes the power of the SWS by adding to the system the following functionalities:

- **Power to reason:** The machine is able to “understand” what the Web Service is doing. It is able to interpret the messages that are interchanged. The messages are no more only pure data structure but are structured in such a way that the data can be analyzed and transformed (mediated). For example, if a SWS receives information of a "client" but expected a "person", he is able to infer that a client is a person and is able to extract the right information. The whole information space is structured and coherent. Axioms are responsible of maintaining this space coherent and reasoners are the medium to do this. As described in the functional requirements, the brokerage system needs a system that is able to handle a unique point entry. The reasoning capabilities given by SWS fit perfectly to these requirements, the queries can be interpreted by the system and the system can identify what the user's wishes are. Additionally, as the machine is able to interpret what the input of the user is, it is able to help the user at the moment of maintaining the system. For example, a company has a brokerage application that allows the user to buy stocks. However, for marketing purposes the number of times that the customer can buy depends on the profile of the customer. In this case, the maintainer of the brokerage system could want to add a new type of client. With a SWS-based system, this is highly simple, as the only action to do is to add the concept of this new type in the ontology and add the information on how many operations he can do. No additional development efforts are needed and no risk of adding errors in the application is run. If the maintainer makes an error in adding the information inside the ontology, the reasoner will warn him before he put it in production.

- **Automatic discovery:** Formally describing what each Web Service does adds the functionality of automatically discovering them. This means that a query written in a formal language can be interpreted so that the correct Web Services with the appropriate functionalities and Quality of Service parameters are found. This allows a better decoupling between: what the user wants and what the system proposes. By separating these two parts, the system gets more flexible.This SWS functionality responds to the functional requirements about the processing of complex queries. The user expresses a complex query in Natural Language or through a Web Interface, this query is translated into the corresponding formal language and this formal query can then be used to retrieve the best Web Services. More than one Web Service can be accessible for the same functionality; the system takes care at choosing the more adapted one.In terms of maintainability, the separation of the SWS is also important in the sense that adding new duplicated SWS of other providers does not require modifying the application. If new types of SWS are added, some extents must be added to the query generation functionality of the brokerage application. But this task remains relatively easy because of the use of ontologies which takes care of the coherence of the system. The automatic discovery responds to the non functional property of the "extensibility of the information source".
- **Automatic orchestration:** SWS support the automatic orchestration or composition of Web services (Medjahed, 2003). By orchestration, we mean the composition of the Web Services in order to provide a more complex service. As Semantic Web Services are semantically annotated, the system has enough information to handle a user query and respond to it by assembling the Web Services.Automatic orchestration provides an easy way to combine a usable interface to the user, with one entry point that provides the three main requirements:

stock market consultation, costumer information consultation and operational invocation.

SWS Technologies

The Semantic Web Service technologies have been in the last few years under intensive research world-wide. In the actual states, two approaches have been developed. Each one of these approach were part of research projects and their validity was proven by the deployment of concrete use-cases. The two approaches are:

- **Pure Semantic Web Services:** These technologies represent a way to write pure Semantic Web Services. By pure, we mean that they are written directly in a formal language and are independent from any non-Semantic Web Services. Of course, all SWS technologies need to be able to be connected with non-Semantic Web Services (called grounding) in order to support any already developed Web service system. But the idea is to be able to build new Semantic Web Services that will not carry on the "old" non semantic technologies. There are two main technologies based on this approach: OWL-S (Martin et al, 2003) based on the OWL ontology language and WSMO (Fensel et al., 2007) (de Brujin, Bussler et al., 2005)based on the WSML (de Brujin, Fensel et al., 2005) ontology language. The first is mainly a North American development effort, while the second one has been developed within EU-funded projects (Sekt, DIP, Knowledge Web, ASG and SUPER projects). They both are submitted to the W3C and have the necessary specification, development tools and execution engines.
- **Semantic Annotation of Web Services:** The second approach consists in directly annotating the WSDL with semantic information. Two main specification efforts are actually done: WSDL-S (Akkiraju et al., 2005) that is at the Member Submission stage in W3C and SAWSDL(Farell et al., 2007) that is a W3C proposed recommendation. Main advantage of these approaches is that the annotation is done directly in the WSDL / XML Schema. Thus, the evolution of existing systems is facilitated. Other advantage is that these specifications are ontology language independent, thus execution engine can be developed for any chosen ontology language. Both languages have the necessary development tools.

Tools already exist and are operational to model and run Semantic Web Services. Most of them were part of a research project and are freely available on the Internet.For modelling SWS, the following tools are available:

- **WSMO Studio[1]:** A SWS and semantic Business Process Modelling Environment. Also support SAWSDL. As described by the name, this tool supports the WSMO Framework. It is Eclipse-based and the last version is 0.7.2 released on 29/11/2007 (in the moment that this chapter is written: end of 2007)
- **Web Service Modelling Toolkit (WSMT)[2]:** A lightweighted framework for the rapid creation and deployment of the tools for SWS. It supports WSMO Framework. It is Eclipse-based and the last version is 1.4.1 released on 13/09/2007 (in the moment that this chapter is written: end of 2007)
- **Radiant (Gomadam, 2005)[3]:** A WSDL-S / SAWSDL Annotation Tool developed by the University of Georgia. The annotation is made using OWL ontologies. It is Eclipse-based and the last version is 0.9.4beta released on 29/05/2007 (in the moment that this chapter is written: end of 2007).
- **ODE SWS (Corcho, 2003)[4]:** A toolset for design and composition of SWS. It is based

on UMPL and some work has been done to integrate OWL-S.

- **OWL-S IDE (Srinivasan, 2006)[5]:** A development environment supporting a SWS developer through the whole process from the Java generation, to the compilation of OWL-S descriptions, to the deployment and registration with UDDI. The last version is 1.1 released on 26/07/2005 (in the moment that this chapter is written: end of 2007).
- **OWL-S Editor (Elenius, 2005)[6]:** A Protégé Ontology Editor plugin for a easy creation of SWS. The last version was released on 04/11/2004 (in the moment that this chapter is written: end of 2007).

The following SWS Engines are available:

- **WSMX[7]:** The reference implementation of WSMO
- **Internet Reasoning Service III (IRS-III)[8]** (Domingue et al, 2004)**:** A SWS framework, which allows applications to semantically describe and execute Web services.
- **OWL-S tools:** A series of tools WSDL2OWL-S, Java2OWL-S, OWL-S2UDDI, etc. are available at: http://www.daml.ri.cmu.edu/tools/details.html

Through these tools represent good proofs of the viability of the technology. That said, further development would be required to adapt them to the needs of real world system. Professional benchmarks would be needed to identify efficiency and security lacks and allow the development of professional SWS frameworks.

We gave a short overview of the existing SWS technologies and we explain now how these technologies can be applied to brokerage applications.

Approach and Architecture

The approach taken for creating a brokerage application with SWS is to use the SWS engine as a central component of the architecture. By taking advantage of the reasoning capacities of the SWS engine, it is possible to build a simple and extendible Brokerage Application. New Semantic Web Services are added directly in the engine and we minimize the development costs of managing new services. The SWS engine "understands" the semantic of the new added SWS and only few modifications are needed inside the Brokerage Application itself. This approach has been proven during the DIP project on a use-case (see the two screenshots Figure 3 And 4) that simulated a brokerage application with Bankinter and external Web Services.The user can enter a complex query composed of several conditions and one action to be taken. The conditions are connected with logical operators (AND/OR). The conditions can be of the types:

- *If the price of a specific stock is higher than a given price.*
- *If the value of an index is lower than a given value.*
- *If the expert recommendation is equal to a specified one.*
- *If the variation of the value of a given stock is higher.*

For each information that needs to be retrieved, the SWS Engine is responsible for discovering the best suitable Semantic Web Service, eventually by composing multiple Semantic Web Services (orchestration) and invoke the one (/ones) that correspond to the given Quality of Services parameters (time to respond, localization, etc.). If a Semantic Web Services is grounded on some other service systems (like normal Web Services), it is in charge of getting the information and converting it into the semantic language. The Brokerage

Figure 3. First screenshot of the SWS based brokerage prototype

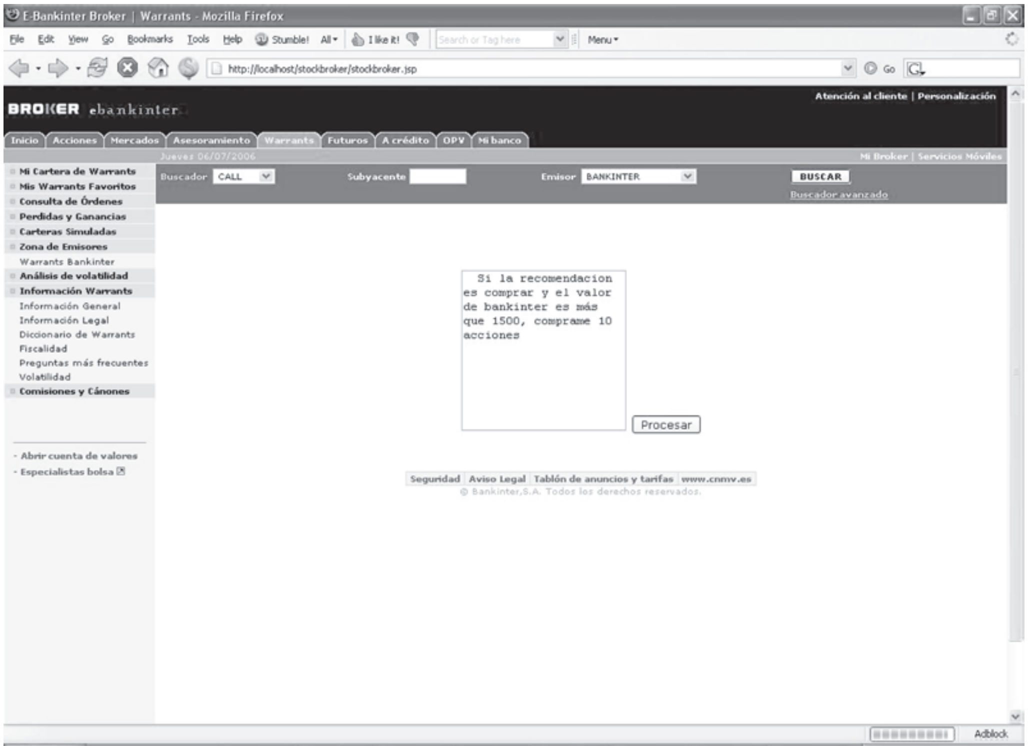
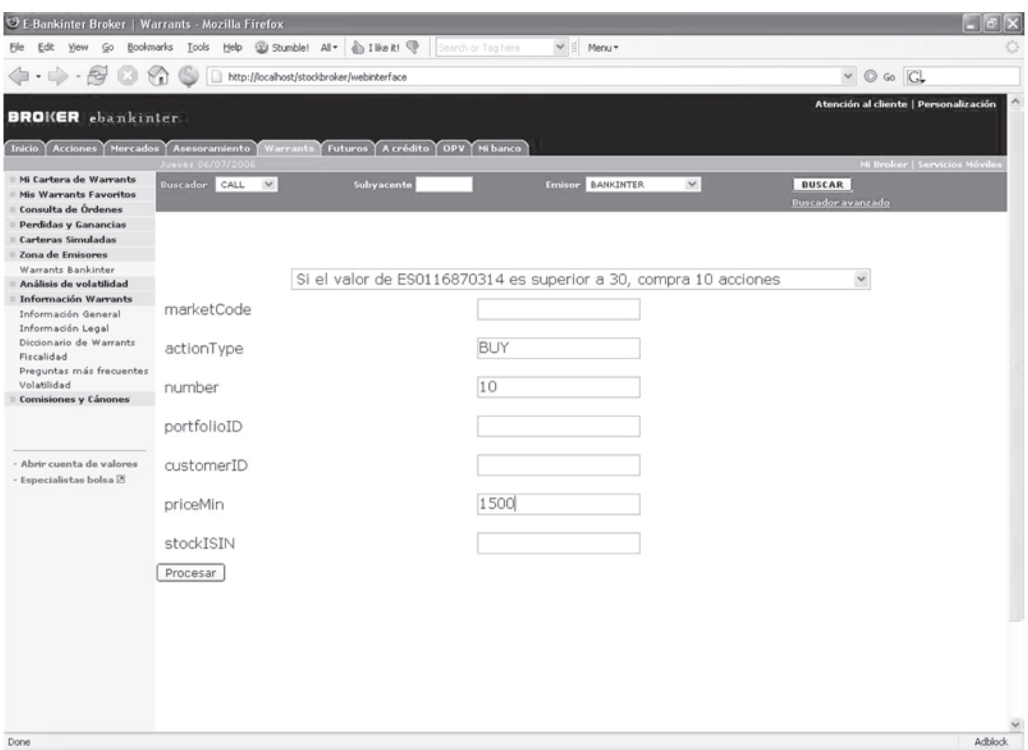


Figure 4. Second screenshot of the SWS based brokerage prototype



application then returns the result of the execution to the Web Interface.

Figure 5 shows the architecture that implements this approach.The three main components are:

- The Web User Interface that should respond to the Usability non-functional requirement.
- The Brokerage Application that should support all the functional requirements and that is in charge of the communication with the SWS Engine.
- The SWS Engine that is in charge of managing the semantic resources: discovering, invoke and orchestrate the SWS.

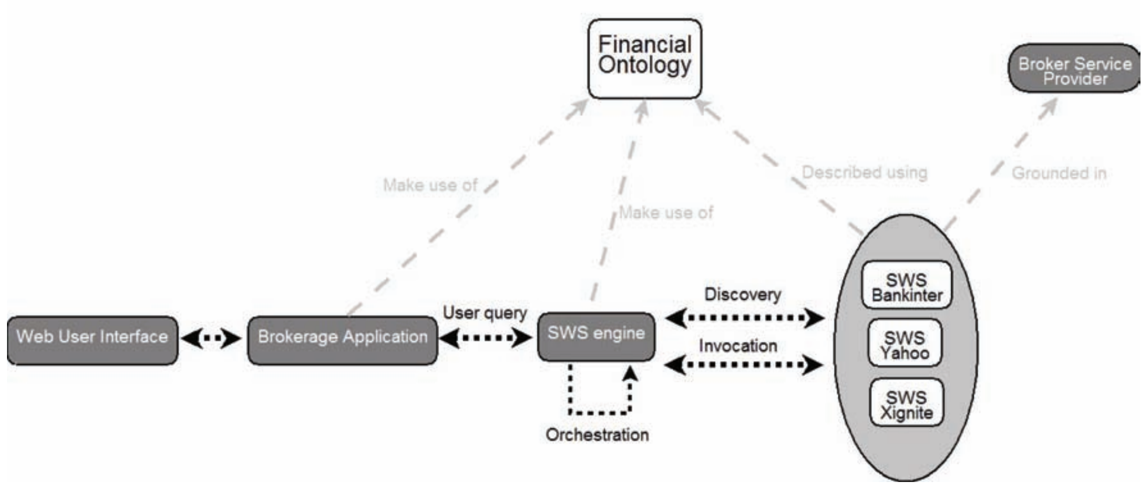
The brokerage application prototype developed in DIP has been developed using J2EE technologies. The application makes use of Natural Language Processing technologies to offer a simplified interface to the user. Receiving one sentence as input, the brokerage application is able to identify what is the user intention and automatically retrieve the information that it needs to invoke the SWS. These parameters are used to generate a WSMO goal (de Brujin, 2005) formally describing the user intention. This goal is the entry point to the SWS Engine.

As SWS Engine, WSMX (Bussler et al., 2005)[9] was chosen over IRIS-III (Domingue et al., 2004) (Cabral et al., 2006)[10] in order to prove the correct implementation of this SWS Engine inside the research project. From the input goal provided by the brokerage application and some optional Quality of Service parameters, the SWS engine discovers the necessary SWS and invokes the retrieved SWS in the right order (orchestration). The brokerage application can then have access to all the information it needs to check the condition provided by the user and if the conditions are validated execute the buy/sell order through another call to the SWS engine.

The Financial ontology, exhaustively described in (López-Cobo, 2008) plays a major role in all the tasks of the SWS engine and is the pillar of the whole brokerage application. It describes the vocabulary of the application and is used to annotate the SWS on both levels: the functionality description (capability) and the interface (message exchange).

By using an architecture based on the Financial ontology and the SWS engine, we provide a flex-

Figure 5. Brokerage application architecture



ible and maintainable application and provide to the brokerage system the whole benefits of using SWS technologies.

In the next section, we describe in a higher level the cost and benefits of adopting such architecture.

COST AND BENEFITS

From a business point of view, the profits of the proposed solution must not be focused on new incomes neither on costs, although they both exist. The resulting application is intended to create a new product, by giving new options to manage their portfolio. These options could have been developed using a more traditional approach but, due to the complex and usually mature architectures used by the financial institutions, the costs would have been significantly higher, both the development and the future sub-applications costs.

Also, a more traditional approach (i.e. without semantic technologies) would have implied agreements with the information providers (data formats, relevant data, how to provide the information, how to access to the data, etc), which will usually implied a one-by-one Trading Partner Agreement (TPA). The semantic layer gives us the ability to smartly read the provided data and therefore to manage it easily. It is also easy to add new providers with this approach. Finally, data is accessed when required and if required, making use of the Web Service advantages.

The costs and benefits, from the technical point of view, must be also considered. The Cost/Benefits ratio in terms of adapting actual systems, although not trivial, is not as dramatic change as the one that was performed in the transition to the Internet era.

Banks were usually based on main frame architectures. The scenario in these last 10-20 years has changed from a exclusive main-frame scenario, where only the bank employees had access to the IT transactional systems, to a Web Based scenario where virtually any customer, anywhere at anytime could be using the bank transactional. This transition meant high investments on scaling the main frame architecture to a 24*7 architecture adding several layers: Web servers, application servers, database servers, security layers...

Fortunately a Web Service scenario is more natural in the actual client-servers environment, therefore, in terms of cost/benefits the investment is lower. The same reasoning can be applied to a SWS scenario, where besides the new Web Service layer to be added, semantic pieces appear to complete the puzzle.

Therefore, taking the chance of adding semantic layers to an existing bank architecture implies low economical risk since no major implications are needed to expand the current architecture.

As a result, and taking into account the business opportunity, the small amount of effort required to create and maintain the service, and the technical prerequisites, the SWS approach emerges as a smart solution to create the new service at a reasonable level of cost, both for the developer and, which is more important, for the final user.

RISK ASSESSMENT

From a business point of view, alternatives are almost always more expensive but it could depend on how each Financial Institution manage its own Stock Market Services. What is more, in several cases the solution could be so complex that it could be considered as ‘nearly impossible’ to develop without studying in depth what is going to be modified, (Stock Market applications are critical tasks for Financial Institutions due to the volatility of the market and the quality assurance that is required in this specific market).

At the same time, the new proposed service is intended to give better utilities to the clients. These kinds of services are actually free, although the final user must manage with them. So the price of the service for the final clients could not be high and thus, it has to be developed at a reasonable cost.There are some risks when the use of SWS is considered:

- **Security issues:** No doubt this is the major functional risk to be considered when deploying SWS technologies. On banking environments security is the mayor column on which all the architecture must be built. All the security issues must be clarified and solved before any transactional application using real customer banking data is deployed to the real world. If this milestone is not achieved all the SWS-based applications will be forced to handle only with public data and the real value of semantic applications will remain as a proof-of-concept not as a real-world-application.
- **Evolution issues:** SWS techniques are in their first steps of use in business environments. As these techniques become more familiar they will evolve and this evolution could mean scalability issues that should be treated as any other scalability issues inside any corporation environment. Although the semantic techniques are mature their wide use could imply changes that would mean changes on the semantic platform. However these two evolution issues are natural to any IT development or to the deploy of any new technology. The IT business, no matter if it is the banking business or any other, has got enough experience to handle these potential risks.

FUTURE TRENDS

The evolution of the Stock Market and its associated services must be forecasted as part of the global social tendencies: people (and investors) are requiring more and more sophisticated products and services allowing them to make their own decisions. If we consider that information aggregators are the Internet 'killer applications' (i.e.: Google, Yahoo, You Tube, etc) investors are expected to make use of Stock Market data aggregators (in fact, the actually use them: Yahoo! Finances, Invertia, etc), making their own buying/selling decisions and finally performing them in their favourite Stock Market site.

None of these services are designed to perform automatic operations that completely fulfil the investor requirements, nor of them are prepared to perform a personalised strategy when data aggregation is required. There are several solutions for professional investors but they are available at a high cost, thus they are only interesting when high volumes are regularly performed (high volume both in terms of number of transactions and in terms of money invested). But there they are not an option for individuals.

Our tests[11] reveal that people usually make use of at least two Stock Market services just to fulfil their information requirements. The SWS is intended to aggregate them and, in the near future, to automatically perform the actions according to the investor strategy, combining sources and retrieving specific data form them.

CONCLUSION

Responding to the need of maintainable and efficient brokerage applications, we have presented in this chapter a novel approach that combines the SOA architecture and the semantic technologies. By using the proposed solution, we resolve the three identified problems of a non-semantic SOA solution: heterogeneity of the vocabulary used in the services which reduces the maintainability and possibilities of evolution of the application, the poor visibility which reduces the possibility for automatic discovery and the lot of manual work that is generated by this poor visibility and the lack of automatic composition possibilities.

We exposed the advantages of a solution based on Semantic Web services: reasoning functionalities, automatic discovery and automatic orchestration. Such functionalities allow us to build an architecture centralized on the SWS engine and have a really flexible system.

The feasibility to build such brokerage application has been demonstrated during the European project called DIP. The output of these projects is a framework called WSMO associated with an ontology language (WSML) and execution engines (WSMX and IRS-III). The European Commission continues to invest money in the research in Semantic Web Services for example in the SUPER project which aims to take advantage of the Semantic Web Services in order to improve Business Process Management Systems. A lot of research is also done in other technologies such as OWL-S. The high activity of these research projects reflects the important interest that should have industrial investor in such technology. The Semantic Web services are ready and continue to be improved.

Companies should consider the benefits of SWS on two levels: the strategic level and the tactical one. On the strategic level, the SWS give the possibility to build highly maintainable applications and profits from a loosely coupled architecture. On a tactical level, a company should see the benefits on other projects where the ontologies that have been created for the SWS are reused and form a common base for the applications. Using such Semantic technologies is interesting for companies because it promotes the homogeneity of the systems inside the company. The intra-company and inter-company applications are then much easier integrable.

REFERENCES

Akkiraju, R., Farrell, J., Miller, J., Nagarajan, M., Schmidt, M., Sheth, A., & Verma, K. (2005) *Web Service Semantics - WSDL-S,*

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 284*, 34–43.

Booth, D., Haas, H., McCabe, F., Newcomer, E., Champion, M., Ferris, C., & Orchard, D. (2004) *Web Services architecture*.

Bussler, C., Cimpian, E., Fensel, D., Gomez, J. M., Haller, A., Haselwanter, T., et al. Zaremba Maciej, Zaremba Michal (2005). *Web Service Execution Environment (WSMX)*.

Cabral, L., Domingue, J., Galizia, S., Gugliotta, A., Norton, B., Tanasescu, V., & Pedrinaci, C. (2006) *IRS-III: A Broker for Semantic Web Services based Applications*, The 5th International Semantic Web Conference (ISWC 2006), Athens, GA, USA, Corcho, O., Fernández-López, M., Gómez-Pérez, A., and Lama, M. (2003). *ODE SWS: A Semantic Web Service Development Environment*. VLDB-Workshop on Semantic Web and Databases, 203-216. Berlin, Germany.

de Brujin, J., Bussler, C., Domingue, J., Fensel, D., Hepp, M., Keller, U., et al. (2005). *Web Service Modeling Ontology (WSMO).*

de Brujin, J., Fensel, D., Keller, U., Kifer, M., Lausen, H., Krummenacher, R., et al. (2005). *Web Service Modeling Language (WSML).*

Domingue, J., Cabral, L., Hakimpour, F., Sell, D., & Motta, E. (2004). IRS-III: A platform, and infrastructure for creating WSMO-based Semantic Web services. In *Proc. Of the Workshop on WSMO Implementations.* CEUR Workshop Proceedings

Donani M.H. (2006). SOA 2006: State Of The Art. *Journal of Object Technology*, 5.

Elenius, D., Denker, G., Martin, D., Gilham, F., Khouri, J., Sadaati, S., & Senanayake, R. The OWL-S editor- A Development Tool for Semantic Web Services. In *The Semantic Web: Research and Applications*. Series: Lecture Notes in Computer Science, Vol. 3532, 78-92. Springer Berlin / Heidelberg 2005. ISBN: 978-3-540-26124-7

Farell, J., & Lausen, H. (2007) *Semantic Annotations for WSDL and XML Schema*.

Fensel, D., Lausen, H., Polleres, A., Brujin, J. d., & Stollberg, M. Romand D., Domingue J. (2007). *Enabling Semantic Web Services – The Web Service Modelin Ontology*. Springer.

Gomadam K., Verma K., Brewer D., Sheth A. P., Miller J. A. (2005) Radiant: A tool for semantic annotation of Web Services, *ISWC 2005*.

López-Cobo, J. M., Cicurel, L., & Losada, S. *Ontology Management in eBanking applications.* In *Ontology Management. Semantic Web, Semantic Web Services, and Business Applications*. Series: Semantic Web and Beyond, Vol. 7. Hepp, M.; De Leenheer, P.; de Moor, A.; Sure, Y. (Eds.) 2008, Approx. ISBN: 978-0-387-69899-1

MacKenzie, M., Laskey, K., McCabe, F., Brown, P. F., & Metz, R. (2006) *Reference Model for Service Oriented Architecture 1.0.*

Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., McIraith, S., et al. (2003). *OWL-S: Semantic Markup for Web Services.*

Medjahed, B., Bouguettaya, A., & Elmagarmid, A. K. (2003). Composing Web services on the Semantic Web. *The VLDB Journal*, *12*(4), 333–351. doi:10.1007/s00778-003-0101-5

Srinivasan, N., Paolucci, M., & Sycara, K. (2006). Semantic Web Service Discovery in the OWL-S IDE. In *Proceedings of the 39th Annual Hawaii international Conference on System Sciences - Volume 06* (January 04 - 07, 2006). HICSS. IEEE Computer Society, Washington, DC, 109.2.

Verma K., Sheth A. (2007). Semantically Annotating a Web Service. *IEEE Computer Society*, March/April 2007, 11, 83-85.

ENDNOTES

1. http://www.wsmostudio.org/download.html
2. https://sourceforge.net/projects/wsmt/
3. http://lsdis.cs.uga.edu/projects/meteor-s/downloads/index.php?page=1
4. http://kw.dia.fi.upm.es/odesws/
5. http://projects.semwebcentral.org/projects/owl-s-ide/
6. http://owlseditor.semwebcentral.org/
7. http://www.wsmx.org/
8. http://kmi.open.ac.uk/projects/irs/
9. http://www.wsmx.org
10. http://kmi.open.ac.uk/projects/irs/
11. Done inside the DIP (http://dip.semanticweb.org/) project, Deliverable 10.10.

This work was previously published in Semantic Web for Business: Cases and Applications, edited by R. García, pp. 336-352, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 4.7
Innovating through the Web:
The Banking Industry Case

Chiara Frigerio
Università Cattolica del Sacro Cuore, Italy

ABSTRACT

In recent years, the financial services industry has been witness to considerable consolidation (Berger & Udell, 2006; De Nicolò, Bartholomew, Zaman, & Zephirin, 2004; Figueira, Neills, & Schoenberg, 2007) and organizational progress in order to sustain two main objectives: efficiency and commercial effectiveness (Epsten, 2005; Sherman & Rupert, 2005). In order to sustain customer-oriented and efficiency strategies, banks have started to explore new ways of conducting their business, introducing areas of innovation in their services, practices, and structures to offer the most complete array of services possible (Quinn et al., 2000). On the other hand, new services and products drive retail banks to explore new ways of producing or delivering these novelties. This is true especially for Internet banking services that offer services to customers 24/7, and it becomes clear that adding new services, that is, trading online or bill payments, is easily and quickly geared towards improving commercial effectiveness. The following chapter aims at describing to what extent the Internet has developed new services and businesses, and what are the main figures of the phenomenon in Europe. Moreover, the Internet has introduced new coordination processes within each financial institution. Let us think about Intranet portal, content management tools, and business process management suites, which are now quite spread in banks due mainly to their technological ease-of-use. Thus, Internet is representing an innovation wave extremely relevant for the financial industry as a whole, and the effects on banks' performance is emerging. What do we expect in the near future? In all probability, the usage of Web-based application will be bigger and bigger also in other contexts of the bank processes, even if some risks could occur when clear strategies and change management practices do not direct the innovation.

INTRODUCTION: INNOVATION AND WEB BANKING

The banking system is undergoing remarkable strategic and organizational change processes, which will consolidate and spread even in the future. Such changes are not to be considered as a contingent change in the design and development criteria of the business, but they point out an innovation character which extends far beyond those boundaries. The application fields of innovation criteria are numerous (e.g., technological, organizational, commercial) even though two main critical issues often hinder

DOI: 10.4018/978-1-60566-024-0.ch012

the creation and spread of innovation: a) the origin of innovation, that is, the characteristics of the innovation source; and b) the possibility to know beforehand the effects of innovation and to keep its sustainability in the long run.

With reference to the subject of the origin of innovation, the most traditional theoretic version counters the organizational process innovation with the innovation of product and service. "Product innovation is given by the introduction of a new product or service to meet the market request or a particular external customer, while process innovation is defined as a new element which is introduced into the organizational processes in order to produce a product or to distribute a service" (Damanpour & Gpalakrishanan, 2001, pp. 47-48). Organizations develop product or service innovation in order to increase their market shares or to improve their strategic positioning, while they innovate in the process in order to achieve economies of scale and increase profitability (Utterback & Abernathy, 1975). The innovation of the first type allows the development of the first phases of the life cycle of a product/service, therefore it involves relevant investments and high risks. On the contrary, organizations that operate process innovations generally develop themselves at the time in which the product/service is in a maturity phase, therefore in the long term. For this reason, process innovation generally requires less investment (Anderson a& Tushman, 1991; Barras, 1990).

Moreover, by overcoming the dichotomy between product/service innovation and process innovation it is possible to consolidate an approach which takes into consideration the combination of both drives within an innovation policy (See Figure 1). Therefore, both product/service and process innovation may be combined and may generate added value within a short period and with minor costs (board 3). Therefore, in some cases, the ability to innovate is measured both in the product/service part as well as in the process one. The convergence of the innovation strategic approaches is ensured by organizational mechanisms which have the role of coordinating and managing skills, relations, and innovation processes as a whole.

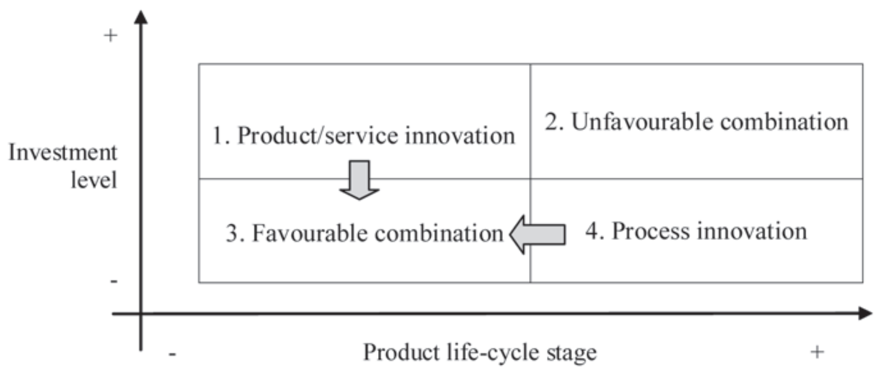
With reference to the second critical issue source, managerial literature associates to innovation the characteristics of "radicality." An innovation is radical when its adoption process over time develops in nonlinear maturity phases, characterized by times of performance growth on discontinuous effects. The main obstacle in the innovation management is given by the difficulty of previously predicting moments in which a performance leap and the relevant extent of the performance growth will take place. The effects of innovation may be even more uncertain if the innovation recipient is a little controllable, being it external (e.g., the customer). As a matter of fact, the lack of control over the external party prevents knowing the cause-effect relationship of the leverages affecting the innovation adoption process. Such leverages may also be generated or limited by solicitations which are outside the bank-to-client relationship, difficult to foresee and to manage.

All the above mentioned introductory considerations may be useful in evaluating one of the main innovation areas developed by the banking industry over the last 20 years, that is, the Web.

As a matter of fact, the introduction of this technology immediately generated a great application interest from financial institutions and was directed mostly towards:

a. Development and management of both relationships with the outside (i.e., stakeholder and clients) and the inside (i.e., cooperators and employees), through the use of the Internet and the Web as innovations in the publication and circulation of contents or services.
b. The development of "core banking" applications and information systems in Web-based environments characterized by integrability and pace of important achievements.

Figure 1. Product/service/process innovation



This chapter aims at exploring the first applicability area, by describing how and in which way the Web development in such a context has been innovative. Specifically, it analyzes first the characteristics of *Internet banking* and then those of *Intranet banking*. By the first term, we refer to the development of distribution channels of product/services alternative to the physical presence, based on Web technology and distributed through Internet. In this area, product innovations went and go hand-in-hand with process/channel innovations. As a matter of fact, certain products specifically created for the Internet channel, like c/a, loans, and financial services, are supported by new management processes, aimed at providing procedure efficiency and standardization. The joint product and process innovation enables to support further innovation strategies aimed, on the one side, at creating increasingly customer behaviour-oriented products and, on the other, at improving processes even in nonefficient areas. To support the theory concerning the mutual relationship of product and service innovation, one may consider the evidence substantiated by the "multichannel service strategies," characterized by the presence of cross-channel and cross-customer product/services and processes.

Therefore, though the evolution of the Internet, the banking phenomenon is continuous and steady; another application area of the Web use, less visible but equally active, is the one of Intranet banking. Intranet is an information support available for members of an organization aimed at providing easy and immediate access to organizational and training information, thus reducing distribution costs and the level of information dispersion which potentially increases the job complexity and employee dissatisfaction.

The development of information technology-based innovations, like Web banking, raises the issue of the importance of an assessment of the threats of technology innovation. It is enough, for instance, to look at the problem of trade (products and channels), organization (production processes and innovation delivery), and information (data and information flows) integration, deriving from the implementation of an Internet strategy. The lack of integration represents one of the main defeating threats of the innovation implementation process (Ciborra, 1996). As a matter of fact, the latter is effective only when it is a part of a whole and is not out of context. Another technology innovation challenge is represented by the speed and the direction of the standard implementation process. Specifically in the Web banking industry, the interoperability requirements are essential. Technology innovations which may vary the widespread standards are particularly critical and difficult to implement.

INTERNET BANKING: A HISTORICAL SUMMARY

Internet banking started growing at the beginning of the 90s and saw its maximum development at the end of that same decade.

Its importance in the definition of bank commercial and marketing strategies was immediately clear. The development of applications which would allow customers to remotely operate by means of their home PC was minimally complex and fairly cheap. But the decisional factor of the main business plans on Internet banking was the evaluation of the transaction marginal efficiency. As a matter of fact, since the first analysis it was clear that the cost of the process of a transaction opened on a virtual channel was considerably lower than the physical one (e.g., that of a promoter branch agency). This scenario initially led some banks and then most of them to adopt an Internet strategy.

Especially in the leading banking groups in Europe, the potentials made available by Internet led to the creation of units fully dedicated to the management of virtual customers (both retail and corporate). Such units were often legally autonomous companies, that is, subsidiaries of the parent bank. However, the strategy of direct banks (i.e., those which were exclusively present on the virtual channel) does not have a long life. As a matter of fact, on the one hand, besides the totally dedicated channel, banking groups keep the possibility of operating their accounts via alternative channels (often by using the same virtual bank technology). On the other hand, certain companies born to represent the mere virtual channel of their parent bank start creating, little by little, within a multichannel perspective "light" branches with employees able to support customers. This was, for example, the Italian case of Banca 121. The current situation is characterized by a higher number of universal banks integrating Internet service into their offer. Only a few operators are solely virtual and they are generally specialized in online trading. This last service has long been considered by most operators as the killer-application of Internet banking. At a time when investments in negotiable securities had high returns, online trading developed considerably and, in certain cases, it replaced the figure of the financial promoter or of the financial manager. Subsequently, the development of further services, not only linked to brokerage, allowed also the development of more complete Internet strategies for retail banking (e.g., c/a transactions, circulars, requests for consulting support) and also for private banking, characterized by sophisticated and demanding clients.

The e-commerce phenomenon for banks has not yet given the expected results. Some banks have activated e-commerce initiatives both in the business-to-business (B2B) as well as in the business-to-consumer (B2C) areas. English and Spanish banks proved to be the most active in this situation. Besides payment management, often accompanied by the creation of special credit cards in the case of B2C, banks also offer partnership formal guarantee services (especially in the case of B2B portals) and of special financing. We have progressively witnessed the passage from an activity of incubator and partnership, still present in the less developed areas, to the activity of facilitating or aggregating, though the few existing cases do not allow us to consider e-commerce for banks as a consolidated phenomenon.

THE PRESENCE ONLINE: AN ANALYSIS OF WEB SITES

A recent survey carried out in 2003 allowed the comparison of the Web world experience of the

European banking key-players, with particular reference to strategies and behaviours (Carignani & Frigerio, 2003).

The survey targeted seven European countries (i.e., France, Germany, Italy, Scandinavian Countries, Spain, Switzerland, UK), selected among those having a higher level of Internet usage penetration and characterized by important investments through alternative channels. To this purpose, the survey analyzed: first the German market, considered as the widest and most developed market of the European scenario; the British one, whose financial and banking tradition is, in many aspects, similar to the US banking system; the Scandinavian system, important for the extremely widespread Internet use among the population; and the Swiss one, characterized by a peculiar banking tradition in retail and private banking. Other significant experiences closer to the Italian context are the one of France, characterized by a good spread of alternative channels, and the one of Spain which, regardless of the disadvantage compared to its European competitors, is now coming up strongly into the direct banking sector with interesting offers.

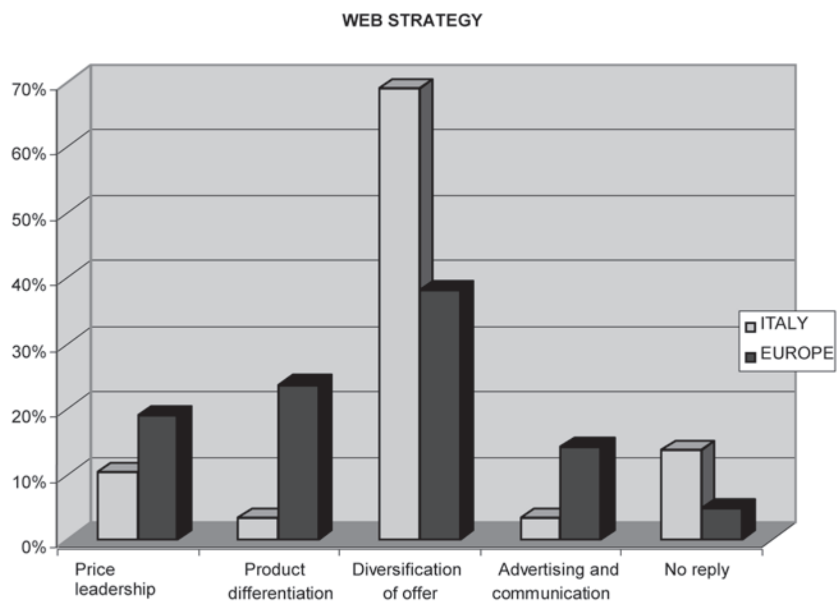
With regards to the individual banks selected within each chosen country, in Germany, whose banking system still represents today the reference model of all operators of the Old Continent, and in the UK, the survey investigated traditional banks which chose to invest hugely in new channels by implementing rather aggressive strategies, and the most significant examples of direct banks born as spin-offs of the major banking groups of their respective countries. This way it is possible to monitor and compare the direct banking initiatives of both leading banks and followers. With reference to the French market, the survey focused merely on the main traditional banks in terms of customers, as well as on volumes and on the three most successful initiatives in the e-banking field. In the Scandinavian banking industry, which, as previously mentioned, is strongly geared towards the use of innovative channels, the attention was focused on those banking institutions which have adopted the most interesting and innovative initiatives from a technological point of view and which, for this reason, may represent a reference model for banks of other countries like ours, where Internet penetration is still low, though growing, and the population is still little accustomed to use technological channels. Considering the central role given to the Italian situation within our research, a wide sample of banking institutions was selected. In particular, attention was given to those banks which form part of the major groups of the country, as well as to single institutions having a considerable volume of customers.

From the analysis of data provided in the field of the strategic choice made, both Italian and European banks show the trend of looking at electronic channels as a way of diversifying their offer. The search for new services which may be offered online, thus contributing to increasing the added value given by the bank, is pursued by 68% of Italian banks and 38% of European banks. Price leadership is a primary value just for 10% of Italian banks (typically those which are merely remote) and of 19% of foreign banks. This value makes even more sense if it is compared to the result given by the same survey carried out in 2000. As a matter of fact, in that occasion, it appeared that price reduction represented the main objective of Italian banks, followed by the advertising channel. Moreover, 24% of European banks are aiming at product differentiation, by introducing new products and services which have a different configuration for the Web.

Going more into details with regards to the implemented strategies, the types of services offered have been analyzed, also comparing the offer of other transfer services by remote banking.

All European and Italian banks which took part in the survey declared they offer online transfer services, information, and value-added services. Call centers basically provide the same offer variety. Actually, the range of operations which may be carried out through the use of the most recent

Figure 2. Internet banking strategies



e-banking technologies is not so wide. Among banks which are equipped with Wap technology, only 70% are able to offer transfer services but only 30% more innovative services, which bring a value added to the overall offer of the bank. SMS messages are used mainly as an information tool. Moreover, the few banking institutions which have extended to the Web TV/digital television their interaction means with customers are trying to develop transfer services which are mostly of transactional type.

The online trading expansion has pushed banks to disclose market information. This information area is mostly explored by foreign banking institutions; to this effect, 94% of European sites, compared to 84% of Italian sites, dispose of a dedicated area. The majority of banks offer services for customers while visitors may consult only stock exchange quotations. The service is almost exclusively supplied by a specialized information provider.

Banks also offer many value-added services; simulation tools and communication services (e.g., forum, newsletter, mailing) are among the most available services. Seventy-four percent of Italian Web sites and 53% of foreign Web sites taking part in the sample offer financial simulators which support customers in the most efficient portfolio choices. Here again, one can notice the will of our virtual banks to aim at clients information. In Europe, the percentage is higher in Switzerland and Great Britain where banking institutions have a long tradition in managing savings. Discussion forums are present in some banks, especially in Central Europe, but they are also rapidly developing in southern banking institutions.

Financial information is the most widely value-added service (100%) offered by Italian banks, followed by product information (93%), market trend graphics (90%), and online help like FAQ, e-mail, and demos, which represent 86%, 86%, and 79% of the surveyed population, respectively. On the other end of the rating scale are multilingual supports (10%), trading simulators, and online financial consultancy (24% each). The latter, in particular, shows a certain growth compared to the previous survey, but is still too weak to be considered representative.

The European sample, on the other hand, shows a somehow different trend. The multilingual support is present in all European countries, with the exception of Great Britain, and the same goes for financial stimulators. To the possibility of asking for advice by mail or to finding an answer to frequent doubts, thanks to the FAQs, European financial institutions prefer to offer financial calculators and asset allocation instruments (though still in embryo).

From the institutional analysis of the Web site, it shows that 94% of banks of the sample deem it indispensable to offer visitors and customers information about the bank. A good 92% of Italian sites describe the structure of the banking institution and its history, while just 79% of foreign banks provide this kind of content. Moreover, listed banks offer the possibility to consult and download the latest balances and relating reports, as well as the most significant press articles. Italian banking institutions seem to give this information area a higher value rather than their foreign counterpart, as well as with regards to the array of news available on the site. All the analyzed Web sites contain their product catalogue. In this regard, there is a difference in price listing: while 83% of Italian Web sites make reference to prices, only 21% of European Web sites make this information available to visitors. Generally speaking, however, the information available concerning this issue is poor and incomplete with regards to contents and presentation clarity.

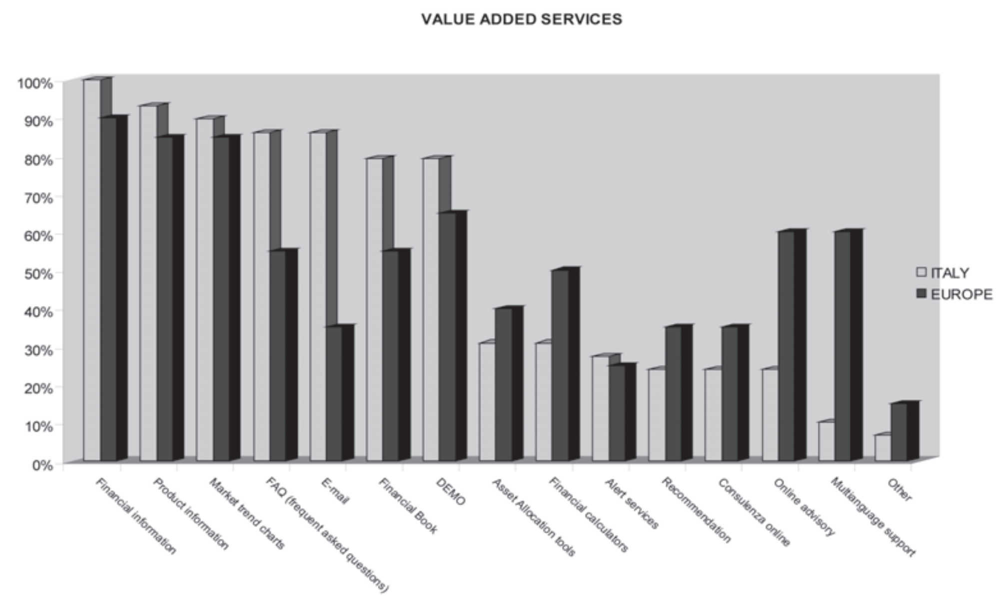
Access speed for the display of the Internet banking service and of online trading has considerably increased (8.75 seconds on average). Italian sites have a higher uploading speed compared to European ones, thanks to the reduced presence of simulators and other value-added services and to the improvement of pages design. In order to optimize the page opening time, HP got equipped with direct service links and improvements were made in the site design. Italian banks got higher scores in the design of Web pages (on average 4.1 against a mean European score of 3.73). The number of images per page has reduced in order to increase browsing speed; legible characters are used, as well as dark tones for the background and light tones for writing. Well aware that the browsing difficulty constitutes a discouraging factor for the average net surfer, banks are paying particular attention to this aspect of their site.

INTRANET BANKING AS AN ORGANIZATIONAL INTEGRATION TOOL

As we were saying, the study on the effects of the use of the Web technologies in the financial and banking industry cannot overlook the intercompany application scope. Banks are traditionally considered as complex, large-size organizations, characterized by geographical widespread location and organizational and product differentiation. Due to the frequent merge and acquisitions transactions (M&As), the organizational complexity has lately grown. As a matter of fact, the search for economies of scale in a minimally differentiated market characterized by end customers being highly sensitive as far as the "price" variable is concerned, pushed the bank top management to increase the volume of products and services offered and to be present in a wide range of markets. The trend is therefore to look for economies through merge and integration projects whose main objective is to increase the company size and to extend the bank geographical boundaries even in international contexts. The search for these situations of strong boost towards volumes has an immediate effect on the competition of the industry, which becomes, on the one hand, increasingly global and, on the other, increasingly based on efficiency and innovation parameters.

Against a rather widespread strategic choice of dimensional growth within the banking industry, the distinguishing factor, on the one hand, and challenging factor, on the other, is represented by the integration pace issue, that is, the time needed

Figure 3. Value-added services offered by online banks



to carry out strategic plans. Therefore, the success factor consists of the ability to create the conditions so that organizational systems and mechanisms change and reach a sound configuration in the shortest time possible.

Within this scenario, a role of paramount importance is played by technologies and information systems in triggering changes through the specification of common practices and procedures in the operations and information management. Lorence and Lorsch (1967) had already pointed out that the need for integration originates from organizational differentiation.

Increasing abilities aimed at managing organizational interdependencies are considered one of the most evident effects of the spread of information technology (Agliati, 1996). In this context the importance of an extensive application of different ICT solutions is stressed, aimed at facing both operational interdependence (i.e., linked to intermediate and final input and output flows which are created as an effect of division of labour) and information interdependence (i.e., referring to the need to exchange information in order to face and solve problems encountered during task performance) (Ferraro, 2000).

A common aspect refers to the progressive increase of internal communication flows channeled through *groupware* applications and, in case of larger organizations, the presence of dedicated Intranet and company portals. The networks (namely *Intranet*) increase the capability of collecting, processing, and redistributing "codified" knowledge. The same infrastructures allow users to exchange messages and different texts, forward requests, consult specialized document files, benefit from remote training and coaching courses, and so forth. The use of such instruments for communication purposes is led by the growing use of solutions and services coming from third parties (i.e., external suppliers or companies referring to the same banking group).

The introduction of the Internet paradigm for the communication and exchange management is not so strong; as an example, the use of *Extranet* in the relationships with external partners is still marginal. Sometimes the relationship with certain *outsourcers* (with the exclusion only of persons in charge of the management information system) uses the support offered by similar tools.

Entrusting the operational phases to outside bodies, with the consequent need to carry out systematic controls on the work performed by the supplier and on the creation of working *teams* of specialists coming from the extra-banking world, may be the reason behind the development, in the future, of computerized applications addressed to external counterparts both at the top and at the bottom of the company organizational limits. Should this trend gain ground, we may witness a further expansion of the limits of information systems, with clear repercussions on the bodies in charge of the operation and management of the said systems.

Numerous banks have realized how *Intranets* may increase the visibility of the organization, since they make clear the existence of knowledge, *routine*, and interconnections between all parties contributing to the same process. In certain fields the analysis perspective is widened and this ends up by making the same organization function *transparently* towards the different players.

INTRANET APPLICATIONS: THE CASE OF CONTENT MANAGEMENT AND OF BUSINESS PROCESS MANAGEMENT

Information systems, in their totality, are essential tools aimed at achieving coordination. Proof of this are the investments that financial institutions make in order to adopt innovative technologies alongside the most traditional corporate information systems.

A survey carried out by CeTIF[1] on a sample of 9 banking groups and 2 service centers, ended in March 2005, showed the existence of significant innovative projects in the fields of content management (CM) and business process management (BPM). By the first term we refer to a set of solutions that automate, partially or fully, the process of creation and publication of contents (*CMt*) and of documents (*document management*) for users inside the bank (Intranet, internal portals) and outside the bank (Extranet, Web etc.). By the second term we refer to the set of *workflow* technologies which manage and monitor the business process, by integrating it with traditional applications.

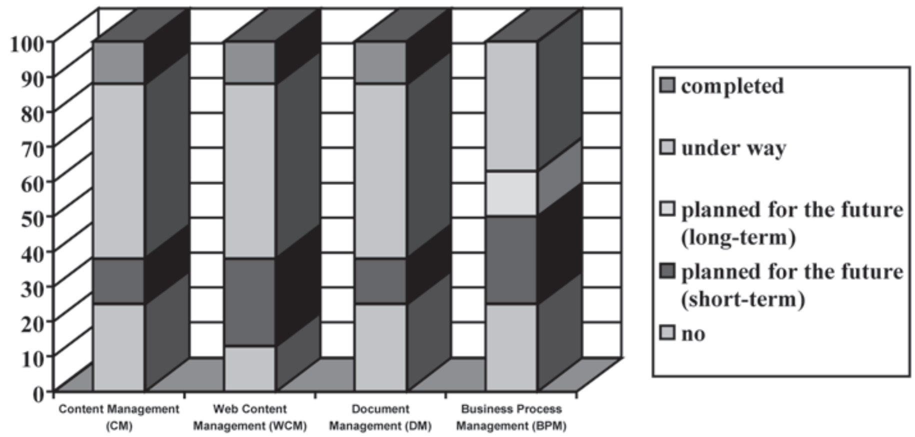
Both such technologies may be considered innovative business integration tools, though their role may be considered of coordination support and of automation (or, to a certain extent, of decision support), respectively.

However, their choice must be explained in the light of the organizational change that major credit institutions have started in the past few years. Many of the cases considered show that the implementation of CM or BPM solutions was followed up by information system integrations or migrations (48%), merge or acquisitions (23%), and, in the case of BPM, after process revision projects (25%). Though the implementation of both projects derived from common organizational needs, the data clearly show a different trend of the first ones compared to the second ones (Figure 4).

Sixty-fiver percent of interviewed banks declared that they are presently carrying out, or they will soon carry out, CM projects, while only 54% of the sample declared that they are presently carrying out or they are planning to carry out in the short term BPM projects. About 10% of banks have planned investments in BPM in the long term. Therefore it becomes evident that CM projects are more widespread than BPMs among financial institutions, which means that CM support technology is more widespread than BPM support technology.

When considering the volume of the projects, we realize that if, on the one hand, CM may have reached a good circulation because it is applied quite frequently to performance information and product/market information processes (pic. 3), on

Figure 4. CM and BPM projects in financial institutions



the other hand, BPM is spread only in the field of credits and information processes (38%), finance and treasury (25%), and auditing and risks (23%). In certain cases (i.e., purchase invoice cycle, information processes, collectionsm and payments), the BPM extends also outside the boundaries of the credit institution, involving also customers or, often, suppliers.

This Difference May be Due to Different Factors

First of all, the factors may be the causes and the relating expected results which boosted the innovation introduction in both technological applications. The survey shows that both CM and BPM projects are mostly induced by exogenous market forces and by the consequent need to reduce the so called "*time to market*" (71%). If, on the one side, the market pressure and the introduction of the customer-oriented culture get, in theory, banks to look for ways to improve internal processes and communication integration, on the other side, reality shows that the equipment with tools bringing immediate and recognizable advantages, mostly outside the bank itself, was made easier at first. Therefore, CM projects, aiming at facilitating information publication and maintenance to the outside public and, only in a second phase, inside the financial institution, have anticipated the BPM projects. The latter are seen as management excellence solutions within processes (68% of interviewed institutions) and as tools implemented to overcome difficulties due to the need for technical modernization of the traditional information systems (44%). The results expected from the two solutions seem to be very different, with consequent differences in the implementation technical difficulties, in costs, and in the identification of organizational effects. If, on the one hand, CM solutions are applied by the bank alongside the existing ones, without the need of organizational changes and important technological integrations, on the other hand, BPM solutions are based on rather more complex technologies, characterized by organizational uncertainty.

Even the analysis of the obstacles faced by CM and BPM projects may be useful to understand the origin of the different spread of such projects. Aggregated data show a common sensitivity towards the cost factor (about 24% of the sample indicated costs as the main obstacle to such initiatives) and the relating difficulty in defining the return on investment (ROI). Another strong obstacle is due to the cultural and organizational reluctance (78%) to the change brought about by these technological innovations. On the other hand, difficulties relating to technologies (18%) do not seem to be a fundamental inhibitor factor. The empiric evidence show, however, a difference between the obstacles to the CM and BPM projects, with particular reference to the cultural and organizational aspects. The reasons are to be found, first of all, in the characteristics of managers and users of these technological innovations and, secondly, in their implementation modalities.

In the case of CM, 44% of respondents state that the people in charge of these projects are active in the lines of business (LOB), while 32% are active in the human resource management (HR) or in the organization. At present, only 18% of financial institutions have "created" a new professional position (i.e., *knowledge officer*) which centralizes content management in one organizational area; the new trend seems to confirm the growth of this organization solution. BPM projects, instead, are managed by the LOB in 24% of the cases, and by the IT area in 32% of the cases. When looking at the characteristics of users of the above mentioned technologies, important differences arise. CM solutions are more and more spread among customers rather than being centrally used. This entail the support of the training structure whose task is to create professional people able to manage the content publication process, as well as, in certain cases (about 50% of the sample), the relevant monitoring process. Even BPM are spread amongst users that manage the processes which make the object of this study. In particular, with regards to processes, a need of coordination and rationalization of business procedures in the most interdependent areas arose. Bank information systems are made up by parallel and integrated applications, often based on different and complex information structures. However, there is no need to create *ad hoc* professional figures which manage the BPM process and project. As a matter of fact, BPM tools often go hand in hand with or replace transactional information systems which play the main role by completing their integration and automation capacity.

Moreover, CM and BPM projects differentiate for their application modality. While CM is characterized by a *big-bang* implementation logic, which introduces technologies in a sole phase, BPM is often characterized by a modular implementation system.

AN INNOVATION EVOLUTION PERSPECTIVE

It is estimated that the difference in the present status of CM and BPM will fade more and more due to the growing awareness that internal processes are the very same business apparatus; this is due even in the light of the recent regulations on operational risks and also because the two technologies may eventually complement each other.

In particular, the survey shows that the differences between the two technologies may be due, on the one hand, to a lesser knowledge of the BPM technological tool compared to the CM one, and on the other, to the lack, limitedly to the BPM projects, of the domino (or *isomorphism)* effect which is common during the technological innovation application phases and which leads to a resemblance of technological choices by financial institutions.

We hereby would like to hypothesize a hidden link between the two technologies and the presence of an "incremental" effect between them. In order to verify this hypothesis, one must look first of all at the technological and organizational innovation which CM on the one side and BPM on the other bring about.

CM technological innovation may be summarized with the introduction of documental integration tools for archive management and in distributed environments. From a functional point of view, CM allows the automated management of the lifecycle of a document, as well as the need to trace the documents and information contained therein. For this reason, such tools are considered as tools enabling the spread of the internal knowledge of the credit institution.

From a technological point of view, BPM introduces innovation integration between the information systems. Moreover, from a functional point of view, they lead to the identification of process interdependencies and to their management harmonization. Moreover, it is a preparatory tool aimed at the process mapping, which can be used for a range of activities (e.g., risk management, auditing system, definition of internal regulations, etc.).

According to this short synthesis, an evolution perspective for the two technologies seems to take form. The common characteristics may be outlined as follows:

- **Process concept:** Both technologies imply a strong idea of organization through processes. In the case of CM, this merely refers to the document creation and publication processes, while in the case of BPM, this concept is also extended to other business processes.
- **Culture of change:** It shows that the organizational change following the introduction of such technologies is difficult to manage, though it is characterized by different factors. The bank should be able to change and, for this purpose, the management *sponsorship* is important.
- **Organizational and technological monitoring:** The technology innovation seems to be linked to the need of monitoring its implementation from both the technological and organizational point of view.
- **Technological flexibility:** Technological innovation goes hand in hand with the flexibility concept, seen as a capability to "adapt" to the company structures and systems.

We therefore believe that BPM is an evolution and extension of the CM concept (Figure 5).

CM projects are therefore destined to evolve towards BPM projects extended to other processes, whenever the financial institution is able to recognize the CM as a technological innovation tool.

SOME USEFUL INSIGHTS

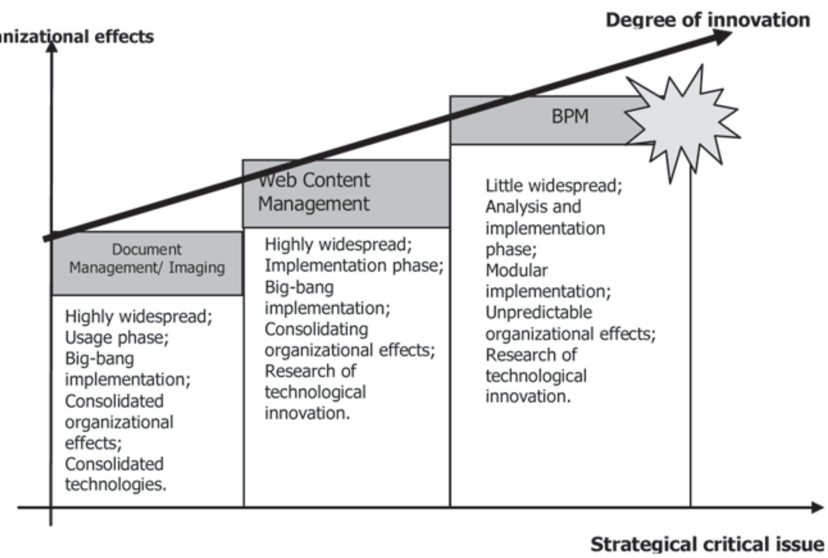
The considerations made on Web-based technologies, both Internet and Intranet, enable one to draw interesting conclusions already outlined at the beginning of this chapter and concerning the innovation characteristics.

Empirical evidence leads us to focus on how the theory may explain the following:

1. Which are the technological innovation signs? Can they always be measured? Does technological innovation progresses through stages or is it an on-going process?
2. Is technological innovation unique? Does it make sense to speak about innovation of each technology sector (automation, decision-making support, coordination)?
3. What is the relationship between technological innovation and organizational efficiency/effectiveness?

Innovation is a synonym of "change or improving evolution of a situation" and surely not a novelty. Some writers consider technological innovation as technology flexibility or as an improvement of performance. Here we would rather exclude such definitions, since the definitions of flexibility and performance may be ambiguous and "contaminated" by other concurrent factors (Maggi, 2003). For this reason, it is therefore believed that technological innovation is not measurable per se. However, the innovation process may be seen as an *on-going process*, more or less fast, characterized by a constant change in technology, players, and structures. This concept

Figure 5. The technological innovation stages



leads us to consider technological innovation as a natural process, only little conscious of the fact that it is stimulated by the implementation of certain technologies which may affect the pace (by increasing it or stopping it) and the process awareness.

Hence, we conclude that the answer to the second question is negative. The division in classes, given at the beginning of this chapter and recognized as valid by the organizational literature, seems not to be relevant in the case of innovation. We could at least "defend" such classification by construing it as typology, that is, as a set of types (rather than classes) which do not have precise demarcation limits and which constitute some archetypes. The concept of technological innovation, however, remains an extensive concept, not referable to any typology.

The third questions leads to wider considerations. Though the concept of organizational efficiency and technological innovation has been highlighted by many interlocutors, the gathered data, however, do not provide empirical evidence. The reason shall be found in the lack of a longitudinal analysis over a longer temporal span, for the determination of effects produced by information technology. However, a qualitative consideration can be drawn: The two technologies showed different efficiency values against the same technology use. This proves, as a first estimate, that information technology does not have a certain and determining role for reaching organizational efficiency, though it may influence both the information need, as well as the information processing capability. Therefore, its effects seem mediated by organizational and contextual variables, leading to different values with regards to effectiveness.

Therefore, the objective of this chapter is to define to what extent information technology may affect the organizational efficiency, that is, the capability to reduce the operating costs of the organizational structure. On this point, the interviewed persons stressed the high importance given to the role of information technology. IT enables the processing of great volumes of data, a deeper control, the cost cutting in terms of communication and transactions. Hence a new question arises: Does any information technology have an impact on transaction costs and on operating costs? The answer seems to be yes, though the impact magnitude on efficiency is different. While automation and communication technologies highly influence the economic management of the institution, decision-making support technologies have an indirect impact, more difficult to asses beforehand.

CONCLUSION

This study dealt with technology innovation, with particular reference to Web technologies. The different empirical evidence shows that technology may be a tool to improve the service/product offered, as well as an efficiency tool. As a matter of fact, the implementation of Internet technologies brought about, on the one hand, the development of a considerable number of Internet strategies for products and services of Italian banks. On the other hand, Internet technologies supported the coordination as well as the organizational integration phases of the same banks.

Process and content innovation mainly implies a change in the cultural attitude and, for the time being, it is still determined mostly by endogenous variables.

The analysis of the collected data leads us to draw some last conclusions:

- Information systems are affected by information technology; in order to determine the information system, one cannot do without knowing the development and implementation status of the technology.
- Information systems are affected by the task uncertainty, by coordination and control mechanisms, and by communication tools; the organizational (micro-) variables affect the definition of the information system.
- The information system is an organizational planning factor; it is able to affect the organizational efficiency and therefore the organizational structure.
- Besides information systems, the organizational structure is affected by other variables; this observation leads us to consider the absence of technologic determinism.
- Information technology has a mediated role in defining the organizational effectiveness and a direct role in the organizational efficiency management.

However, the role played by the said technologies affects the organizational structures in a different way.

REFERENCES

Agliati, M. (Ed.). (1996). *Tecnologia dell'informazione e sistema amministrativo*. Milan: Egea.

Anderson, P., & Tushman, M. L. (1991). Managing through cycles of technological change. *Research Technology Management*, *34*(3), 26–31.

Barras, R. (1990). Interactive innovation in financial and business services: The vanguard of the service revolution. *Research Policy*, *19*, 215–237. doi:10.1016/0048-7333(90)90037-7

Berger, A. N., & Udell, G. F. (2006). A more complete conceptual framework for SME finance. *Journal of Banking & Finance*, *30*(11), 2945–2966. doi:10.1016/j.jbankfin.2006.05.008

Buckingham, R., Hirschheim, R. A., Land, F. F., & Tully, C. J. (1987). Information systems curriculum. A basis for course design. In Buckingham et al. (Ed.), *Information systems education: Recommendations and implementation* (pp. 72-113). Cambridge: CUP.

Carignani, A., & Frigerio, C. (2003). *European Web banking strategies*. Università Cattolica, Milan, CeTIF Research Centre.

Ciborra, C. U. (1996). Improvisation and information technology in organizations. In *Proceedings of the 17th International Conference on Information Systems*, Cleveland, OH (pp. 369-380).

Damanpour, F., & Gopalakrishnan, S. (2001). The dynamics of the adoption of product and process innovation in organizations. *Journal of Management Studies*, *38*(1), 45–65. doi:10.1111/1467-6486.00227

De Nicolò, G., Bartholomew, P., Zaman, J., & Zephirin, M. (2004). Bank consolidation, internationalization, and conglomeration: Trends and implications for financial risk. *Financial Markets, Institutions, & . Instruments*, *13*(4), 173–217.

Epstein, M. J. (2005). The determinants and evaluation of merger success. *Business Horizons, Elsevier, 48*, 37–46. doi:10.1016/j.bushor.2004.10.001

Ferraro, F. (2000). L'analisi organizzativa: l'individuo. In R. Mercurio & F. Testa (Eds.), *Organizzazione, assetto e relazioni nel sistema di business* (pp. 52-76). Torino: Giappichelli.

Figueira, C., Neills, J., & Schoenberg, R. (2007). Travel abroad or stay at home? Investigating the patterns of bank industry M&As in the EU. *European Business Review*, *19*(1), 23–39. doi:10.1108/09555340710714135

Frigerio, C. (2005). *Process IT Innovation in banking*. Università Cattolica, Milan, CeTIF Research Centre.

Lawrence, P. R., & Lorsch, J. W. (1967). *Organization and environment, graduate school of business administration*. Boston: Harvard University.

Maggi, B. (2003). *De l'Agir Organisationnel*. Toulose: Editions Octares.

Quinn, J. B., Anderson, P., & Finkelstein, S. (2005). Leveraging intellect. *The Academy of Management Executive*, *19*(4), 78–94.

Sherman, H. D., & Rupert, T. J. (2006). Do bank mergers have hidden or foregone value? Realized and unrealized operating synergies in one bank merger. *European Journal of Operational Research*, *168*(1), 253–268. doi:10.1016/j.ejor.2004.05.002

Utterback, J., & Abernathy, W. (1975). A dynamic model of product and process innovation. *Omega*, *3*(6), 639–656. doi:10.1016/0305-0483(75)90068-7

ENDNOTE

1. The research mentioned was carried out by the author between March 2004 and October 2004 through a quali-quantitative analysis, in the form of semi-structured questionnaires and interviews (Frigerio, 2005).

This work was previously published in Business Web Strategy: Design, Alignment, and Application, edited by L. Al-Hakim; M. Memmola, pp. 219-234, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 4.8
Semantic Web for Media Convergence:
A Newspaper Case

Ferran Perdrix
Universitat de Lleida, Spain & Diari Segre Media Group, Spain

Juan Manuel Gimeno
Universitat de Lleida, Spain

Rosa Gil
Universitat de Lleida, Spain

Marta Oliva
Universitat de Lleida, Spain

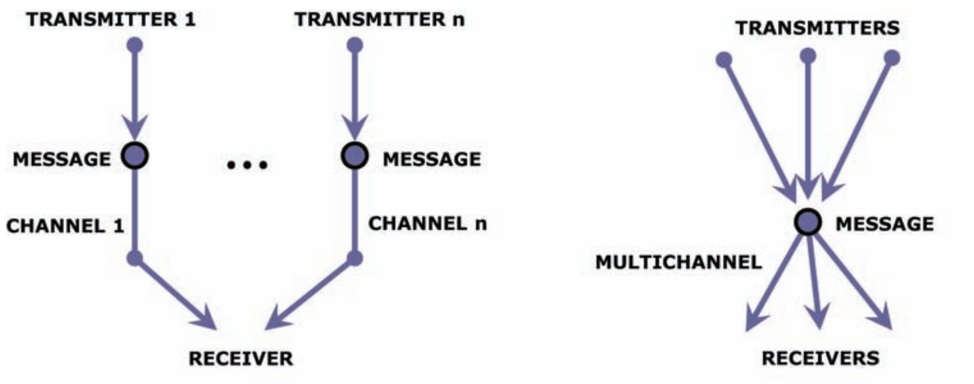
Roberto García
Universitat de Lleida, Spain

ABSTRACT

Newspapers in the digitalisation and Internet era are evolving from mono-channel and static communication mediums to highly dynamic and multi-channel ones, where the different channels converge into a unified news editorial office. Advanced computerised support is needed in order to cope with the requirements arising from convergent multimedia news management, production and delivery. Such advanced services require machines to be aware of a greater part of the underlying semantics. Ontologies are a clear candidate to put this semantics into play, and Semantic Web technologies the best choice for Web-wide information integration. However, newspapers have made great investments in their current news management systems so a smooth transition is required in order to reduce implementation costs. Our proposal is to build an ontological framework based on existing journalism and multimedia standards and to translate existing metadata to the Semantic Web. Once in a semantic space, data integration and news management and retrieval are facilitated enormously. For instance, Semantic Web tools are being developed in the context of a media house that are capable of dealing with the different kinds of media managed in the media house in an integrated and transparent way.

DOI: 10.4018/978-1-60566-066-0.ch009

Figure 1. Traditional news information flux (left) and the new trend of convergent news flux (right)



CURRENT SITUATION

Web news publishing is evolving fast, as the majority of Internet services, and nowadays this service is trying to adapt information to a way that best fits users' interests in order to increase its use. With that, newspapers are expecting to profit more from their news sites. In parallel, many of the newspaper companies are changing into news media houses. They own radio stations and video production companies that produce content unsupported by traditional newspapers, but that is delivered by Web newspapers or new mobile services. Initially, Web news was a mere reproduction of those in the printed edition. Nowadays, they are constantly updated and provide new services for those users interested on reaching this information as soon as possible and enjoying new ways of interaction with them (Eriksen & Ihlström, 2000; Lundberg, 2002; Ihlström, Lundberg, & Perdrix, 2003).

Consequently, news industry communication model is changing from the traditional one shown on the left of Figure 1 to the one shown in the right. In the former, each channel is considered separately (press, TV, radio, Internet, mobile phones…) and implies his way creating his own message, transmitting over this channel and using his own interface in order to show the message to the receivers. On the other hand, the latter is based on an information convergence flux. In this model, transmitters make information in collaboration with other transmitters and produce messages that include as media as it is necessary (video, text, audio, images…). Finally, receivers choose the channel that best fits their needs in order to get access to messages.

The previous situation is the one faced in the context of the Diari Segre Media Group[1], which is a journalism holding that in the last years has been facing this convergence trend. This holding started 25 years ago with a newspaper edition. Today produces three press editions in two languages, three radio stations, six television regional channels and several Internet Websites. Nowadays, all the editorial staff is applying the convergence of information flux approach during news generation and management. Therefore, they are required to be versatile journalists because they cannot be specialized in any concrete media. They must deal with video, image and text edition. Moreover, they must write in different ways, for instance for press news or for radio or TV voiceover.

On the other hand, the Diari Segre archive system is changing to a new repository build from the combination of text, images, video and audio files. In this sense, archive management is becoming a big issue and it requires deep improvements in terms of content search, relations among news (e.g. historical relations among news items) or information retrieval interfaces. The archive system must be a very productive and comprehensive tool in order to assist journalists while they create new content. This business case details how Semantic Web technologies are being explored in the context of the Diari Segre Media Group in order to face this new challenges.

In general, it has been observed that media houses must adapt to the requirements imposed by this new model. First of all, there are changes in how they reach consumers. News are build up from a combination of different content types (video, audio, the traditional text plus images, etc.) and are delivered to users through different channels and adapted to many kinds of devices (PC, PDA, smart phones, etc.). Therefore, formats must be selected and adapted according to the device and connection the user is using. These operations include transcoding of formats, resizing of images or recoding for higher levels of compression. Moreover, multi-channel distribution must take into account that for each channel one must define its own content, aesthetic and interaction model. These characteristics define what an interactive channel is (McDonald, 2004).

However, changes are not just restricted to the relation with consumers. Digital media eliminates many time and space restrictions and changes editorial team routines. Moreover, all different media converge into a unified news object that is produced by interdisciplinary teams. Consequently, more efficient and effective means for news management are needed in order to facilitate coordination and production of these multimedia assets.

The news industry is currently using content management solutions for these means, but the additional requirements of a convergent editorial office stress the need for advanced knowledge management and information retrieval. Currently, there are specific standardisation efforts in the journalism domain, together with more generic ones in the multimedia domain, which carry some uniformity to the newspaper content management systems. However, as it is introduced in the following subsections, they just provide data structures and schemas that facilitate systems interoperability. They do not facilitate knowledge management and information retrieval tasks. These tasks are currently carried out mainly by the documentation department, who is in charge of the news archival process using legacy tools.

Journalism Metadata

One of the main standardization frameworks in the journalism domain is the International Press Telecommunications Council[2], an international consortium of news agencies, editors and newspapers distributors. IPTC has developed standards like the Information Interchange Model[3], NewsCodes[4] (formerly the Subject Reference System), the News Industry Text Format[5] or NewsML[6].

Currently, almost all of them have evolved towards XML-based standards to represent and manage news along their whole lifecycle, including their creation, exchange and consumption. For instance, NewsML is used to represent news as multimedia packages and NITF deals with document structure, i.e. paragraphs, headlines, etc. On the other hand, the Subject Reference System (SRS), now part of IPTC NewsCodes, is a subject classification hierarchy with three levels and seventeen categories in its first level.

Moreover, a new family of these journalism standards has been just proposed, expanding the range of available metadata. The new suite, known as the IPTC G2, is actually a series of specifications and XML components that can be shared among all IPTC G2 components for maximum efficiency.

IPTC G2 Standards make possible the integration of any news item with text, photos, graphics, video or other media. The News Architecture model (NAR) is used in order to package any combination of these items. Moreover, it makes stronger use of IPTC's robust metadata taxonomy suite, which is based on NewsCodes, and better interacts with other IPTC-G2 standards.

This standard contains hooks for managing news items, and its flexibility allows news providers to choose whether to support all of the IPTC G2-standards XML tags or a compact subset. It's the cost-effective way of managing news, whether for a Web site, news aggregator, newspaper, radio or television station.

Multimedia Metadata

All the previous initiatives are centred on the journalism specific aspects of a semantic newspaper. However, as has been pointed out, newspapers are evolving towards the digital multimedia domain. Therefore, they stress more and more their multimedia management requirements.

In the multimedia metadata domain, as it is extensively shown in the literature (Hunter, 2003; Einhoff, Casademont, Perdrix, & Noll, 2005), the MPEG-7 (Salembier, & Smith, 2002) standard constitutes the greatest effort for multimedia description. It is divided into four main components: the Description Definition Language (DDL, the basic building blocks for the MPEG-7 metadata language), Audio (the descriptive elements for audio), Visual (those for video) and the Multimedia Description Schemes (MDS, the descriptors for capturing the semantic aspects of multimedia contents, e.g. places, actors, objects, events, etc.).

In addition to MPEG-7, which concentrates on content description, MPEG-21 defines an open framework for multimedia delivery and consumption. This standard must be also considered because it focuses on the content management issues for full delivery and consumption chain, from content creators' applications to end-users' players. The different MPEG-21 parts deal with diverse aspects like Digital Rights Management or Digital Items, the definition of a fundamental content unit for distribution and transaction very useful for convergent media management.

Problem Statement

As has been pointed out in the description of the current situation in many media houses, archivists classify news using a proprietary hierarchical thesaurus while journalists search this information when they need to inform themselves on subjects, histories or events. This search can be performed in extreme situations, e.g., lack of time, or lack of knowledge in relation to the archive system. This is reflected in the way journalists formulate their queries. The gap between archivists' and journalists' mental models implies that more flexible content categorization and search systems are needed. This trend is even bigger when we consider cross-media content production and coordination in order to get multimedia and multichannel news.Therefore, in order to take advantage of the possibilities offered by the digital medium to exploit a newspaper archive, the aspects that can be improved include:

- Keyword search falling short in expressive power
- Weak interrelation between archive items: users may need to combine several indirect queries manually before they can get answers to complex queries
- Lack of a commonly adopted standard representation for sharing archive news across newspapers
- Lack of internal consensus for content description terminology among reporters and archivists
- Lack of involvement of journalist in the archiving process

These shortcomings are difficult to deal with if the existing standards are used as provided. The main standards that have been presented, both in the journalism and multimedia domains, are based on XML and specified by XML Schemas. The more significant case is the MPEG-7 one. It is based on a set of XML Schemas that define 1182 elements, 417 attributes and 377 complex types. NewsML and NITF are also very big standards, they define more than 100 elements, and the NewsCodes hierarchy of subjects defines more than one thousand different subjects.

The complexity of these standards makes it very difficult to manage them. Moreover, the use of XML technologies implies that a great part of the semantics remains implicit. Therefore, each time an application is developed, semantics must be extracted from the standard and re-implemented.

For instance, if we use XQuery in order to retrieve MPEG-7 SegmentType descriptions from an XML database, we must be aware of the hierarchy of segment types and implement an XQuery that has to cover any kind of multimedia segment, i.e. VideoSegmentType, AnalyticClipType, AudiSegmentType, etc.

If the intended interpretation of the segments structure was available for computerised means, semantic queries would benefit from the corresponding formal semantics. Consequently, a semantic query for SegmentType will retrieve all subclasses without requiring additional developing efforts. This is not possible with XML tools because, although XML Schemas capture some semantics of the domain they model, XML tools are based on syntax. The captured semantics remain implicit from the XML processing tools point of view. Therefore, when an XQuery searches for a SegmentType, the XQuery processor has no way to know that there are many other kinds of segment types that can appear in its place, i.e. they are more concrete kinds of segments.

The previous example only illustrates one kind of difficulty derived from the use of just syntax-aware tools. Another example is that the lack of explicit semantics makes MPEG-7 very difficult to extend in an independent way, i.e. third party extensions. The same applies for MPEG-21 or the journalism standards. Moreover, standards from both worlds share many concepts so it would be possible, and easier, to integrate them once their implicit semantics are available from a computer processing point of view.

Proposed Solution

In this chapter, we explore Semantic Web technologies (Berners-Lee, Hendler & Lassila, 2001) as a way to overcome many of the challenges of digital and convergent media houses. The size and complexity of the stored information, and the time limitations for cataloguing, describing and ordering the incoming information, make newspaper archives a relatively disorganised and difficult to manage corpus. In this sense, they share many of the characteristics and problems of the World Wide Web, and therefore the solutions proposed in the Semantic Web vision are pertinent here.

In order to implement more advanced newspaper content management applications, they should be more informed about the content they are managing. They are not just files with some weak interrelations. There is a lot of knowledge embedded in these pieces of content and in their interrelationships. In order to make computers aware of it, their implicit semantics must be formalised, for instance using ontologies. Semantic Web technologies facilitate the building blocks for Web ontologies, which add the facilities for Web-wide ontology sharing and integration. The latter is a key feature for convergent and globalised media houses.

In order to build an ontological infrastructure for the Semantic Newspaper, it is important to consider the state of the art of the metadata initiatives in the journalism domain, which have been introduced in the current situation description section. Additionally, digital newspapers have stressed the requirements of multimedia management. Digital news is managed as multimedia packages that integrate text, images, video, audio, etc. Therefore, it is also important to consider the current situation in the more general multimedia metadata domain.

We have undertaken the application of the Semantic Web proposals to the newspapers world by following a smooth transition strategy (Haustein, & Pleumann, 2002). This strategy advises about keeping compatibility, at least initially, with current newspaper content management systems and journalism and multimedia standards. Consequently, we have rooted our proposed approach on existing journalism and multimedia standards and provide a methodology to move them, together with existing data, to the Semantic Web domain.

Objectives

The objective is then to design a Semantic Web-based platform that is an extension of previously working systems in mass media companies, particularly in the context of the Diari Segre Media Group. The manual creation of semantic instances for news items, at a regular daily pace, is indeed a feasible goal as long as this process is integrated into existing systems and it just causes a slightly greater work load while producing observable benefits. Consequently, the introduction of new semantic documentation tools requires a careful work of analysis, design, testing and balancing of the additional burden that such tools may impose on archivists, journalists or readers.

In order to produce a semantic platform that seamlessly integrates into newspapers content management systems, the first objective is to develop an ontological framework based on existing standards. Once this ontological infrastructure based on existing journalism and multimedia standards is developed, the objective is then to put it into practice in the context of an architecture based on Semantic Web tools for semantic integration, querying and reasoning. However, all this effort must end up reaching users through applications that offer to them the extra benefits of semantic metadata while avoiding them the burden of dealing with the underlying extra complexity.

Overview

The proposed solution is detailed in Section 3. First of all, Section 3.1 presents the methodology that produces an ontological framework based on existing standards. This methodology is based on two mappings. The first one from XML Schema, the language used in most of the considered standards, to ontologies based on the Semantic Web language Web Ontology Language (OWL) (McGuinness & Harmelen, 2004). The second one is based on the previous one and makes it possible to map from XML metadata, based on XML Schemas previously mapped to OWL, to Semantic Web metadata, based on the Resource Description Framework (RDF) (Becket, 2004).

The ontologies produced using this methodology constitute the foundation on top of which an architecture based on Semantic Web technologies is built. This architecture, described in Section 3.2, takes profit from the semantics formalised by these ontologies and loads Semantic Web metadata based on them in order to offer services like semantic integration, semantics queries or logic reasoning. These services are used in order to build applications that facilitate managing heterogeneous media repositories and the underlying knowledge. One example of such an application is given in Section 3.3.

The described application builds on top of a text-to-speech and a semantic annotation tool. The generated annotations are based on existing standards ontologies and loaded into the proposed semantic architecture, which makes it possible to manage audio, audiovisual and text content in an integrated way. However, the key point here is to offer all the semantic services to users in a usable and accessible way. To this end, the application is based on a user interface that provides an object-action interaction paradigm best suited for heterogeneous information spaces. The interface does not solely facilitate content management, it also allows browsing the underlying domain knowledge, formalised using specialised ontologies, and constitutes a useful tool in media houses in order to facilitate news tracking and producing new content.

SOLUTION DETAILS

This section provides a detailed description of the proposed solution. The different modules are described in the following subsections starting from the methodology used in order to benefit from existing standards and produce ontologies that formalise them. These ontologies make possible to develop an architecture that takes profit from their semantics in order to offer advanced services like semantic integration, querying and reasoning. Finally, these services are used in order to build an application that makes the benefits emerging from semantic metadata and ontologies available for end users.

XML SEMANTICS REUSE METHODOLOGY

In order to put into practice the smooth transition strategy, the first step has been to reuse existing standards in the journalism and multimedia fields, which have been for long very active in standardization.

However, as has been highlighted in current situation analysis, all the more recent standards are based on XML and lack formal semantics that facilitate applying a Semantic Web approach. Therefore, in order to facilitate the transition from current standards and applications to the semantic world, we have applied the XML Semantics Reuse methodology (García, 2006).

The main caveat of semantic multimedia metadata is that it is sparse and expensive to produce. If we want to increase the availability of semantic multimedia metadata and, in general, of semantic metadata, the more direct solution is to benefit from the great amount of metadata that has been already produced using XML, which is extensively used by many newspaper content management systems.

There are many attempts to move metadata from the XML domain to the Semantic Web. Some of them just model the XML tree using the RDF primitives (Klein, 2002). Others concentrate on modelling the knowledge implicit in XML languages definitions, i.e. DTDs or the XML Schemas, using Web ontology languages (Amann, Beer, Fundulak, & Scholl, 2002; Cruz, Xiao, & Hsu, 2004). Finally, there are attempts to encode XML semantics integrating RDF into XML documents (Lakshmanan, & Sadri, 2003; Patel-Schneider, & Simeon, 2002).

However, none of them facilitates an extensive transfer of XML metadata to the Semantic Web in a general and transparent way. Their main problem is that the XML Schema implicit semantics are not made explicit when XML metadata instantiating this schemas is mapped. Therefore, they do not benefit from the XML semantics and produce RDF metadata almost as semantics-blind as the original XML. Or, on the other hand, they capture these semantics but they use additional ad-hoc semantic constructs that produce less transparent metadata.

Therefore, we propose the XML Semantics Reuse methodology, which is implemented by the ReDeFer project[7] as an XML Schema to OWL plus and XML to RDF mapping tool. This methodology combines an XML Schema to Web ontology mapping, called XSD2OWL, with a transparent mapping from XML to RDF, XML2RDF. The ontologies generated by XSD2OWL are used during the XML to RDF step in order to generate semantic metadata that makes XML Schema semantics explicit. Both steps are detailed next.

XSD2OWL Mapping

The XML Schema to OWL mapping is responsible for capturing the schema implicit semantics. This semantics are determined by the combination of

Table 1. XSD2OWL mappings from XML Schema building blocks to OWL ones plus an explanation of why they are interpreted as equivalent modelling constructs

XML Schema	OWL	Explanation
element | attribute	rdf:Property owl:DatatypeProperty owl:ObjectProperty	Named relation between nodes or nodes and values
element@substitutionGroup	rdfs:subPropertyOf	Relation can appear in place of a more general one
element@type	rdfs:range	The relation range kind
complexType|group |attributeGroup	owl:Class	Relations and contextual restrictions package
complexType//element	owl:Restriction	Contextualised restriction of a relation
extension@base restriction@base	rdfs:subClassOf	Package concretises the base package
@maxOccurs @minOccurs	owl:maxCardinality owl:minCardinality	Restrict the number of occurrences of a relation
Sequence choice	owl:intersectionOf owl:unionOf	Combination of relations in a context

XML Schema constructs. The mapping is based on translating these constructs to the OWL ones that best capture their semantics. These translations are detailed in Table 1.

The XSD2OWL mapping is quite transparent and captures a great part XML Schema semantics. The same names used for XML constructs are used for OWL ones, although in the new namespace defined for the ontology. Therefore, XSD2OWL produces OWL ontologies that make explicit the semantics of the corresponding XML Schemas. The only caveats are the implicit order conveyed by xsd:sequence and the exclusivity of *xsd:choice*.

For the first problem, *owl:intersectionOf* does not retain its operands order, there is no clear solution that retains the great level of transparency that has been achieved. The use of RDF Lists might impose order but introduces ad-hoc constructs not present in the original metadata. Moreover, as has been demonstrated in practise, the element ordering does not contribute much from a semantic point of view. For the second problem, owl:unionOf is an inclusive union, the solution is to use the disjointness OWL construct, *owl:disjointWith*, between all union operands in order to make it exclusive.

To conclude, one important aspect is that the resulting OWL ontology may be OWL-Full depending on the input XML Schema. This is due to the fact that, in some cases, the XSD2OWL translator must employ *rdf:Property* for those xsd:elements that have both data type and object type ranges.

XML2RDF Mapping

Once all the metadata XML Schemas are available as mapped OWL ontologies, it is time to map the XML metadata that instantiates them. The intention is to produce RDF metadata as transparently as possible. Therefore, a structure-mapping approach has been selected (Klein, 2002). It is also possible to take a model-mapping approach (Tous, García, Rodríguez, & Delgado, 2005).

XML model-mapping is based on representing the XML information set using semantic tools. This approach is better when XML metadata is semantically exploited for concrete purposes. However, when the objective is semantic metadata that can be easily integrated, it is better to take a more transparent approach.

Transparency is achieved in structure-mapping models because they only try to represent the XML metadata structure, i.e. a tree, using RDF. The RDF model is based on the graph so it is easy to model a tree using it. Moreover, we do not need to worry about the semantics loose produced by structure-mapping. We have formalised the underlying semantics into the corresponding ontologies and we will attach them to RDF metadata using the instantiation relation *rdf:type*.

The structure-mapping is based on translating XML metadata instances to RDF ones that instantiate the corresponding constructs in OWL. The more basic translation is between relation instances, from *xsd:elements* and *xsd:attributes* to *rdf:Properties*. Concretely, *owl:ObjectProperties* for node to node relations and *owl:DatatypeProperties* for node to values relations.

However, in some cases, it would be necessary to use *rdf:Properties* for *xsd:elements* that have both data type and object type values. Values are kept during the translation as simple types and RDF blank nodes are introduced in the RDF model in order to serve as source and destination for properties. They will remain blank for the moment until they are enriched with semantic information.

The resulting RDF graph model contains all that we can obtain from the XML tree. It is already semantically enriched due to the *rdf:type* relation that connects each RDF properties to the *owl:ObjectProperty* or *owl:DatatypeProperty* it instantiates. It can be enriched further if the blank nodes are related to the *owl:Class* that defines the package of properties and associated restrictions they contain, i.e. the corresponding *xsd:complexType*. This semantic decoration of the graph is formalised using *rdf:type* relations from blank nodes to the corresponding OWL classes.

At this point we have obtained a semantics-enabled representation of the input metadata. The instantiation relations can now be used to apply OWL semantics to metadata. Therefore, the semantics derived from further enrichments of the ontologies, e.g. integration links between different ontologies or semantic rules, are automatically propagated to instance metadata due to inference.

However, before continuing to the next section, it is important to point out that these mappings have been validated in different ways. First, we have used OWL validators in order to check the resulting ontologies, not just the MPEG-7 Ontology but also many others (García, Gil, & Delgado, 2007; García, Gil, Gallego, & Delgado, 2005). Second, our MPEG-7 ontology has been compared to Hunter's (2001) and Tsinaraki's ones (2004).

Both ontologies, Hunter's and Tsinaraki's, provide a partial mapping of MPEG-7 to Web ontologies. The former concentrates on the kinds of content defined by MPEG-7 and the latter on two parts of MPEG-7, the Multimedia Description Schemes (MDS) and the Visual metadata structures. It has been tested that they constitute subsets of the ontology that we propose.

Finally, the XSD2OWL and XML2RDF mappings have been tested in conjunction. Testing XML instances have been mapped to RDF, guided by the corresponding OWL ontologies from the used XML Schemas, and then back to XML. Then, the original and derived XML instances have been compared using their canonical version in order to correct mapping problems.

Ontological Infrastructure

As a result of applying the XML Semantics Reuse methodology, we have obtained a set of ontologies that reuse the semantics of the underlying standards, as they are formalised through the corresponding XML Schemas. All the ontologies related to journalism standards, i.e. NewsCodes NITF and NewsML, are available from the Semantic Newspaper site[8]. This site also contains some of the ontologies for the MPEG-21 useful for news modelling as convergent multimedia units. The MPEG-7 Ontology is available from the MPEG-7 Ontology site[9]. These are the ontologies that are going to be used as the basis for the semantic newspaper info-structure:

- **NewsCodes subjects ontology:** An OWL ontology for the subjects' part of the IPTC NewsCodes. It is a simple taxonomy of subjects but it is implemented with OWL in order to facilitate the integration of the subjects' taxonomy in the global ontological framework.
- **NITF 3.3 ontology:** An OWL ontology that captures the semantics of the XML Schema specification of the NITF standard. It contains some classes and many properties dealing with document structure, i.e. paragraphs, subheadlines, etc., but also some metadata properties about copyright, authorship, issue dates, etc.
- **NewsML 1.2 ontology:** The OWL ontology resulting from mapping the NewsML 1.2 XML Schema. Basically, it includes a set of properties useful to define the news structure as a multimedia package, i.e. news envelope, components, items, etc.
- **MPEG-7 ontology:** The XSD2OWL mapping has been applied to the MPEG-7 XML Schemas producing an ontology that has 2372 classes and 975 properties, which are targeted towards describing multimedia at all detail levels, from content based descriptors to semantic ones.
- **MPEG-21 digital item ontologies:** A digital item (DI) is defined as the fundamental unit for distribution and transaction in MPEG-21.

System Architecture

Based on the previous XML world to Semantic Web domain mappings, we have built up a system architecture that facilitates journalism and multimedia metadata integration and retrieval. The architecture is sketched in Figure 2. The MPEG-7 OWL ontology, generated by XSD2OWL, constitutes the basic ontological framework for semantic multimedia metadata integration and appears at the centre of the architecture. In parallel, there are the journalism ontologies. The multimedia related concepts from the journalism ontologies are connected to the MPEG-7 ontology, which acts as an upper ontology for multimedia. Other ontologies and XML Schemas can also be easily incorporated using the XSD2OWL module.

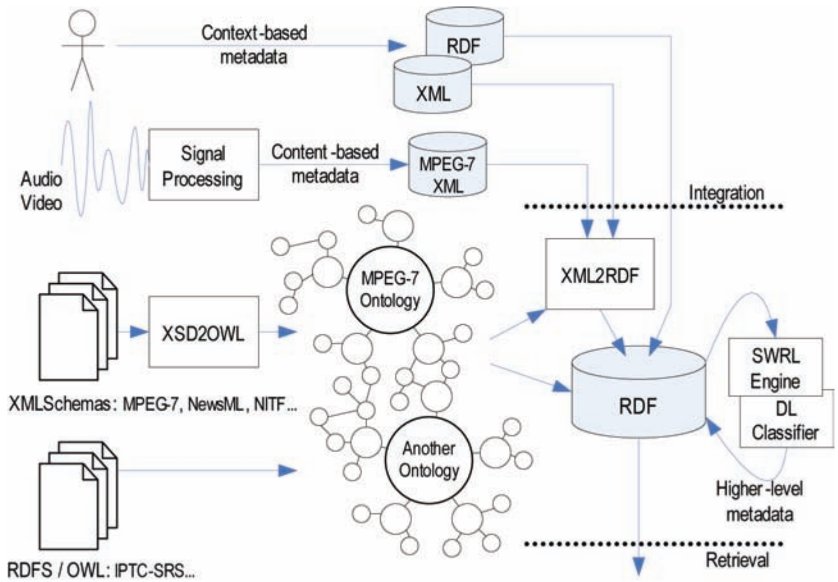
Semantic metadata can be directly fed into the system together with XML metadata, which is made semantic using the XML2RDF module. For instance, XML MPEG-7 metadata has a great importance because it is commonly used for low-level visual and audio content descriptors automatically extracted from its underlying signals. This kind of metadata can be used as the basis for audio and video description and retrieval.

In addition to content-based metadata, there is context-based metadata. This kind of metadata higher level and it usually, in this context, related to journalism metadata. It is generated by the system users (journalist, photographers, cameramen, etc.). For instance, there are issue dates, news subjects, titles, authors, etc.

This kind of metadata can come directly from semantic sources but, usually, it is going to come from legacy XML sources based on the standards' XML Schemas. Therefore, in order to integrate them, they will pass through the XML2RDF component. This component, in conjunction with the ontologies previously mapped from the corresponding XML Schemas, generates the RDF metadata that can be then integrated in the common RDF framework.

This framework has the persistence support of a RDF store, where metadata and ontologies reside. Once all metadata has been put together, the semantic integration can take place, as shown in the next section.

Figure 2. News metadata integration and retrieval architecture



Semantic Integration Outline

As mentioned in the introduction, one of the main problems in nowadays media houses is that of heterogeneous data integration. Even within a single organization, data from disparate sources must be integrated. Our approach to solve this problem is based on Web ontologies and, as the focus is on multimedia and journalism metadata integration, our integration base are the MPEG-7, MPEG-21 and the journalism ontologies.

In order to benefit from the system architecture presented before, when semantic metadata based on different schemes has to be integrated, the XML Schemas are first mapped to OWL. Once this first step has been done, these schemas can be integrated into the ontological framework using OWL semantic relations for equivalence and inclusion: *subClassOf*, *subPropertyOf*, *equivalentClass*, *equivalentProperty*, *sameIndividualAs*, etc. These relations allows simple integration relations, for more complex integration steps that require changes in data structures it is possible to use Semantic Web rules (Horrocks, Patel-Schneider, Boley, Tabet, Grosof, & Dean, 2004).

These relationships capture the semantics of the data integration. Then, once metadata is incorporated into the system and semantically-decorated, the integration is automatically performed by applying inference. Table 2 shows some of these mappings, performed once all metadata has been moved to the semantic space.

First, there are four examples of semantic mappings among the NITF Ontology, the NewsML Ontology and the IPTC Subjects Ontology. The first mapping tells that all values for the *nitf:tobject.subject* property are from class *subj:Subject*. The second one that the property *nitf:tobject.subject.detail* is equivalent to *subj:explanation*. The third one that all *nitf:body* instances are also *newsml:DataContent* instances and the fourth one that all *newsml:Subject* are *subj:Subject*. Finally, there is also a mapping that is performed during the XML to RDF translation. It is necessary in order to recognise an implicit identifier, *nitf:tobject.subject.refnum* is mapped to rdf:ID in order to make this recognise this identifier in the context of NITF and make it explicit in the context of RDF.

Table 2. Journalism and multimedia metadata integration mapping examples

Semantic Mappings
∀ nitf:tobject.subject . subj:Subject
nitf:tobject.subject.detail ≡ subj:explanation
nitf:body ⊆ newsml:DataContent
newsml:Subject ≡ subj:Subject
XML2RDF Mappings
nitf:tobject.subject.refnum → rdf:ID

SEMANTIC MEDIA INTEGRATION FROM HUMAN SPEECH

This section introduces a tool, build on top of the ontological infrastructure described in the previous sections, geared towards a convergent and integrated news management in the context of a media house. As has been previously introduced, the diversification of content in media houses, who must deal in an integrated way with different modalities (text, image, graphics, video, audio, etc.), carries new management challenges. Semantic metadata and ontologies are a key facilitator in order to enable convergent and integrated media management.

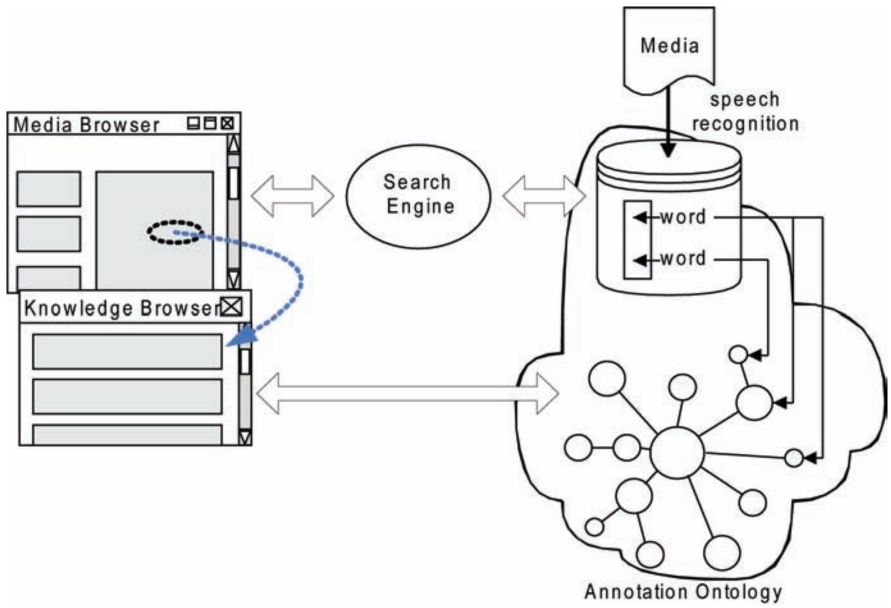
In the news domain, news companies like the Diari Segre Media Group are turning into news media houses, owning radio stations and video production companies that produce content not supported by the print medium, but which can be delivered through Internet newspapers. Such new perspectives in the area of digital content call for a revision of mainstream search and retrieval technologies currently oriented to text and based on keywords. The main limitation of mainstream text IR systems is that their ability to represent meanings is based on counting word occurrences, regardless of the relation between words (Salton, & McGill, 1983). Most research beyond this limitation has remained in the scope of linguistic (Salton, & McGill, 1983) or statistic (Vorhees, 1994) information.

On the other end, IR is addressed in the Semantic Web field from a much more formal perspective (Castells, Fernández, & Vallet, 2007). In the Semantic Web vision, the search space consists of a totally formalized corpus, where all the information units are unambiguously typed, interrelated, and described by logic axioms in domain ontologies. Such tools enabled the development of semantic-based retrieval technologies that support search by meanings rather than keywords, providing users with more powerful retrieval capabilities to find their way through in increasingly massive search spaces.

Semantic Web based news annotation and retrieval has already been applied in the Diari Segre Media Group in the context of the Neptuno research project (Castells, Perdrix, Pulido, Rico, Benjamins, Contreras, & Lorés, 2004). However, this is a partial solution as it just deals with textual content. The objective of the tool described in this section is to show how these techniques can also be applied to content with embedded human-speech tracks. The final result is a tool based on Semantic Web technologies and methodologies that allows managing text and audiovisual content in an integrated and efficient way. Consequently, the integration of human speech processing technologies in the semantic-based approach extends the semantic retrieval capabilities to audio content. The research is being undertaken in the context of the S5T research project[10].

As shown in Figure 3, this tool is based on a human speech recognition process inspired

Figure 3. Architecture for the Semantic Media Integration from Human Speech Tool



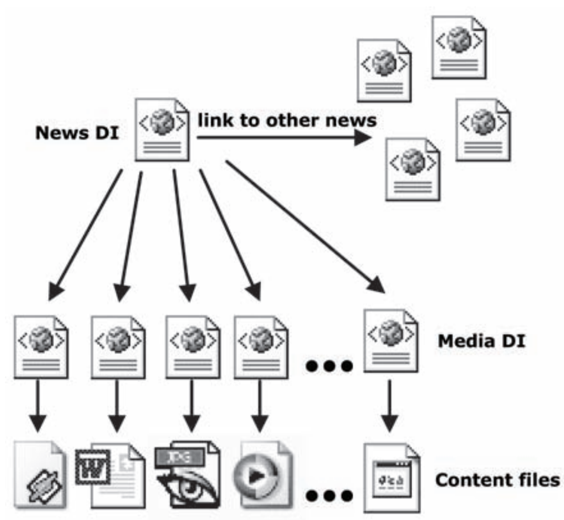
by (Kim, Jung, & Chung, 2004) that generates the corresponding transcripts for the radio and television contents. From this preliminary process, it is possible benefit from the same semi-automatic annotation process in order to generate the semantic annotations for audio, audiovisual and textual content. Keywords detected during speech recognition are mapped to concepts in the ontologies describing the domain covered by audiovisual and textual content, for instance the politics domain for news talking about this subject. Specifically, when the keyword forms of a concept are uttered in a piece of speech, the content is annotated with that concept. Polysemic words and other ambiguities are treated by a set of heuristics. More details about the annotation and semantic query resolution processes are available from (Cuayahuitl, & Serridge, 2002).

Once audio and textual contents have been semantically annotated (Tejedor, García, Fernández, López, Perdrix, Macías, et al., 2007), it is possible to provide a unified set of interfaces, rooted on the semantic capabilities provided by the annotations. These interfaces, intended for journalists and archivist, are shown on the left of Figure 3. They exploit the semantic richness of the underlying ontologies upon which the search system is built. Semantic queries are resolved, using semantic annotations as has been previously described, and retrieve content items and pieces of these contents. News contents are packaged together using annotations based on the MPEG-21 and MPEG-7 ontologies, as it is described in Section 3.3.1. Content items are presented to the user through the Media Browser, detailed in Section 3.3.2, and the underlying semantic annotations and the ontologies used to generate these annotations can be browsed using the Knowledge Browser, described in Section 3.3.3.

Semantic News Packaging Using MPEG Ontologies

Actually, in an editorial office there are a lot of applications producing media in several formats. This is an issue that requires a common structure to facilitate management. The first step is to treat each unit of information, in this case each new,

Figure 4. Content DI structure



as a single object. Consequently, when searching something upon this structure, all related content is retrieved together.

Another interesting issue is that news can be linked to other news. This link between news allows the creation of information threads. A news composition metadata system has been developed using concepts from the MPEG-21 and MPEG-7 ontologies. It comprises three hierarchical levels as shown in Figure 4.

The lower level comprises content files, in whatever format they are. The mid level is formed by metadata descriptors (what, when, where, how, who is involved, author, etc.) for each file, mainly based on concepts from the MPEG-7 ontology generated using the methodology described in Section 3.1. They are called the Media Digital Items (Media DI).

These semantic descriptors are based on the MPEG-7 Ontology and facilitate automated management of the different kinds of content that build up a news item in a convergent media house. For instance, it is possible to generate semantic queries that benefit from the content hierarchy defined in MPEG-7 and formalised in the ontology. This way, it is possible to pose generic queries for any kind of segment (e.g. *AudioSegmentType*, *VideoSegmentType*…) because all of them are formalised as subclasses of *SegmentType* and the implicit semantics can be directly used by a semantic query engine.

Table 3 shows a piece of metadata that describes an audio segment of a Diari Segre Media Group news item used in the S5T project. This semantic metadata is generated from the corresponding XML MPEG-7 metadata using the XML to RDF mapping and takes profit from the MPEG-7 OWL ontology in order to make the MPEG-7 semantics explicit. Therefore, this kind of metadata can be processed using semantic queries independently from the concrete type of segment. Consequently, it is possible to develop applications that process in an integrated and convergent way the different kinds of contents that build up a new.

The top level in the hierarchy is based on descriptors that model news and put together all the different pieces of content that conform them. These objects are called News Digital Items (News DI). There is one News DI for each news item and all of them are based on MPEG-21 metadata. The part of the standard that defines digital items (DI) is used for that. DI is the fundamental unit defined in MPEG-21 for content distribution and transaction, very useful for convergent

Table 3. MPEG-7 Ontology description for a audio segment generated from XML MPEG-7 metadata fragment

```
<?xml version=”1.0”?>
<rdf:RDF
xmlns:mpeg7=”http://rhizomik.net/ontologies/2006/03/Mpeg7-2001.owl#”>
<mpeg7:AudioType rdf:about=”http://rhizomik.net/audio/2007-01-13.mp3”>
<mpeg7:Audio>
<mpeg7:AudioSegmentType>
<mpeg7:MediaTime>
<mpeg7:MediaTimeType>
<mpeg7:MediaTimePoint
rdf:datatype=”&xsd;time”>01:27.0</mpeg7:MediaTimePoint>
<mpeg7:MediaDuration
rdf:datatype=”&xsd;time”>P5S</mpeg7:MediaDuration>
</mpeg7:MediaTimeType>
</mpeg7:MediaTime>
</mpeg7:AudioSegmentType>
</mpeg7:Audio>
</mpeg7:AudioType>
</rdf:RDF>
```

media management. As in the case of MPEG-7 metadata, RDF semantic metadata is generated from XML using the semantics made explicit by the MPEG-21 ontologies. This way, it is possible to implement generic processes also at the news level using semantic queries.

On top of the previous semantic descriptors at the media and news level, it is possible to develop an application for integrated and convergent news management in the media house. The application is based on two specialised interfaces described in the next subsections. They benefit from the ontological infrastructure detailed in this chapter, which is complemented with ontologies for the concrete news domain. However, the application remains independent from the concrete domain.

Media Browser

The Media Browser, shown in Figure 5, takes profit from the MPEG-21 metadata for news and MPEG-7 metadata for media in order to implement a generic browser for the different kinds of media that constitute a news item in a convergent newspaper. This interface allows navigating them and presents the retrieved pieces of content and the available RDF metadata describing them. These descriptions are based on a generic rendering of RDF data as interactive HTML for increased usability (García, & Gil, 2006).

The multimedia metadata is based on the Dublin Core schema for editorial metadata and IPTC News Codes for subjects. For content-based metadata, especially the content decomposition depending on the audio transcript, MPEG-7 metadata is used for media segmentation, as it was shown in Table 3. In addition to the editorial metadata and the segments decomposition, a specialized audiovisual view is presented. This view allows rendering the content, i.e. audio and video, and interacting with audiovisual content through a click-able version of the audio transcript.

Two kinds of interactions are possible from the transcript. First, it is possible to click any word in the transcript that has been indexed in order to perform a keyword-based query for all content in the database where that keyword appears. Second, the transcript is enriched with links to the ontology used for semantic annotation. Each word in the transcript whose meaning is represented by an ontology concept is linked to a description of that concept, which is shown by the Knowledge Browser detailed in the next section. The whole interaction is performed through the user Web

Figure 5. Media Browser interface presenting content metadata (left) and the annotated transcript (right)

browser using AJAX in order to improve the interactive capabilities of the interface.

For instance, the transcript includes the name of a politician that has been indexed and modelled in the ontology. Consequently, it can be clicked in order to get all the multimedia content where the name appears or, alternatively, to browse all the knowledge about that politician encoded in the corresponding domain ontology.

Knowledge Browser

This interface is used to allow the user browsing the knowledge structures employed to annotate content, i.e. the underlying ontologies. The same RDF data to interactive HTML rendering used in the Media Browser is used here. Consequently, following the politician example in the previous section, when the user looks for the available knowledge about that person and interactive view of the RDF data modelling him is shown. This way, the user can benefit from the modelling effort and, for instance, be aware of the politician party, that he is a member of the parliament, etc.

This interface constitutes a knowledge browser so the link to the politician party or the parliament can be followed and additional knowledge can be retrieved, for instance a list of all the members of the parliament. In addition to this recursive navigation of all the domain knowledge, at any browsing step, it is also possible to get all the multimedia content annotated using the concept currently being browsed. This step would carry the user back to the Media Browser.

Thanks to this dual browsing experience, the user can navigate through audiovisual content using the Media Browser and through the underlying semantic models using the Knowledge Browser in a complementary an inter-weaved way. Finally, as for the Media Browser, the Knowledge Browser is also implemented using AJAX so the whole interactive experience can be enjoyed using a Web browser.

ALTERNATIVES

There are other existing initiatives that try to move journalism and multimedia metadata to the Semantic Web world. In the journalism field, the Neptuno (Castells, Perdrix, Pulido, Rico, Benjamins, Contreras, et al., 2004) and NEWS (Fernández, Blázquez, Fisteus, Sánchez, Sintek, Bernardi, et al., 2006) projects can be highlighted.

Both projects have developed ontologies based on existing standards (IPTC SRS, NITF or NewsML) but from an ad-hoc and limited point of view. Therefore, in order to smooth the transition from the previous legacy systems, more complex and complete mappings should be developed and maintained.

The same can be said for the existing attempts to produce semantic multimedia metadata. Chronologically, the first attempts to make MPEG-7 metadata semantics explicit where carried out, during the MPEG-7 standardisation process, by Jane Hunter (2001). The proposal used RDF to formalise a small part of MPEG-7, and later incorporated some DAML+OIL construct to further detail their semantics (Hunter, 2001). More recent approaches (Hausenblas, 2007) are based on the Web Ontology Language (McGuinness & Harmelen, 2004), but are also constrained to a part of the whole MPEG-7 standard, the Multimedia Description Scheme (MDS) for the ontology proposed at (Tsinaraki, Polydoros, & Christodoulakis, 2004).

An alternative to standards-based metadata are folksonomies (Vanderwal, 2007). Mainly used in social bookmarking software (e.g. del.icio.us, Flickr, YouTube), they allow the easy creation of user driven vocabularies in order to annotate resources. The main advantage of folksonomies is the low entry barrier: all terms are acceptable as metada, so no knowledge of the established standards is needed. Its main drawback is the lack of control over the vocabulary used to annotate resources, so resource combination and reasoning becomes almost impossible. Some systems combine social and semantic metadata and try to infer a formal ontology from the tags used in the folksonomy (Herzog, Luger & Herzog, 2007). In our case we believe that it is better to use standard ontologies both from multimedia and journalism fields than open and uncontrolled vocabularies.

Moreover, none of the proposed ontologies, for journalism of multimedia metadata, is accompanied by a methodology that allows mapping existing XML metadata based on the corresponding standards to semantic metadata. Consequently, it is difficult to put them into practice as there is a lack of metadata to play with. On the other hand, there is a great amount of existing XML metadata and a lot of tools based on XML technologies. For example, the new Milenium Quay[11] cross-media archive system from PROTEC, the worldwide leadership in cross-media software platforms, is XML-based. This software is focused on flexibility using several XML tags and mappings, increasing interoperability with other archiving systems. The XML-based products are clearly a trend in this scope. Every day, new products from the main software companies are appearing, which deal with different steps in all the news life-cycle, from production to consumption.

Nowadays, commercial tools based on XML technologies constitute the clear option in newspaper media houses. Current initiatives based on Semantic Web tools are constrained due to the lack of “real” data to work with; they constitute a too abrupt breaking from legacy systems. Moreover, they are prototypes with little functionality. Consequently, we do not see the semantic tools as an alternative to legacy systems, at least in the short term. On the contrary, we think that they constitute additional modules that can help dealing with the extra requirements derived from media heterogeneity, multichannel distribution and knowledge management issues.

The proposed methodology facilitates the production of semantic metadata from existing legacy systems, although it is simple metadata as the source is XML metadata that is not intended for carrying complex semantics. In any case, it constitutes a first and smooth step toward adding semantic-enabled tools to existing newspaper content management systems. From this point, more complex semantics and processing can be added without breaking continuity with the investments that media houses have done in their current systems.

COST AND BENEFITS

One of the biggest challenges in media houses is to attach metadata to all the generated content in order to facilitate management. However, this is easier in this context as in many media houses there is a department specialized in this work, which is carried out by archivists. Consequently, the additional costs arising from the application of Semantic Web technologies are mitigated due to the existence of this department. It is already in charge of indexation, categorization and content semantic enrichment.

Consequently, though there are many organizational and philosophy changes that modify how this task is currently carried out, it is not necessary to add new resources to perform this effort. The volume of information is another important aspect to consider. All Semantic Web approaches in this field propose an automatic or semi-automatic annotation processes.

The degree of automation attained using Semantic Web tools allows archivists spending less time in the more time consuming and mechanical tasks, e.g. the annotation of audio contents which can be performed with the help of speech-to-text tools as in the S5T project example presented in Section 3.3. Consequently, archivists can spend their time refining more concrete and specific metadata details and leave other aspects like categorization or annotation to partially or totally automatic tools. The overall outcome is that, with this computer and human complementary work, it is possible to archive big amounts of content without introducing extra costs.

Semantic metadata also provides improvements in content navigability and searching, maybe in all information retrieval tasks. This fact implies a better level of productivity in the media house, e.g. while performing event tracking through a set of news in order to produce a new content. However, it is also important to take into account the gap between journalists' and archivists' mental models, which is reflected in the way archivists categorise content and journalist perform queries.

This gap is a clear threat to productivity, although the flexibility of semantic structures makes it possible to relate concepts from different mental models in order to attain a more integrated and shared view (Abelló, García, Gil, Oliva, & Perdrix, 2006), which improves the content retrieval results and consequently improves productivity.

Moreover, the combination of semantic metadata and ontologies, together with tools like the ones presented for project S5T, make it possible for journalists to navigate between content metadata and ontology concepts and benefit from an integrated and shared knowledge management effort. This feature mitigates current gaps among editorial staff that seriously reduce the possibilities of media production.

Another point of interest is the possibility that journalists produce some metadata during the content generation process. Nowadays, journalists do not consider this activity part of their job. Consequently, this task might introduce additional costs that have not been faced at the current stage of development. This remains a future issue that requires deep organisational changes, which are not present yet in most editorial staffs, even if they are trying to follow the media convergence philosophy.

To conclude, there are also the development costs necessary in order to integrate the Semantic Web tools into current media houses. As has been already noted, the choice of a smooth transition approach reduces the development costs. This approach is based on the XSD2OWL and XML2RDF mappings detailed in Section 3.1.

Consequently, it is not necessary to develop a full newspaper content management system based on Semantic Web tools. On the contrary, existing systems based on XML technologies, as it is the common case, are used as the development platform on top of which semantic tools are deployed. This approach also improves interoperability with other media houses that also use XML technologies, though the interoperation is performed at the semantic level once source metadata has been mapped to semantic metadata.

RISK ASSESSMENT

In one hand we can consider some relevant positive aspects from the proposed solution. In fact, we are introducing knowledge management into the newspaper content archive system. The proposal implies a more flexible archive system with significant improvements in search and navigation. Compatibility with current standards is kept while the archive system allows searching across media and the underlying terms and domain knowledge. Finally, the integrated view on content provides seamless access to any kind of archived resources, which could be text, audio, video streaming, photographs, etc. Consequently, separate search engines for each kind of media are no longer necessary and global queries make it possible to retrieve any kind of resources.

This feature represents an important improvement in the retrieval process but also in the archiving one. The introduction of a semi-automatic annotation process produces changes in the archivist work. They could expend more time refining semantic annotation and including new metadata. Existing human resources in the archive department should spend the same amount of time than they currently do. However, they should obtain better quality results while they populate the archive with all the semantically annotated content. The overall result is that the archive becomes a knowledge management system.

On other hand, we need to take into account some weaknesses in this approach. Nowadays, Semantic Web technologies are mainly prototypes under development. This implies problems when you try to build a complete industrial platform based on them. Scalability appears as the main problem as it was experienced during the Neptuno research project (Castells et al., 2004) also in the journalism domain.

There is a lack of implementations supporting massive content storage and management. In other words, experimental solutions cannot be applied to real system considering, as our experience has shown, more than 1 million of items, i.e. news, photos or videos. This amount can be generated in 2 or 3 months in a small news media company. A part from the lack of implementations, there is also the lack of technical staff with Semantic Web development skills.

Despite all these inconveniences, there is the opportunity to create a platform for media convergence and editorial staff tasks integration. It can become an open platform that can manage future challenges in media houses and that is adaptable to different models based on specific organizational structures. Moreover, this platform may make it possible to offer new content interaction paradigms, especially through the World Wide Web channel.

One of these potential paradigms has already started to be explored in the S5T project. Currently, it offers integrated and complementary browsing among content and the terms of the underlying domain of knowledge, e.g. politics. However, this tool is currently intended just for the editorial staff. We anticipate a future tool that makes this kind of interaction available from the Diari Segre Web site to all of its Web users. This tool would provide an integrated access point to different kinds of contents, like text or news podcasts, but also to the underlying knowledge that models events, histories, personalities, etc.

There are some threats too. First of all, any organizational change, like changing the way the archive department works or giving unprecedented annotation responsibilities to journalists, constitutes an important risk. Changes inside an organization never be easy and must be well done and follow very closely if you want to make them successful. Sometimes, the effort-satisfaction ratio may be perceived as not justified by for some journalist or archivists. Consequently, they may react against the organisational changes required in order to implement rich semantic metadata.

FUTURE TRENDS

The more relevant future trend is that the Semantic Web is starting to be recognised as a consolidated discipline and a set of technologies and methodologies that are going to have a great impact in the future of enterprise information systems (King, 2007). The more important consequence of this consolidation is that many commercial tools are appearing. They are solid tools that can be used in order to build enterprise semantic information systems with a high degree of scalability.

As has been shown, the benefits of semantic metadata are being put into practice in the Diari Segre Media Group, a newspaper that is becoming a convergent media house with press, radio, television and a World Wide Web portal. As has been detailed, a set of semantics-aware tools have been developed. They are intended for journalist and archivists in the media house, but they can be also adapted to the general public at the portal.

Making the Diari Segre semantic tools publicly available is one of the greatest opportunities and in the future, with the help of solid enterprise semantic platforms, is the issue where the greatest effort is going to be placed. In general, a bigger users base puts extra requirements about the particular needs that each user might have. This is due to the fact that each user may have a different vision about the domain of knowledge or about searching and browsing strategies. In this sense, we need some degree of personalisation beyond the much more closed approach that has been taken in order to deploy these tools for the editorial staff.

Personalisation ranges from interfaces, to processes or query construction approaches applying static or dynamic profiles. Static profiles could be completed by users in when they first register. Dynamic profiles must be collected by the system based on the user system usage (Castells et al., 2007). Per user profiles introduce a great amount of complexity, which can be mitigated building groups of similar profiles, for instance groups based on the user role.

Moreover, to collect system usage information while users navigate through the underlying conceptual structures makes it possible to discover new implicit relations among concepts with some semantic significance, at least from the user, or group to which the user belongs, point of view. If there are a lot of users following the same navigation path between items, maybe it would be better to add a new conceptual link between the initial and final items. Currently, this kind of relations can only be added manually. In the near future, we could use the power of Semantic Web technologies in order to do this automatically. This would improve user experience while they search or navigate as the underlying conceptual framework would accommodate the particular user view on the domain.

To conclude this section, it is also important to take into account the evolution of the standards upon which the ontological framework has been build. On the short range, the most import novelty is the imminent release of the NewsML G2 standard (Le Meur, 2007). This standard is also based on XML Schemas for language formalisation. Therefore, it should be trivial to generate the corresponding OWL ontologies and to start mapping metadata based on this standard to semantic metadata. More effort will be needed in order to produce the integration rules that will allow integrating this standard into existing legacy systems augmented by Semantic Web tools.

CONCLUSION

This research work has been guided by the need for a semantic journalism and multimedia metadata framework that facilitates semantic newspaper applications development in the context of a convergent media house. It has been detected, as it is widely documented in the bibliography and professional activity, that IPTC and MPEG standards are the best sources for an ontological framework that facilitates a smooth transition from legacy to semantic information systems. MPEG-7, MPEG-21 and most of the IPTC standards are based on XML Schemas and thus they do not have formal semantics.

Our approach contributes a complete and automatic mapping of the whole MPEG-7 standard to OWL, of the media packaging part of MPEG-21 and of the main IPTC standard schemas (NITF, NewsML and NewsCodes) to the corresponding OWL ontologies. Instance metadata is automatically imported from legacy systems through a XML-2RDF mapping, based on the ontologies previously mapped from the standards XML schemas. Once in a semantic space, data integration, which is a crucial factor when several sources of information are available, is facilitated enormously.

Moreover, semantic metadata facilitates the development of applications in the context of media houses that traditional newspapers are becoming. The convergence of different kinds of media, that now constitute multimedia news, poses new management requirements that are easier to cope with if applications are more informed, i.e. aware of the semantics that are implicit in news and the media that constitute them. This is the case for the tools we propose for archivists and journalists, the Media Browser and the Knowledge Browser. These tools reduce the misunderstandings among them and facilitate keeping track of existing news stories and the generation of new content.

REFERENCES

Abelló, A., García, R., Gil, R., Oliva, M., & Perdrix, F. (2006). Semantic Data Integration in a Newspaper Content Management System. In R. Meersman, Z. Tari, & P. Herrero (Eds.), *OTM Workshops 2006*. LNCS Vol. 4277 (pp. 40-41). Berlin/Heidelberg, DE: Springer.

Amann, B., Beer, C., Fundulak, I., & Scholl, M. (2002). Ontology-Based Integration of XML Web Resources. *Proceedings of the 1st International Semantic Web Conference*, ISWC 2002. LNCS Vol. 2342 (pp. 117-131). Berlin/Heidelberg, DE: Springer.

Becket, D. (2004). RDF/XML Syntax Specification. World Wide Web Consortium Recommendation.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 34–43.

Castells, P., Fernández, M., & Vallet, D. (2007). An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. *IEEE Transactions on Knowledge and Data Engineering*, *19*(2), 261–272. doi:10.1109/TKDE.2007.22

Castells, P., Perdrix, F., Pulido, E., Rico, M., Benjamins, R., Contreras, J., et al. (2004). Neptuno: Semantic Web Technologies for a Digital Newspaper Archive. In C. Bussler, J. Davies, D. Fensel, & R. Studer, (Eds.), *The Semantic Web: Research and Applications: First European Semantic Web Symposium*, ESWS 2004, LNCS Vol. 3053 (pp. 445-458). Berlin/Heidelberg, DE: Springer.

Castells, P., Perdrix, F., Pulido, E., Rico, M., Benjamins, R., Contreras, J., & Lorés, J. (2004). *Neptuno: Semantic Web Technologies for a Digital Newspaper Archive*. LNCS Vol. 3053 (pp. 445-458).Berlin/Heidelberg, DE: Springer.

Cruz, I., Xiao, H., & Hsu, F. (2004). An Ontology-based Framework for XML Semantic Integration. *Proceedings of the Eighth International Database Engineering and Applications Symposium*, IDEAS'04, (pp. 217-226). Washington, DC: IEEE Computer Society.

Cuayahuitl, H., & Serridge, B. (2002). Out-of-vocabulary Word Modelling and Rejection for Spanish Keyword Spotting Systems. *Proceedings of the 2nd Mexican International Conference on Artificial Intelligence*.

Einhoff, M., Casademont, J., Perdrix, F., & Noll, S. (2005) ELIN: A MPEG Related News Framework. In M. Grgic (Ed.), *47th International Symposium ELMAR: Focused on Multimedia Systems and Applications* (pp.139-142). Zadar, Croatia: ELMAR.

Eriksen, L. B., & Ihlström, C. (2000). Evolution of the Web News Genre - The Slow Move Beyond the Print Metaphor. In *Proceedings of the 33rd Hawaii international Conference on System Sciences*. IEEE Computer Society Press.

Fernández, N., Blázquez, J. M., Fisteus, J. A., Sánchez, L., Sintek, M., Bernardi, A., et al. (2006). NEWS: Bringing Semantic Web Technologies into News Agencies. *The Semantic Web - ISWC 2006*, LNCS Vol. 4273 (pp. 778-791). Berlin/Heidelberg, DE: Springer.

García, R. (2006). XML Semantics Reuse. In *A Semantic Web Approach to Digital Rights Management*, PhD Thesis (pp. 116-120). TDX.

García, R., & Gil, R. (2006). Improving Human-Semantic Web Interaction: The Rhizomer Experience. *Proceedings of the 3rd Italian Semantic Web Workshop*, SWAP'06, Vol. 201 (pp. 57-64). CEUR Workshop Proceedings.

García, R., Gil, R., & Delgado, J. (2007). A Web ontologies framework for digital rights management. *Artificial Intelligence and Law*, *15*(2), 137–154. doi:10.1007/s10506-007-9032-6

García, R., Gil, R., Gallego, I., & Delgado, J. (2005). Formalising ODRL Semantics using Web Ontologies. In R. Iannella, S. Guth, & C. Serrao, Eds., *Open Digital Rights Language Workshop*, ODRL'2005 (pp. 33-42). Lisbon, Portugal: ADETTI.

Hausenblas, M., Troncy, R., Halaschek-Wiener, C., Bürger, T., & Celma, O. Boll, et al. (2007). *Multimedia Vocabularies on the Semantic Web*. W3C Incubator Group Report, World Wide Web Consortium.

Haustein, S., & Pleumann, J. (2002). Is Participation in the Semantic Web Too Difficult? In *Proceedings of the First International Semantic Web Conference on The Semantic Web*, LNCS Vol. 2342 (pp. 448-453). Berlin/Heidelberg: Springer.

Herzog, C., Luger, M., & Herzog, M. (2007). Combining Social and Semantic Metadata for Search in Document Repository.Bridging the Gap Between Semantic Web and Web 2.0. *International Workshop at the 4th European Semantic Web Conference* in Insbruck, Austria, June 7, 2007.

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004). *SWRL: A Semantic Web Rule Language Combining OWL and RuleML*. W3C Member Submission, World Wide Web Consortium.

Hunter, J. (2001). Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. *Proceedings of the International Semantic Web Working Symposium* (pp. 260-272). Standford, CA.

Hunter, J. (2003). Enhacing the Semantic Interoperability of Multimedia through a Core Ontology. *IEEE Transactions on Circuits and Systems for Video Technology*, *13*(1), 49–58. doi:10.1109/TCSVT.2002.808088

Ihlström, C., Lundberg, J., & Perdrix, F. (2003) Audience of Local Online Newspapers in Sweden, Slovakia and Spain - A Comparative Study. In *Proceedings of HCI International* Vol. 3 (pp. 749-753). Florence, Kentucky: Lawrence Erlbaum Associates.

Kim, J., Jung, H., & Chung, H. (2004). A Keyword Spotting Approach based on Pseudo N-gram Language Model. *Proceedings of the 9th Conf. on Speech and Computer*, SPECOM 2004 (pp. 256-259). Patras, Greece.

King, R. (2007, April 29). Taming the World Wide Web. *Special Report, Business Week*.

Klein, M. C. A. (2002). Interpreting XML Documents via an RDF Schema Ontology. In *Proceedings of the 13th International Workshop on Database and Expert Systems Applications*, DEXA 2002 (pp. 889-894). Washington, DC: IEEE Computer Society.

Lakshmanan, L., & Sadri, F. (2003). Interoperability on XML Data. *Proceedings of the 2nd International Semantic Web Conference*, ICSW'03, LNCS Vol. 2870 (pp. 146-163). Berlin/Heidelberg: Springer.

Le Meur, L. (2007). How NewsML-G2 simplifies and fuels news management. Presented at *XTech 2007: The Ubiquitous Web*, Paris, France.

Lundberg, J. (2002). *The online news genre: Visions and state of the art*. Paper presented at the 34th Annual Congress of the Nordic Ergonomics Society, Sweden.

McDonald, N. (2004). Can HCI shape the future of mass communications. *Interaction*, *11*(2), 44–47. doi:10.1145/971258.971272

McGuinness, D. L., & Harmelen, F. V. (2004). *OWL Web Ontology Language Overview*. World Wide Web Consortium Recommendation.

Patel-Schneider, P., & Simeon, J. (2002). The Yin/Yang Web: XML syntax and RDF semantics. *Proceedings of the 11th International World Wide Web Conference*, WWW'02 (pp. 443-453). ACM Press.

Salembier, P., & Smith, J. (2002). Overview of MPEG-7 multimedia description schemes and schema tools. In B.S. Manjunath, P. Salembier, & T. Sikora (Ed.), *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons.

Salton, G., & McGill, M. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Sawyer, S., & Tapia, A. (2005). The sociotechnical nature of mobile computing work: Evidence from a study of policing in the United States. *International Journal of Technology and Human Interaction*, *1*(3), 1–14.

Tejedor, J., García, R., Fernández, M., López, F., Perdrix, F., Macías, J. A., et al. (2007). Ontology-Based Retrieval of Human Speech. *Proceedings of the 6th International Workshop on Web Semantics*, WebS'07 (in press). IEEE Computer Society Press.

Tous, R., García, R., Rodríguez, E., & Delgado, J. (2005). Arquitecture of a Semantic XPath Processor. In K. Bauknecht, B. Pröll, & H. Werthner, Eds., *E-Commerce and Web Technologies: 6th International Conference*, EC-Web'05, LNCS Vol. 3590 (pp. 1-10). Berlin/Heidelberg, DE: Springer.

Tsinaraki, C., Polydoros, P., & Christodoulakis, S. (2004). Integration of OWL ontologies in MPEG-7 and TVAnytime compliant Semantic Indexing. In A. Persson, & J. Stirna, Eds., *16th International Conference on Advanced Information Systems Engineering*, LNCS Vol. 3084 (pp. 398-413). Berlin/Heidelberg, DE: Springer.

Tsinaraki, C., Polydoros, P., & Christodoulakis, S. (2004). Interoperability support for Ontology-based Video Retrieval Applications. *Proceedings of 3rd International Conference on Image and Video Retrieval, CIVR 2004*. Dublin, Ireland.

Vanderwal, T. (2007) *Folksonomy Coinage and Definition*.

Vorhees, E. (1994). Query expansion using lexical semantic relations. *Proceedings of the 17th ACM Conf. on Research and Development in Information Retrieval*, ACM Press.

ADDITIONAL READING

Kompatsiaris, Y., & Hobson, P. (Eds.). (2008). *Semantic Multimedia and Ontologies: Theory and Applications*. Berlin/Heidelberg, DE: Springer.

ENDNOTES

1 http://www.diarisegre.com
2 IPTC, http://www.iptc.org
3 IIM, http://www.iptc.org/IIM
4 http://www.iptc.org/NewsCodes
5 NITF, http://www.nitf.org
6 http://www.newsml.org
7 http://rhizomik.net/redefer
8 http://rhizomik.net/semanticnewspaper
9 http://rhizomik.net/ontologies/mpeg7ontos
10 http://nets.ii.uam.es/s5t
11 Milenium Quay, http://www.mileniumcross-media.com

This work was previously published in Semantic Web for Business: Cases and Applications, edited by R. García, pp. 170-193, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 4.9
Applying Semantic Web to E-Tourism

Danica Damljanović
University of Sheffield, UK

Vladan Devedžić
University of Belgrade, Serbia

ABSTRACT

Traditional E-Tourism applications store data internally in a form that is not interoperable with similar systems. Hence, tourist agents spend plenty of time updating data about vacation packages in order to provide good service to their clients. On the other hand, their clients spend plenty of time searching for the 'perfect' vacation package as the data about tourist offers are not integrated and are available from different spots on the Web. We developed Travel Guides - a prototype system for tourism management to illustrate how semantic web technologies combined with traditional E-Tourism applications: a.) help integration of tourism sources dispersed on the Web b) enable creating sophisticated user profiles. Maintaining quality user profiles enables system personalization and adaptivity of the content shown to the user. The core of this system is in ontologies – they enable machine readable and machine understandable representation of the data and more importantly reasoning.

INTRODUCTION

A mandatory step on the way to the desired vacation destination is usually contacting tourist agencies. Presentations of tourist destinations on the Web make a huge amount of data. These data are accessible to individuals through the official presentations of the tourist agencies, cities, municipalities, sport alliances, etc. These sites are available to everyone, but still, the problem is to find useful information without wasting time. On the other hand, plenty of systems on the Web are maintained regularly to provide tourists with up-to-date information. These systems require a lot of efforts from humans - especially in travel agencies where they want to offer tourists a good service.

We present Travel Guides – a prototype system that is combining Semantic Web technologies with those used in mainstream applications (cp. Djuric, Devedzic & Gasevic, 2007) in order to enable data exchange between different E-Tourism systems and thus:

- Ease the process of maintaining the systems for tourist agencies
- Ease the process of searching for perfect vacation packages for tourists

The core of Travel Guides system is in ontologies. We have developed domain ontology for tourism and described the most important design principles in this chapter.

As ontologies enable presenting data in a machine-readable form thus offering easy exchange of data between different applications, this would lead to increased interoperability and decreased efforts tourist agents make to update the data in their systems. To illustrate increased interoperability we initialized our knowledge base using data imported from some other system. We built an environment to enable transferring segments of any knowledge base to the other by selecting some criteria - this transfer is possible even if the knowledge bases rely on different ontologies.

Ontology-aware systems provide the possibility to perform semantic search – the user can search the destinations covered by Travel Guides using several criteria related to travelling (e.g., accommodation rating, budget, activities and interests: concerts, clubbing, art, sports, shopping, etc.). For even more sophisticated search results we introduce user profiles created based on data that system possesses about the user. These data are analysed by a reasoner, and the heuristics is residing inside the ontology.

The chapter is organized as follows: in next section we describe different systems that are developed in the area of tourism which use semantic web technologies. In the central section we first discuss problems that are present in existing E-Tourism systems, and then describe how we solve some of these problems with Travel Guides: we give details of the design of the domain ontology, the creation of the knowledge base and finally system architecture. To illustrate Travel Guides environment we give an example of using this system by providing some screenshots. Finally, we conclude and give the ideas of future work and also future research directions in the field.

BACKGROUND

E-Tourism comprises electronic services which include (Aichholzer, Spitzenberger & Winkler, 2003):

- Information Services (IS), e.g. destination, hotel information.
- Communication Services (CS), e.g. discussion forums, blogs.
- Transaction Services (TS), e.g. booking, payment.

Among these three services *Information Services* are the most present on the Web. Hotels usually have their Web sites with details about the type of accommodation, location, and contact information. Some of these Web sites even offer *Transaction Services* so that it is possible to access the prices and availability of the accommodation for the requested period and perform booking and payment.

Transaction Services are usually concentrated on sites of Web tourist agencies such as Expedia, Travelocity, Lastminute, etc. These Websites sometimes include *Communication Services* in the form of forums where people who visited hotels give their opinion and reviews. With emerging popularity of social web applications many sites specialize in CS only (e.g., www.43places.com).

However, for complete details about a certain destination (e.g., activities, climate, monuments, and events) one often must search for several sources. Apparently all of these sources are dispersed on different places on the Internet and there is an "information gap" between them. The best way to bridge this gap would be to enable communication between different tourist applications.

For *Transaction Services* this is already partly achieved by using Web portals that serve as mediators between tourists and tourist agencies. These portals (e.g., Bookings.com) gather vacation packages from different vendors and use Web services to perform booking and sometimes payment. *Communication Services* are tightly coupled with *Information Services*, in a way that the integration of the first implies the integration of the latter. Henriksson (2005) discusses that the one of the main reasons for lack of interoperability in the area of tourism is the tourism product itself: immaterial, heterogeneous and non-persistent. Travel Guides demonstrates how Semantic Web technologies can be used to enable communication between *Information Services* dispersed on the Web. This would lead to easier exchange of communication services, thus resulting in better quality of E-Tourism and increased interoperability.

Hepp, Siorpaes and Bachlechner (2006) claim that "Everything is there, but we only have insufficient methods of finding and processing what's already on the Web" (p. 2). This statement reveals some of the reasons why Semantic Web is not frequently applied in real-time applications: Web today contains content understandable to humans hence only humans can analyse it. To retrieve information from applications using computer programs (e.g., intelligent agents) two conditions must be satisfied: 1) data must be in a machine-readable form 2) applications must use technologies that provide information retrieval from this kind of data.

Many academic institutions are making efforts to find methods for computer processing of human language. GATE (General Architecture for Text Engineering) is an infrastructure for developing and deploying software components that process human language (Cunningham, 2002). It can annotate documents by recognizing concepts such as: locations, persons, organizations and dates. It can be extended to annotate some domain-related concepts, such as hotels and beaches.

The most common approaches for applying Semantic Web in E-Tourism are:

1. Making applications from scratch using recommended standards
2. Using ontologies as mediators to merge already existing systems
3. Performing annotations in respect to the ontology of already existing Web content

One of the first developed E-Tourism systems was onTour (http://ontour.deri.org/) developed by DERI (Siorpaes & Bachlechner, 2006; Prantner, 2004) where they built a prototype system from scratch and stored their data in the knowledge base created based on the ontology. They developed domain ontology following the World Tourism Organization standards, although they considered a very limited amount of concepts and relations. Later on, they took over the ontology developed as a part of Harmonize project and now planning to develop an advanced E-Tourism Semantic Web portal to connect the customers and virtual travel agents (Jentzsch, 2005).

The idea of Harmonize project was to integrate Semantic Web technologies and merge tourist electronic markets yet avoiding forcing tourist agencies to change their already existing information systems, but to merge them using ontology as a mediator (Dell'erba, Fodor, Hopken, & Werthner, 2005).

The third approach is very challenging for researchers as with the current state of the Web it is not easy to add semantics to the data without changing the technologies used to develop the Web applications. Cardoso (2006) presents a system that creates vacation packages dynamically using previously annotated data in respect to the ontology. This is performed with a service that constructs itinerary by combining user preferences with flights, car rentals, hotel, and activities on-fly. In 2005, Cardoso founded a lab for research in the area of Semantic Web appliance in E-Tourism. The main project called SEED (Semantic E-Tourism Dynamic packaging) aims to illustrate the appliance of Web services and Semantic Web in the area of tourism. One of the main objectives of this project is the development of OTIS ontology (Ontology for Tourism Information Systems). Although they discuss the comprised concepts of this ontology, its development is not yet finished, and could not be further discussed in this chapter.

On the other side, Hepp et al. (2006) claim that there are not enough data in the domain of tourism available on the Web - at least for Tyrol, Austria. Their experiment revealed that existing data on the Web are incomplete: the availability of the accommodation and the prices are very often inaccessible.

Additionally, most of E-Tourism portals store their data internally, which means that they are not accessible by search engines on the Web. Using Semantic Web services, e.g. Web Service Modelling Ontology - WSMO (Roman et al., 2005) or OWL-based Web service ontology - OWL-S (Smith & Alesso, 2005) it would be possible to access data from data-intensive applications. SATINE project is about deploying semantic travel Web services. In (Dogac et al., 2004) they present how to exploit semantics through Web service registries.

Semantic Web services might be a good solution for performing E-Tourism *Transaction Services*, and also for performing E-Tourism *Information Services*, as they enable integrating homogenous data and applications. However, using Semantic Web services, as they are applied nowadays, will not reduce every-day efforts made by tourist agents who are responsible for providing current data about vacation packages and destinations. Data about different destinations are not static – they change over time and thus require E-Tourism systems to be updated. With the current state of the development of E-Tourism applications, each travel agency performs data update individually.

In Travel Guides we employ Semantic Web technologies by combining the first and the third approach. We use the first approach to build the core of the system, and to initialize the repository, whereas in later phase we propose using annotation tools such as GATE to perform semi-automatic annotation of documents and update of knowledge base accordingly. Some of the existing Knowledge Management platforms such as KIM (Popov, Kiryakov, Ognyanoff, Manov & Kirilov, 2004) use GATE for performing automatic annotation of documents and knowledge base enrichment. Due to the very old and well-known problem of syntactic ambiguity (Church & Patil, 1982) of human language widely present inside the Web content that is used in the process of annotation, we argue that the role of human is irreplaceable.

The core of the Travel Guides system is in ontologies. Many ontologies have been already developed in the area of tourism. Bachlechner (2004) has made a long list of the areas that need to be covered by E-Tourism relevant ontologies and made a brief analysis of the developed domain and upper level ontologies. Another good summary of E-Tourism related ontologies is given in (Jentzsch, 2005).

However, no ontology includes all concepts and relations between them in such a way that it can be used without any modifications, although some of them such as Mondeca's (http://www.mondeca.com) or OnTour's ontology (Prantner, 2004) are developed following World Tourism Organization standards. While developing Travel Guides ontology we tried to comprise all possible concepts that are related to the area of tourism and also - tourists. Concepts and relations that describe user's activities and interests coupled with built-in reasoner enable identifying the user as a particular type: some tourists enjoy comfort during vacation, whereas others don't care about the type of the accommodation but more about the outdoor activities or the scenery that is nearby.

Most of the developed ontology-aware systems nowadays propose using a RDF repository instead of using conventional databases (Stollberg, Zhdanova & Fensel, 2004). RDF repositories are not built to replace conventional databases, but to add a refinement which is not supported by conventional databases, specifically – to enable representing machine-readable data and reasoning. In Travel Guides system we distinguish between data that are stored in RDF repositories and those that are stored in conventional databases. In RDF repositories we store machine-understandable data used in the process of reasoning, and relational databases are used to store and retrieve all other data – those that are not important in this process and also being specific for each travel agency which means they are not *sharable*. We propose *sharable* data to be those that could be easily exchangeable between applications. This way, applications can share a unique repository which means that if, for instance, a new hotel is built on a certain destination and one tourist agency updates the repository, all others can use it immediately.

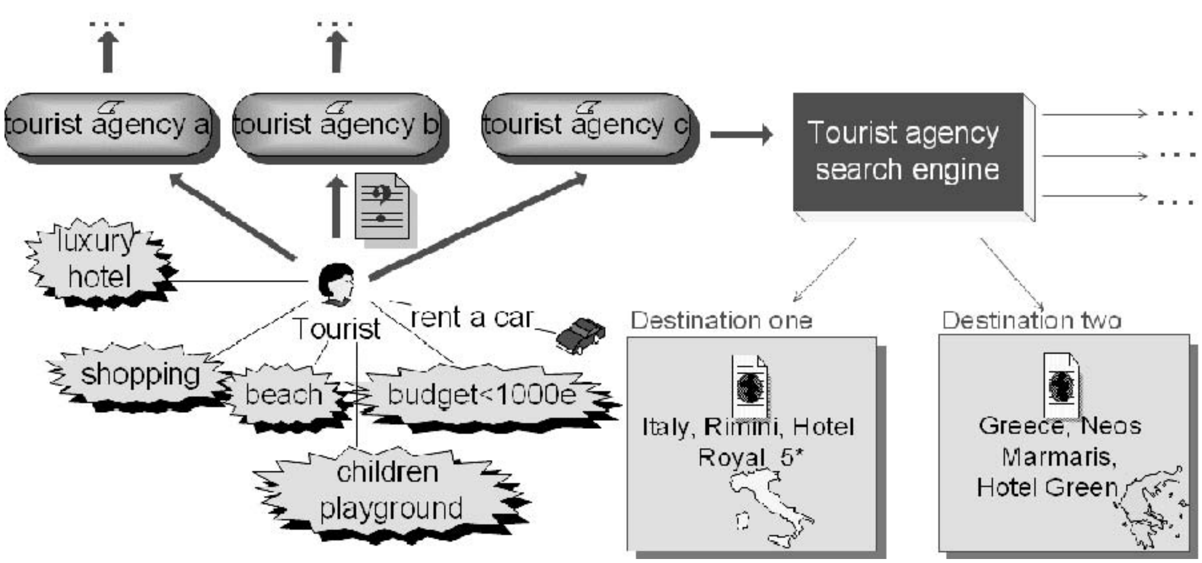
We suggest this approach as Semantic Web technologies nowadays are still weak to handle a huge amount of data, and could not be compared by performance with relational databases in the terms of transaction handling, security, optimization and scalability (cf. Guo, Pan & Heflin, 2004).

APPLYING SEMANTIC WEB TO E-TOURISM

E-Tourism Today

Searching for information on a desired spot for vacation is usually a very time-consuming. Figure 1 depicts the most frequent scenario which starts with the vague ideas of the user interested in travelling, and ends with the list of tourist destinations. In most of cases the user is aware of a few criteria that should be fulfilled (the distance from the shopping centre, sandy beach, a

Figure 1. The usual scenario of searching the Internet for a 'perfect' vacation package



possibility to rent a car, etc.), as well as of some individual constraints (prices, departure times, etc.). After processing the user's query (using these criteria as input data), the search engine of a tourist agency will most likely return a list of vacation packages. It is up to the user to choose the most appropriate one. If the user is not satisfied with the result, the procedure is repeated, with another tourist agency. This scenario is restarted in N iterations until the user gets the desired result. The essential disadvantage of this system is a lack of the integrated and ordered collection of the tourist deals. Tourist deals are dispersed on the Web and being offered from different tourist agencies each of which maintain their system independently.

Additional problem with existing E-Tourism applications is the lack of interactivity. It is always the user who provides the criteria for the search/query and who analyzes the results returned.

The problem of dispersed information about tourist deals would be reduced totally if all vacation packages would be gathered at one place - the Web portal. This assumption could not be taken as realistic, but apparently the distribution of the tourist offers would be decreased by adding the tourist offers of each tourist agency into the portal. Although the portals are more sophisticated than simple Web applications, they usually do not compensate for the lack of interactivity.

Some of the popular Web tourist agencies, such as Expedia, expand their communication with the user by offering various services based on user selection during visiting their site. Namely, they track user actions (mouse clicks) so that when user browse throughout the site the list of user recently visited places is always available. In case the user provides his personal e-mail they send some special offers or advertisements occasionally. Although their intentions are to improve communication with their clients, this kind of service can be irritating sometimes. Developing more sophisticated user profiles would help developing more personalized systems thus avoiding spamming the user with an unattractive content.

Creating user profiles is widely used in many applications nowadays, not only in the area of tourism. In order to create his/her profile, the user is usually prompted to register and fill in few forms with some personal info such as location, year of birth, interests, etc. Filling these forms sometimes can take a lot of time and thus carries the risk of 'refusing' the user. The best way is to request a minimum data from the user on his first log in, and then update his data later step by step.

Travel Guides

Disadvantages of traditional E-Tourism systems imply requirements for a new system that is focusing on the integration of *Information Services* present on the Web and also on introducing sophisticated *user profiles*. Travel Guides prototype system has been developed to satisfy these requirements by combining semantic web technologies with those used in traditional E-Tourism systems.

Using semantic web technologies enables representing the data in machine-readable form. Such a representation enable easier integration of tourist resources as data exchange between applications is feasible. Integration of tourist resources would decrease efforts tourist agents make in tourist agencies to maintain these data. The final result would affect the tourist who will be able to search for details about destinations from the single point on the Web.

In Travel Guides we introduce more sophisticated user profiles – these are to enable personalization of the Web content and to act as agents who work for users, while not spamming them with commercial content and advertisements. For example, if during registration the user enters that he is interested in extreme sports, and later moves on to the search form where he does not specify any sport requirement, the return results could be filtered in a manner that the first listed are those that are flagged as "adventurer destinations".

Developing sophisticated user profiles requires analyse of the user behaviour while visiting the portal. This behaviour is determined by the data the system collects about the user: his personal data, interests, activities, and also the data that system tracks while 'observing' the user: user selection, mouse clicks, and the like. To be able to constantly analyze the user's profile the portal requires intelligent reasoning. To make a tourism portal capable of intelligent reasoning, it is necessary to build some initial and appropriate knowledge in the system, as well as to maintain the knowledge automatically from time to time and during the user's interaction with the system. Simply saving every single click of the user could not be enough to make a good-quality user profile. It is much more suitable to use a built-in reasoner to infer the user's preferences and intentions from the observations.

Any practical implementation of the aforementioned requirements leads to representing essential knowledge about the domain (tourism) and the portal users (user profiles) in a machine-readable and machine-understandable form. In other words, it is necessary to develop and use a set of ontologies to represent all important concepts and their relations.

When ontologies are developed, it is necessary to populate the knowledge base with instances of concepts from the ontology and with relevant relations. After some knowledge is created, it needs to be coupled with a built-in inference engine to support reasoning. Finally, it is essential to enable input in the system from the user as reasoning requires some input data to be processed.

In the next sections we describe the ontology we developed to satisfy the requirements, followed by the knowledge base creation and the architecture of the system that enables processing the input from end users.

Travel Guides Ontologies

The Travel Guides Ontologies are written in OWL (Antoniou & Harmelen, 2004) and developed using Protégé (Horridge, Knublauch, Rector, Stevens & Wroe, 2004). To develop a well-designed ontology, it was important to:

1. *Include all important terms in the area of tourism* to represent destinations in general, excluding data specific for any tourist agency. For instance, information about a city name, its latitude and longitude, and the country it belongs to is to be included here.
2. *Classify user interests and activities* so that they can be expressed in the manner of a collection of user profiles, and identify the concepts to represent them.
3. *Identify concepts to represent the facts about destinations* that are specific for each tourist agency. This information is extracted from expert knowledge, where an expert is a tourist agent who would be able to classify destinations according to the different criteria; for instance, if the destination is a family destination, a romantic destination, etc. After identifying these concepts they need to be connected with other relevant concepts, e.g. create relations between destination types and relevant user profile types.

Representing aforementioned three steps in a manner of a formal representation of concepts and relations results in the creation of the following:

1. *The World ontology*, with concepts and relations from the real world: geographical terms, locations with coordinates, land types, time and date, time zone, currency, languages, and all other terms that are expressing concepts that are in a way related to tourism or tourists, but not to vacation

packages that could be offered by some tourist agencies. This ontology should also contain the general concepts necessary for expression of semantic annotation, indexing, and retrieval (Kiryakov et al., 2003).

2. *The User ontology* containing concepts related to the users – the travellers who visit the Travel Guides portal. This ontology describes user interests and activities, age groups, favourite travel companies, and other data about different user profiles.
3. *The Travel (Tourism) ontology* contains concepts related to vacation packages, types of vacations, and traveller types w.r.t. various tourist destinations. It includes all terms being specific to vacation packages offered in tourist agencies and being important for travellers, like the type of accommodation, food service type, transport service, room types in a hotel, and the like. It is this ontology that makes a connection between users and destinations. This is accomplished by creating user profiles for the users, and determining the type of destination for each vacation package. Finally, user profiles are linked to relevant destination types (and vice versa).

After the evaluation of existing domain and upper-level ontologies, we have found that the one that suits the Travel Guides the best is the PROTON ontology (Terziev, Kiryakov and Manov, 2005). PROTON upper-level ontology includes four modules, each of which is a separate ontology. For the purpose of Travel Guides development, the Upper module of PROTON (Terziev et al., 2005) was used as *the World ontology*. This module was extended to fit *the Tourism (Travel) ontology*. The PROTON Knowledge management module (Terziev et al., 2005) was extended to serve as *the User ontology*.

The World Ontology

The PROTON upper level ontology contains all concepts required by Travel Guides World ontology. In addition, it contains concepts and relations necessary for information extraction, retrieval and semantic annotation. PROTON class we used the most frequently in our World ontology is the class **Location**. Figure 2 depicts the hierarchy of the class **Location** and its subclasses in the PROTON Upper Level ontology.

The classes and properties from PROTON used in Travel Guides are shown in Figure 3. Following aliases have been used instead of full namespaces: pkm for PROTON Knowledge Management, psys for PROTON System Module, ptou for PROTON Upper Module, and ptop for Proton Top module.

For more information about PROTON ontology we refer reader to (Terziev et al., 2005).

The User Ontology

PROTON Knowledge Management (KM) ontology has been extended to suit *the User ontology* needs. The most frequently used classes are: **User**, **UserProfile**, and **Topic**. According to the PROTON documentation, **Protont:Topic** (the PROTON top module class) is “any sort of a topic or a theme, explicitly defined for classification purposes”. For the needs of Travel Guides, **protont:Topic** class has been extended to represent user interests and activities. Its important relations and concepts are depicted in Figure 4.

For determining user profile types, the age and the user preferred travel company is of a great importance hence relevant concepts have been created inside the ontology: **AgeGroup** is a representation of the first and the **TravelCompany** is a representation of the latter (Figure 5). For example, if the user selects that he/she travels with family very often, he/she could be considered as a **FamilyType**.

Figure 2. The Location class and its subclasses in the PROTON Upper Level ontology

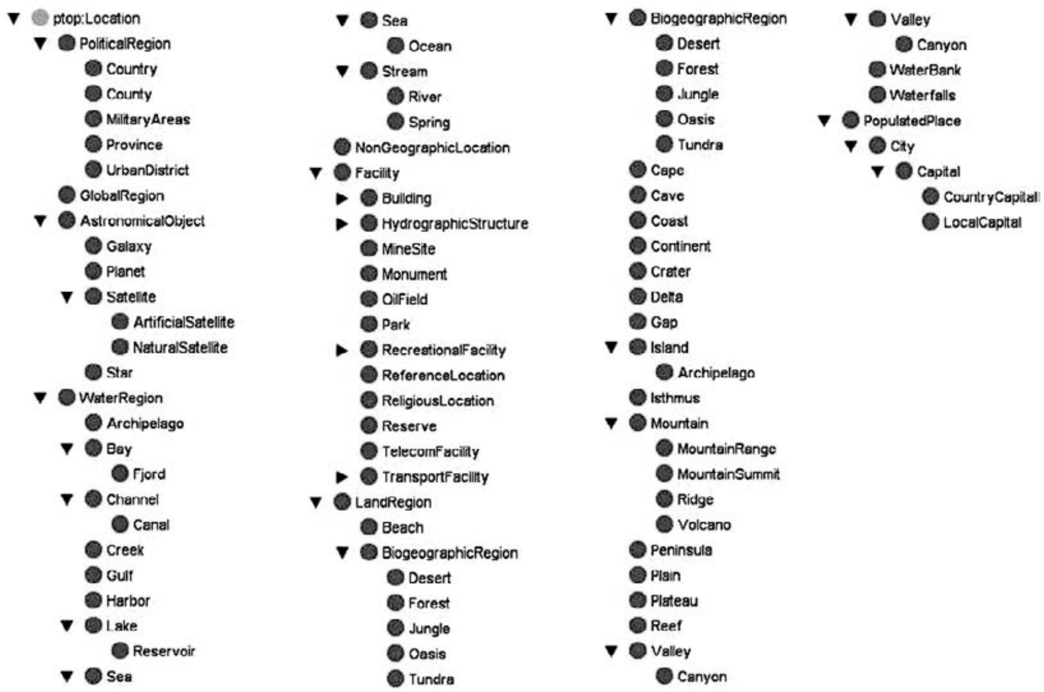
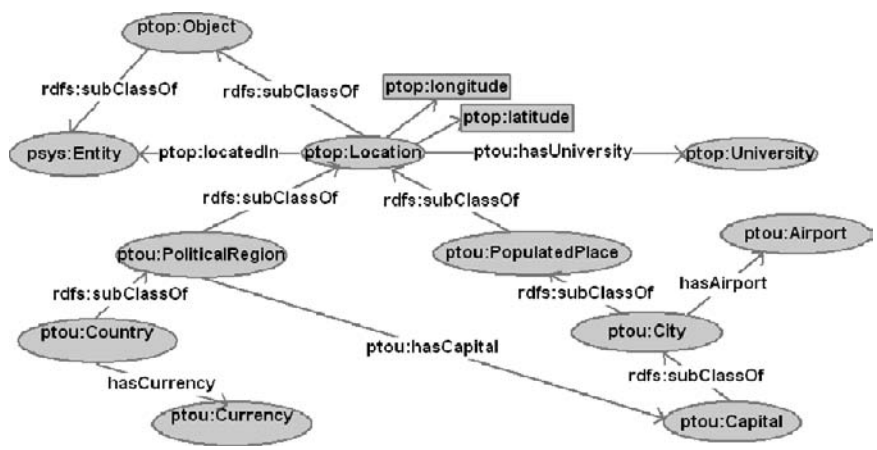


Figure 3. The classes and properties from the PROTON ontology frequently used in Travel Guides



The **UserProfile** class is extended to represent various types of tourists. These profiles are made based on user interests and activities. Figure 6 depicts various types of user profiles.

In practice, many tourists would be determined to belong to more than one type of user profiles. For this purpose, there is a property *weight* that could be assigned for each type of user profile. It is this property that reflects the importance of certain profiles. For example:

User *hasUserProfile* **Adventurer** (*weight* = 2),
User *hasUserProfile* **ClubbingType** (*weight* = 1).

Figure 4. The most important concepts and their relation in the User ontology

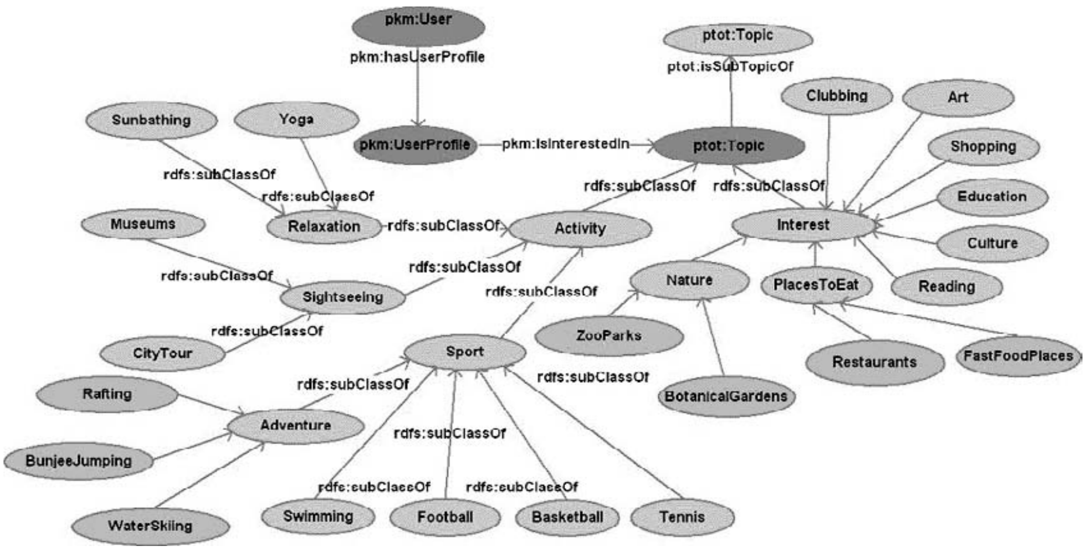
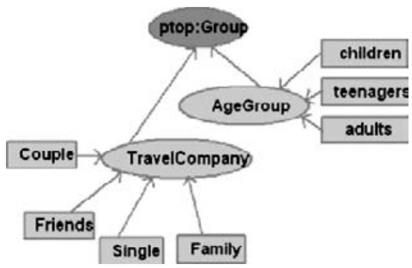


Figure 5. Extension of PROTON Group class



In this case the user profile is a mixture of the **Adventurer** and the **Clubbing** type, but due to the *weight* values adventure destinations have a priority over those that are "flagged" as great-night-life destinations.

Travel Guides User Ontology is available on-line at http:// goodoldai.org.yu/ns/upproton.owl.

Tourism (Travel) Ontology

In order to design the domain ontology for the area of tourism as well as to "link" tourist destination types to the user profile types, we extended the PROTON Upper module ontology. The class **Offer** is extended with the subclass of **TouristOffer** representing a synonym term for vacation package offered in a tourist agency. Figure 7 depicts the **TouristOffer** class and types of destinations assigned to tourist offers. These types are used as indicators of types of tourist offers which are later being assigned to relevant user profile types.

Figure 8 depicts classes and relations between them in *the Travel ontology*. Since the Travel ontology is an extension of the PROTON Upper module ontology, there are some concepts and relations from PROTON that are frequently used. They all have appropriate prefixes.

As shown on Figure 8, a vacation package being an instance of **TouristOffer** class *isAttractiveFor* certain type of **UserProfile**, where this type is determined by user's interests and activities.

Travel Guides Travel Ontology is available on-line at http://goodoldai.org.yu/ns/tgproton.owl.

Travel Guides Knowledge Base

Due to a huge amount of data that is stored inside the knowledge base (KB), it is essential that its structure allows easy maintenance. To meet this requirement, we represent the KB as a collection of *.owl* files (Figure 9).The circle on the top represents the core and contains concepts such as continents and countries used by all other parts of the KB.

Figure 6. Subclasses of PROTON UserProfile class

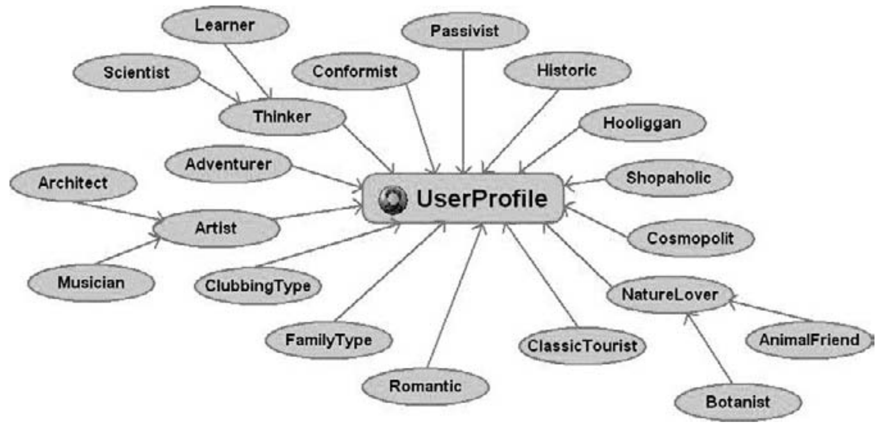
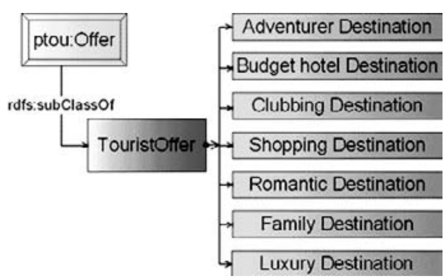


Figure 7. The extended ptou:Offer class in Travel Guides Tourism ontology



The other parts are independent *.owl* files that are country specific and contain all destinations inside the country, all hotels on the destinations and finally all vacation packages related to the hotels. For the clarity of the presentation Figure 9 depicts only 3 elements of the KB apart from the core. Ideally, the number of these elements is equal to the number of existing countries.

To alleviate the creation, extensions, and maintenance of the KB, and also to address the interoperability issue, we explored some other ontology-based systems that include instances of concepts that are of interest to Travel Guides system. We have built an environment that enables exploiting instances of classes (concepts) and relations of the arbitrary KB in accordance to the predefined criteria. We considered using KIM KB and also WordNet (Fellbaum, 1998). As KIM KB contains more data that are of interest to Travel Guides system and also is built based on the ontology whose core is PROTON ontology (Popov et al, 2004), we successfully exploited it to build our core (continents and countries). This core is available online at http:// goodoldai.org.yu/ns/travel_wkb.owl, and is used to initialize other elements of the KB.

This way we avoided entering permanent data about various destinations manually, and also showed that it is possible to share the knowledge between different platforms when it is represented using RDF structure and achieve interoperability - the content of one application can be of use inside the other application, even if they are based on different ontologies. Our environment for knowledge base exploitation is applicable for any knowledge base and ontology; the only precondition is selection of criteria that will define the statements to be extracted.

Apart from many concepts (e.g., organizations and persons), KIM Platform KB includes data about continents, countries and many cities. The environment created inside Travel Guides enables extracting of concepts by selecting some of the criteria, e.g., name of the property. We selected *hasCapital*, as this property has class **Country** as a domain and class **City** as a range. Our environment extracts not only the concepts that are directly related to the predefined property, but also

Figure 8. Concepts and relations in the Travel Guides Travel ontology

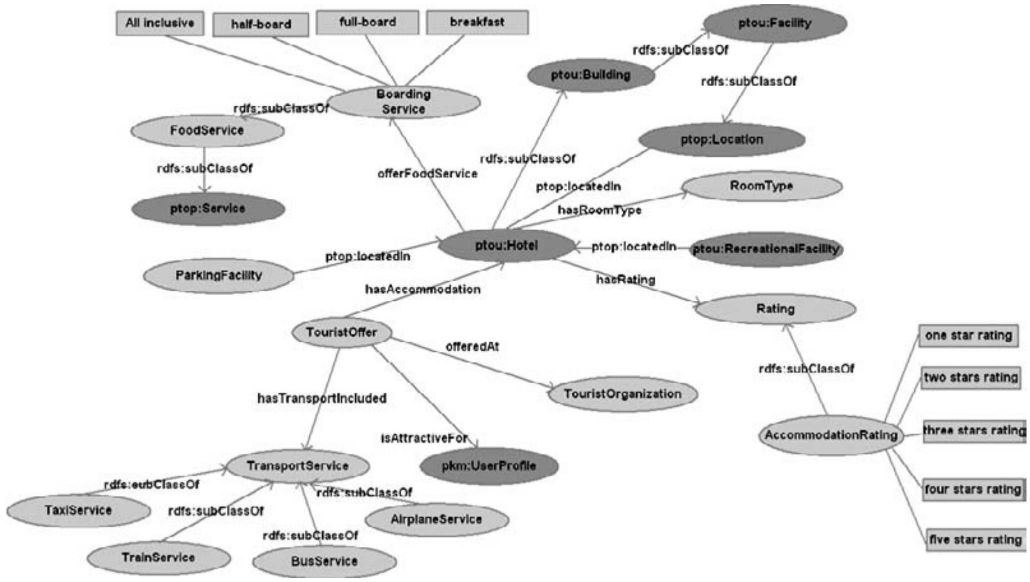
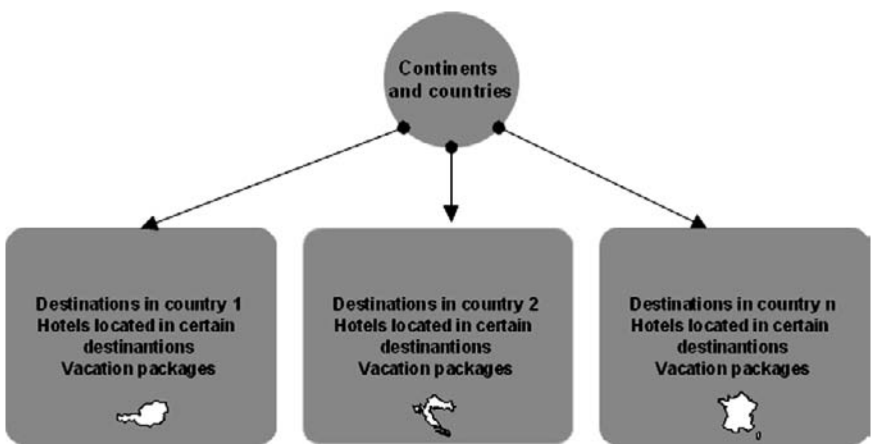


Figure 9. Organization of the knowledge base inside the Travel Guides system



all other statements that are the result of transitive relations of this property. For example, if defined relation **Country** *isLocatedIn* **Continent** exists, statements that represent this relation will also be extracted.

Figure 10 depicts some of the classes and relations whose instances are imported during the KB extraction.

Ideally, the knowledge base should contain descriptions of all destinations that could (but need not necessarily) be included in the offers of the tourist agencies connected to the portal.

The Portal Architecture

This section gives details about the architecture of Travel Guides system (Figure 11) and its design. The system comprises following four modules:

Figure 10. Classes and relations whose concepts are imported during KB extraction

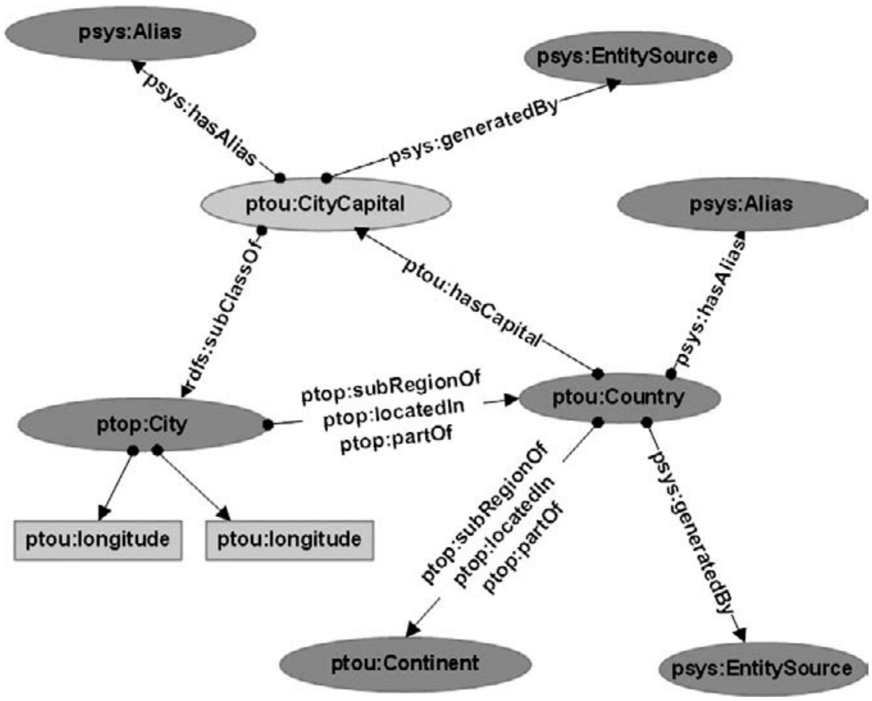
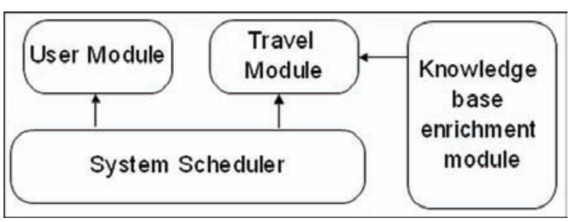


Figure 11. Travel Guides Architecture



1. **User Module:** For generating user profiles and maintaining user data.
2. **Travel Module:** For generating and maintaining vacation packages and all other data related to vacation packages and destinations.
3. **System Scheduler Module:** For update of the knowledge base. It communicates with:
   - User Module when updating user profiles
   - Travel Module when updating data about vacation packages and destinations.
4. **Knowledge base enrichment module:** For knowledge base enrichment based on annotations in respect to the ontology. It communicates with Travel Module to update knowledge base with new instances and relations between them.

Following are details about key modules.

User Module

User Controller (Figure 12) accepts requests from the user (via *User registration form*) and fires appropriate actions. Actions (at the presentation layer) are directly connected to the business layer of the system represented by *User Manager* (UM). The UM has the following roles:

- Store and retrieve data about the user.
- Observe and track the activities of the user during his visit to the portal.

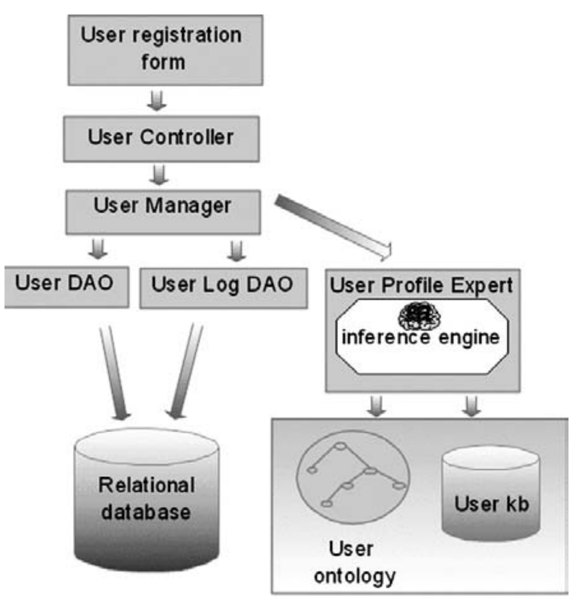
For manipulation with data stored in the database UM uses the *User DAO* (User Data Access Object). These data are user details that are not subject to frequent changes and are not important for determining the user profile: the username, password, first name, last name, address, birth date, phone and email.

For logging user activities during visiting the portal UM uses *User Log DAO*.

When reasoning over the available data about the user and determining user profile types UM use the User Profile Expert. The *User Profile Expert* is aware of the *User ontology* and also of the User profile knowledge base (*User kb*) that contains instances of classes and relations from the User ontology.

The data about users are collected in two ways:

Figure 12. User module components



1. Using User interface: the user is prompted to fill the forms to input data about him/herself. These data are: gender, birth date, social data (single, couple, family with kids, friends), the user's location, profession, education, languages, interests and activities (art, museums, sightseeing, sports, exploring new places during vacation, animals, eating out, nightlife, shopping, trying local food/experiencing local customs/habits, natural beauties, books), budget, visited destinations.
2. The system collects data about the user's interests and preferences while the user is reading about or searching for vacation packages using the portal. Each time the user clicks on some of the vacation package details, the system stores his/her action in the database, and analyse it later on.

Travel Module

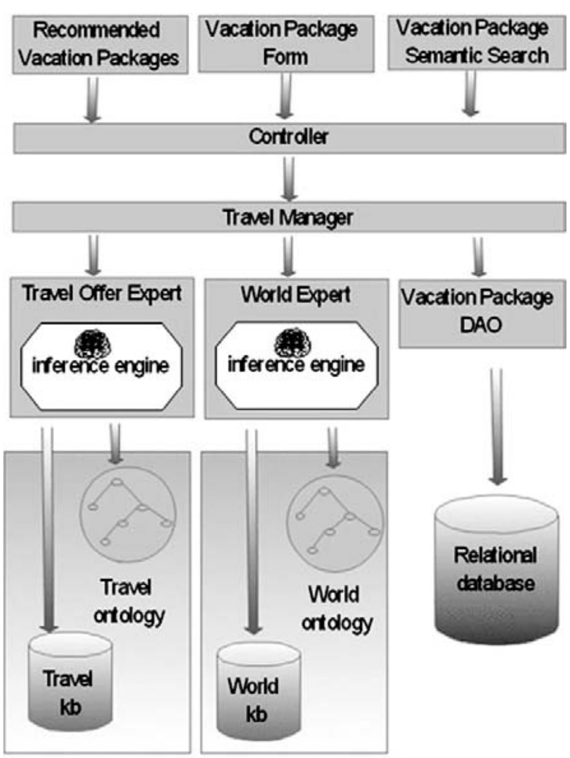
Travel Module generates and maintains data about vacation packages, destinations and related concepts. The User interface of Travel Module component comprises following forms (Figure 13):

1. *Recommended Vacation Packages* form: This form shows the list of vacation packages that the user has not explicitly searched for - system generates this list automatically based on the user profile.
2. *Vacation Packages Form*: This form is important for travel agents when updating vacation packages data.
3. *Vacation Package Semantic Search Form*: This form enables semantic search of vacation packages.

Each of the available forms communicates with the *Controller* who dispatches the requested actions to the *Travel Manager (TM)*. The Travel Manager is responsible for fetching, storing and updating the data related to vacation packages. It includes a mechanism for storing and retrieving data from the database using *Vacation Package DAO* (Data Access Object). The data stored in the database are those that are subject to frequent changes and are not important in the process of reasoning: start date, end date, prices (accommodation price, food service price, and transport price), benefits, discounts and documents that contain textual descriptions with details about the vacation packages. Some of these data are used in the second phase of retrieving a 'perfect' vacation package, when the role of the inference engine is not important. Retrieving a 'perfect' vacation package is performed in two steps:

1. Matching the user's wishes with certain destinations – the user profile is matched with certain types of destinations. To perform this TM uses the *Travel Offer Expert* (TOE) and the *World Expert* (WE) components.
2. The list of destinations retrieved in the first step is filtered using the constraints the user provided (for example, the start/end dates of the vacation). TM filters retrieved result using the *Vacation Package DAO.*

Figure 13. Travel Module Components



TOE and *WE* components include inference engines. These inference engines are aware of the ontologies and knowledge bases: TOE works with *Travel ontology* and a knowledge base (*Travel kb*) created based on this ontology. WE uses the *World ontology* and the knowledge base (*World kb*) created based on it.

After the initial knowledge base is deployed into Travel Guides application, its further update could be performed semi-automatically by *Knowledge base enrichment module (KBEM)* deployed inside Travel Guides. For example, when a new hotel is built, the knowledge base should be enriched with this information. This can be performed either by:

- Using the Travel Guides environment, where a tourist agent or administrator manually enters the name and other data about the new hotel (Figure 13).
- Performing annotation of the relevant content with regards to the Travel Guides ontology, semi-automatically (Figure 14).

Knowledge Base Enrichment Module (KBEM)

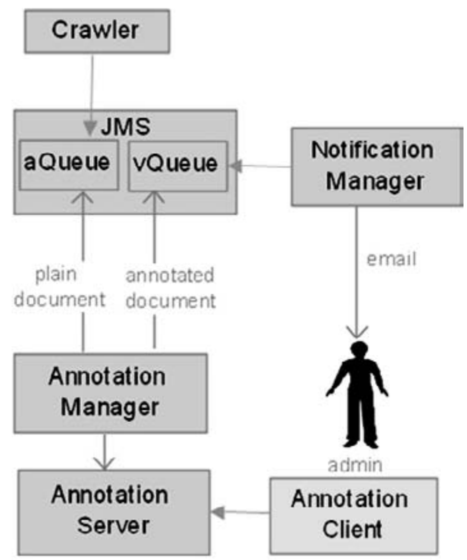
Semi-automatic annotation process starts with *Crawler* actions. Crawler searches the Internet and finds potentially interesting sites with details about destinations, hotels, beaches, new activities in a hotel, news about some destinations, popular events, etc. The result (HTML pages) is transformed into *.txt* format and redirected to JMS (Java Message Service) to wait in a queue for annotation process (*aQueue*). JMS API is a messaging standard that allows application components based on the Java 2 Platform, Enterprise Edition (J2EE) to create, send, receive, and read messages. It enables distributed communication that is loosely coupled, reliable, and asynchronous.

Annotation Manager consumes these plain documents and connects to the *Annotation Server* to perform process of annotation with regards to Travel Guides ontologies. After the annotation process is completed, the annotated documents are sent to JMS to wait in a queue for verification (*vQueue*). The *Notification Manager* consumes these massages and sends an e-mail to the administrator with the details about annotated documents (e.g., location of the annotated documents). The administrator starts *Annotation Interface* and performs the process of verification. The output of the annotation process is correctly annotated documents.

Retrieved annotations that refer to the new concepts/instances could be further used to enrich the KB and also for semantic search over the knowledge store that includes processed documents. Similar approach uses KIM Platform: they provide querying of the knowledge store that includes not only the knowledge base created w.r.t. ontologies, but also annotated documents (Popov et al., 2004).

Annotation of documents performed by KBEM would be simplified in case that verification step is skipped. The implementation of the system would also be simpler. In addition, there would not be a human influence, but the machine would do everything by itself. This would lead to many missed annotations, though. A machine cannot always notice some "minor" refinements as humans can. For example, if in the title "Maria's sand" the machine notices "Maria" and finds it in the list of female first names, it will annotate it as an instance of a class **Woman**. "Maria" can be an instance of a woman, but in this context it is a part of the name of a beach. These kinds of mistakes would happen frequently, and the machine would annotate them in wrong ways, if it does it automatically without any verification.

Figure 14. Knowledge base enrichment module inside the Travel Guides



An Example of Using Travel Guides

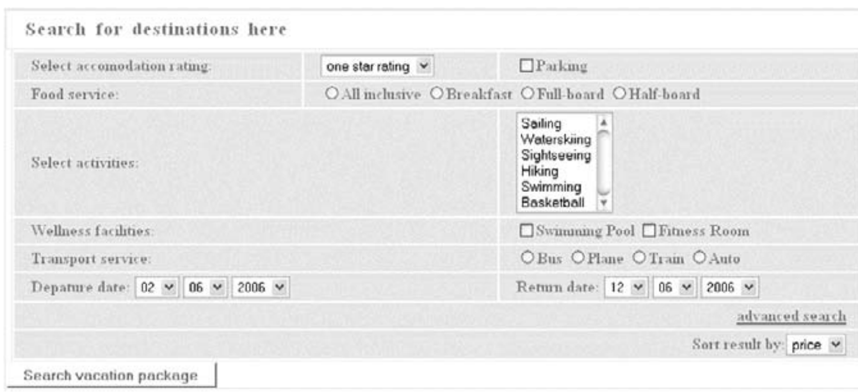
Travel Guides users are divided in 3 groups, each of which contributing to the knowledge base in its own way.

End users (i.e., tourists) visit this portal to search for useful information. They can feed the system with their personal data, locations, and interests, which then get analyzed by the system in order to create/update user profiles. Note that the system also uses logged data about each user's activities (mouse clicks) when updating the user's profile. User profile form for feeding the system with user personal information, activities and interests is depicted on Figure 15.

On the left hand side there is a section with results of system personalization. This section provides a list of potentially interested destinations for the tourist. The section is created based on the user profile analyse, which means that offered

Figure 15. The User profile form in Travel Guides

Figure 16. The Search form in Travel Guides



destinations should be matching user wishes, interests and activities. To explicitly search for a 'perfect' vacation package the user uses form shown on Figure 16.

Tourist agents create vacation packages and similar offers in tourist agencies. They feed and update the database with new vacation packages and also knowledge base with new information about destinations. To do this, they fill appropriate forms and save the filled-in information (Figure 17).

To successfully fill in this form and save the vacation package, the hotel has to be selected. If the hotel does not exist in the system, it has to be entered before creating the new vacation package. Figure 18 depicts a form for entering a new hotel into the KB.

Portal administrators mediate the knowledge base updates with destinations not covered by the tourist agencies connected to the portal. This process is very similar to the process conducted by tourist agents. The major difference is that this part of the knowledge base contains mostly static and permanent information about some geographical locations, such as countries, their capitals, mountains, rivers, seas, etc. all over the world. The idea is that tourist agents can use this part of the knowledge base as the basis for creating new vacation packages and other tourist offers.

Figure 17. The Vacation package form in Travel Guides

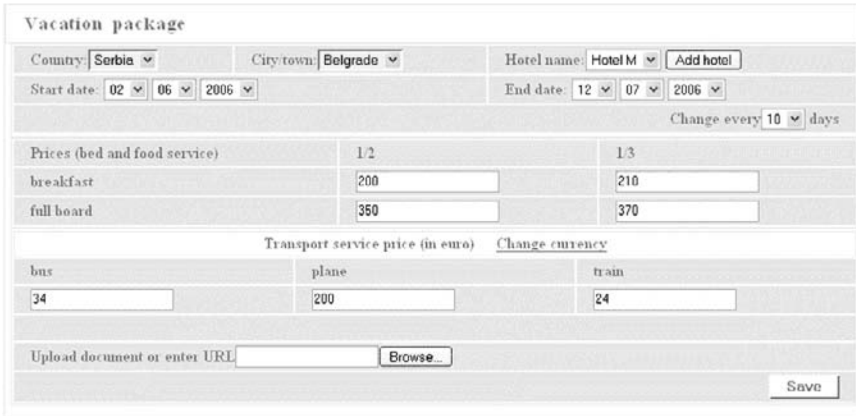
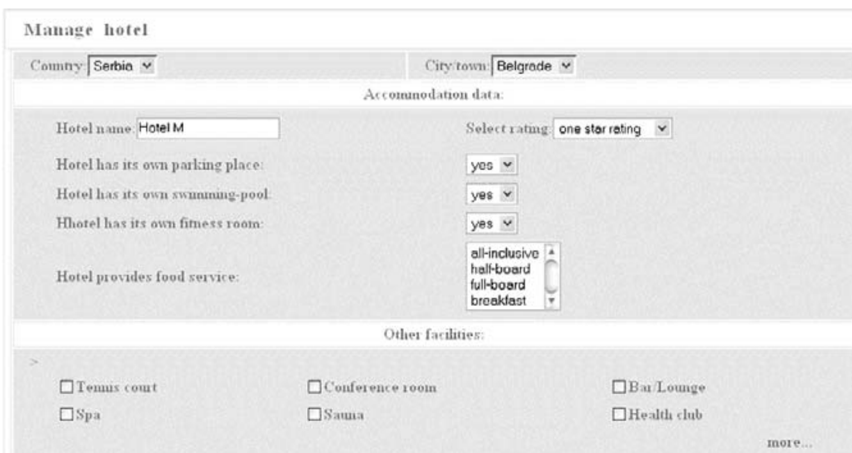


Figure 18. Entering a new hotel using the Travel Guides environment



CONCLUSION

Representing tourism-related data in a machine-readable form can help the integration of E-Tourism Information Services. If tourism sources would be centralized in a unique repository, the maintenance efforts would be significantly decreased. Integration of all E-Tourism sources would result in the possibility to search for tourist deals from one place – this would drastically reduce the time tourists spend while searching various tourism-related Web sites.

Built-in heuristics inside ontologies and use of a reasoner enable implying the user profile types for different tourists w.r.t. their activities and interests. Coupled with the destination types which are derived from the specific vacation package descriptions, user profiles can improve the process of searching for the perfect vacation package. Additionally, building a good quality user profiles provides personalization of dynamically created content.

The system's prototype described here includes a limited collection of vacation packages. The main precondition for its evaluation and usability would be feeding it with vacation packages from real tourist agencies.

As Travel Guides focus on integration of Information Services, such as information about destinations, hotels and the like, it would be worth exploring the possibility to integrate such a system with existing applications that offer Transactional Services, so that it can be possible to book and pay for recommended vacation packages after searching repository with available tourist offers covered by Travel Guides. In addition, there are opportunities to extend Travel Guides or to develop an independent module for integration of Communication Services, so that tourists can contribute to the system knowledge about the destinations and express their experience as well.

Finally, as the current version of Travel Guides ontology supports only representing hotel accommodation, there is a space for future improvements that include extending types of accommodations with hostels, private apartments for rent, and campgrounds.

FUTURE RESEARCH DIRECTIONS

Integrating semantic web technologies in traditional existing Web applications has a lot of space for improvement. The most popular way to perform this integration is by employing ontologies as they enable presenting data in machine readable form, reasoning and running intelligent agents, semantic Web services and semantic search. Each of these is partly applied in E-Tourism applications nowadays. However, current state of the art in this field is not mature enough to be used in industry, meaning that there is lots of space for different research topics, some of which could be implied from reading this chapter.

Reasoning over ontologies is very expensive due to the state of development of current inference engines. Development of better and faster reasoner is a precondition for using ontologies in large scale applications. At the moment, only few ontology-based systems exist in the area of tourism, among which Mondeca (www.mondeca.com) is applying the most of them to tourism in different regions in France. Their ontologies define the structure of data they are working with but the use of a reasoner is on the minimum level.

Emerging popularity of social web applications raises another interesting field of research, specifically information retrieval from user created content. Existing Natural Language Processing Tools are still weak to extract and retrieve meaningful answers based on the understanding of the query given in a form of natural language. For example, searching a social web application (e.g., a forum with reviews of different hotels), it would be hard to find 'the hotel in the posh area' using mainstream search engines as some of the posts might talk about luxury hotels, but not using 'posh' to describe them. Developing Natural Language Processing tools that could analyse text so that machines can understand it is a field with lots of research opportunities that would contribute not only to the E-Tourism applications, but to all applications on the Web.

Improving the process of automatic annotation and developing algorithms for training such a process would be another important contribution. Up to date, only Named Entities (e.g., organisations, persons, locations) are known to be automatically retrieved to the reasonable level of accuracy. Additionally, as current systems for performing annotation process usually require the knowledge and understanding of the underlying software such as GATE, research in this field can lead to developing more user-friendly interfaces to allow handling annotations and verifications without any special knowledge of the underlying software. The most natural way would be that similar to using tags in Web 2.0 applications, or any other simple way that requires no training for the user.

REFERENCES

Aichholzer, G., Spitzenberger, M., Winkler, R. (2003, April). Prisma Strategic Guideline 6: eTourism. Retrieved January 13, 2007, from: http://www.prisma-eu.net/deliverables/sg6tourism.pdf

Antoniou, G., Harmelen, F. V. (2004). Web Ontology Language: OWL. In Staab, S., Studer, R. (Eds.): Handbook on Ontologies. International Handbooks on Information Systems, Springer, pp. 67-92.

Bachlechner, D. (October, 2004), D10 v0.2 Ontology Collection in view of an E-Tourism Portal, E-Tourism Working Draft. Retrieved January 15, 2007 from: http://138.232.65.141/deri_at/research/projects/E-Tourism/2004/d10/v0.2/20041005/#Domain

Cardoso, J. (2006). Developing Dynamic Packaging Systems using Semantic Web Technologies. Transactions on Information Science and Applications. Vol. 3(4). 729-736.

Cunningham, H. (2002). GATE, a General Architecture for Text Engineering. Computers and the Humanities. 36 (2). 223–254.

Church, K., Patil, R. (1982). Coping with Syntactic Ambiguity or How to Put the Block in the Box. American Journal of Computational Linguistics, 8(3-4).

Dell'erba, M., Fodor, O. Hopken, W., Werthner, H. (2005). Exploiting Semantic Web technologies for harmonizing E-Markets. Information Technology & Tourism. 7(3-4). 201-219(19).

Djuric, D., Devedzic, V. & Gasevic, D. (2007). Adopting Software Engineering Trends in AI. *IEEE Intelligent Systems*. 22(1). 59-66.

Dogac, A., Kabak ,Y., Laleci, G., Sinir, S., Yildiz, A. Tumer, A. (2004). SATINE Project : Exploiting Web Services in the Travel Industry . eChallenges 2004 (e-2004), 27 - 29 October 2004, Vienna, Austria.

Fellbaum, C. (1998). WordNet - An Electronic Lexical Database. The MIT Press.

Guo, Y; Pan, Z; and Heflin, J. (2004). An Evaluation of Knowledge Base Systems for Large OWL Datasets. The Semantic Web – ISWC 2004: The Proceedings of the Third International Semantic Web Conference, Hiroshima, Japan, November 7-11, 2004. Springer Berlin/Heidelberg. 274-288.

Henriksson, R., (November, 2005), Semantic Web and E-Tourism, Helsinki University, Department of Computer Science. [Online]. Available: http://www.cs.helsinki.fi/u/glinskih/semanticweb/Semantic_Web_and_E-Tourism.pdf

Hepp, M., Siorpaes, K., Bachlechner, D. (2006). Towards the Semantic Web in E-Tourism: Can Annotation Do the Trick? In Proc. of 14th European Conf. on Information System (ECIS 2006), June 12–14, 2006, Gothenburg, Sweden.

Horridge, M., Knublauch, H., Rector, A., Stevens, R., Wroe, C. (2004). A Practical Guide To Building OWL Ontologies Using The Protege-OWL Plugin and CO-ODE Tools Edition 1.0. The University of Manchester, August 2004. [Online]. Available: http://protege.stanford.edu/publications/ontology_development/ontology101.html

Jentzsch, A. (April, 2005) XML Clearing House Report 12: Tourism Standards. Retrieved September 6, 2007, from http://www.xml-clearinghouse.de/reports/Tourism%20Standards.pdf

Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A., Goranov, M., (2003), Semantic Annotation, Indexing, and Retrieval, Lecture Notes in Computer Science, Springer-Verlag. Pages 484-499.

Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A. (2004). KIM - A Semantic Platform For Information Extraction and Retrieval. Journal of Natural Language Engineering, Cambridge University Press. 10 (3-4). 375-392.

Prantner, K. (2004). OnTour: The Ontology [Online]. Retrieved June 2, 2005, from http://E-Tourism.deri.at/ont/docu2004/OnTour%20-%20The%20Ontology.pdf/

Roman D., Keller, U., Lausen, H., Bruijn J. D., Lara, R., Stollberg, M., Polleres, A., Feier, C.,Bussler, C., Fensel, D. (2005). Web Service Modeling Ontology. Applied Ontology. 1(1): 77 - 106.

Siorpaes, K., Bachlechner, D. (2006). OnTour: Tourism Information Retrieval based on YARS. Demos and Posters of the 3rd European Semantic Web Conference (ESWC 2006), Budva, Montenegro, 11th – 14th June, 2006.

Smith, C. F., Alesso, H. P. (2005). Developing Semantic Web Services. A K Peters, Ltd.

Stollberg, M., Zhdanova, A.V., Fensel, D. (2004). “h-TechSight - A Next Generation Knowledge Management Platform”, Journal of Information and Knowledge Management, 3 (1), World Scientific Publishing, 45-66.

Terziev, I., Kiryakov, A., Manov, D. (2005). D1.8.1 Base upper-level ontology (BULO) Guidance, SEKT. Retrieved January, 15th, 2007 from: http://www.deri.at/fileadmin/documents/deliverables/Sekt/sekt-d-1-8-1-Base_upper-level_ontology__BULO__Guidance.pdf

ADDITIONAL READING

Bennett, J. (2006, May 25). The Semantic Web is upon us, says Berners-Lee. Silicon.com research panel: WebWatch. Retrieved January 3, 2007, from: http://networks.silicon.com/Webwatch/0,39024667,39159122,00.htm

Bussler, C. (2003). The Role of Semantic Web Technology in Enterprise Application Integration. IEEE Data Engineering Bulletin. Vol. 26, No. 4, pp. 62-68.

Cardoso, J. (2004). Semantic Web Processes and Ontologies for the Travel Industry. AIS SIGSEMIS Bulletin. Vol. 1, No. 3, pp. 25-28.

Cardoso, J. (2006). Developing An Owl Ontology For e-Tourism. In Cardoso, J. & Sheth, P. A. (Eds.). Semantic Web Services, Processes and Applications (pp. 247-282), Springer.

Davidson, C., Voss, P. (2002). Knowledge Management. Auckland: Tandem.

Davies, J., Weeks, R., Krohn. U. (2003a). QuizRDF: Search Technology for the Semantic Web. Towards the Semantic Web: Ontology-Driven Knowledge Management, pp. 133-43.

Davies, J., Duke, A., Stonkus, A. (2003b). OntoShare: Evolving Ontologies in a Knowledge Sharing System. Towards the Semantic Web: Ontology-Driven Knowledge Management, pp. 161-177.

Djuric, D., Devedžić, V.,Gašević, D. (2007). Adopting Software Engineering Trends in AI. IEEE Intelligent Systems. 22(1). 59-66.

Dzbor, M., Domingue, J., Motta, E. (2003). Magpie - Towards a Semantic Web Browser, In Proc. of the 2nd International Conference (ISWC 2003), pp. 690-705. Florida, USA.

Edwards, S. J., Blythe, P. T., Scott, S., Weihong-Guo, A. (2006). Tourist Information Delivered Through Mobile Devices: Findings from the Image. Information Technology & Tourism. 8 (1). 31-46(16).

Engels R., Lech, T. (2003). Generating Ontologies for the Semantic Web: OntoBuilder. Towards the Semantic Web: Ontology-Driven Knowledge Management, pp. 91-115.

E-Tourism Working Group (2004). Ontology Collection in view of an E-Tourism Portal. October, 2004. Retrieved January 13, 2007, from: http://138.232.65.141/deri_at/research/projects/e-tourism/2004/d10/v0.2/20041005/

Fensel, D., Angele, J., Erdmann, M., Schnurr, H., Staab, S., Studer, R., Witt, A. (1999). On2broker: Semantic-based access to information sources at the WWW. In Proc. of WebNet, pp. 366-371.

Fluit, C., Horst, H., van der Meer, J., Sabou, M., Mika, P. (2003). Spectacle. Towards the Semantic Web: Ontology-Driven Knowledge Management, pp. 145-159.

Hepp, M. (2006). Semantic Web and semantic Web services: father and son or indivisible twins? Internet Computing, IEEE. 10 (2). 85- 88.

Heung, V.C.S. (2003). Internet usage by international travellers: reasons and barriers. International Journal of Contemporary Hospitality Management, 15 (7), 370-378.

Hi-Touch Working Group (2003). Semantic Web methodologies and tools for intraEuropean sustainable tourism [Online]. Retrieved April 6, 2004, from http://www.mondeca.com/articleJITT-hitouch-legrand.pdf/

Kanellopoulos, D., Panagopoulos, A., Psillakis, Z. (2004). Multimedia applications in Tourism: The case of travel plans. Tourism Today. No. 4, pp. 146-156.

Kanellopoulos, D., Panagopoulos, A. (2005). Exploiting tourism destinations' knowledge in a RDF-based P2P network, Hypertext 2005, 1st International Workshop WS4 – Peer to Peer and Service Oriented Hypermedia: Techniques and Systems, ACM Press.

Kanellopoulos, D. (2006). The advent of Semantic web in Tourism Information Systems. *Tourismos: an international multidisciplinary journal of* tourism. 1(2), pp. 75-91.

Kanellopoulos, D. & Kotsiantis, S. (2006). Towards Intelligent Wireless Web Services for Tourism. IJCSNS International Journal of Computer Science and Network Security. 6 (7). 83-90.

Kanellopoulos, D.,Kotsiantis, S., Pintelas, P. (2006), Intelligent Knowledge Management for the Travel Domain ,GESTS International Transactions on Computer Science and Engineering. 30(1). 95-106.

Kiryakov, A. (2006). *OWLIM: balancing between scalable repository and light-weight reasoner.* Presented at the Developer's Track of WWW2006, Edinburgh, Scotland, UK, 23-26 May, 2006.

Maedche, A., Staab S., Stojanovic, N., Studer, R., Sure, Y. (2001). SEmantic PortAL -

The SEAL approach. In D. Fensel, J. Hendler, H. Lieberman, W. Wahlster (Eds.) In Creating the Semantic Web. Boston: MIT Press, MA, Cambridge.

McIlraith, S.A., Son, T.C., & Zeng, H. (2001). Semantic Web Services. IEEE Intelligent Systems 16(2), 46-53.

Missikoff, M., Werthner, H. Höpken, W., Dell'Ebra, M., Fodor, O. Formica, A., Francesco, T. (2003) HARMONISE: Towards Interoperability in the Tourism Domain. In Proc. ENTER 2003, pp. 58-66, Helsinki: Springer.

Passin, B., T.(2004). *Explorer's Guide to the Semantic Web*. Manning Publications Co., Greenwich.

Sakkopoulos, E., Kanellopoulos, D., Tsakalidis, A. (2006). Semantic mining and web service discovery techniques for media resources management. International Journal of Metadata, Semantics and Ontologies. Vol. 1, No. 1, pp. 66-75.

Singh, I., Stearns, B., Johnson, M. and the Enterprise Team (2002): Designing Enterprise Applications with the J2EE Platform, Second Edition. Prentice Hall. pp. 348. Online: http://java.sun.com/blueprints/guidelines/designing_enterprise_applications_2e/app-arch/app-arch2.html

Shadbolt, N., Berners-Lee T., Hall, W. (2006). The Semantic Web Revisited. IEEE Intelligent Systems. 21(3). 96-101.

Stamboulis, Y. Skayannis P. (2003). Innovation Strategies and Technology for Experience-Based Tourism. Tourism Management. Vol. 24, pp. 35-43.

Stojanovic, LJ., Stojanovic N.,Volz, R. (2002). Migrating data-intensive Web sites into the Semantic Web. Proceedings of the 2002 ACM symposium on Applied computing, Madrid, Spain, ACM Press. 1100-1107.

Sycara, K., Klusch, M., Widoff, S., Lu, J. (1999) Dynamic service matchmaking among agents in open information environments. ACM SIGMOD Record. Vol. 28(1), pp. 47-53.

World Tourism Organization, 2001, Thesaurus on Tourism & Leisure Activities: http://pub.world-tourism.org:81/epages/Store.sf/?ObjectPath=/Shops/Infoshop/Products/1218/SubProducts/1218-1

WTO (2002). Thesaurus on Tourism & Leisure Activities of the World Tourism Organization [Online]. Retrieved May 12, 2004, from http://www.world-tourism.org/aciduis ciduisi bla facillum nulla feuguer adignit amet

This work was previously published in The Semantic Web for Knowledge and Data Management: Technologies and Practices, edited by Z. Ma and H. Wang, pp. 243-265, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 4.10
E-Tourism Image:
The Relevance of Networking for Web Sites Destination Marketing

Lluís Prats-Planagumà
Universitat de Girona, Spain

Raquel Camprubí
Universitat de Girona, Spain

ABSTRACT

The competitiveness of tourism destinations is a relevant issue for tourism studies, moreso, is a key element on the daily basis of tourism destinations. In this sense, the management of tourism destinations is essential to maintain competitive advantages. In this chapter tourism destination is considered as a relational network, where interaction and cooperation is needed among tourism agents, to achieve major levels of competitive advantage and a more effective destination management system. In addition, the perceptions of tourists are obtained from two main sources. The first one is the social construction of a tourism destination previous to the visit and the second one is obtained from the interaction between tourists and tourism destination agents during the visit. In this sense, the management of tourism destination to emit a homogenous and collective image is a factor that can reduce the gap if dissatisfaction from the previous and real tourist perception. The authors specifically discuss the importance of a common agreement of tourism agents on virtual tourism images projected through official Web sites, considering that the literature focused mainly in how to promote and sell destinations trough Internet but not in terms of exploiting a destination joint image. Finally, in order to analyze the integration of a tourism product and determine their consequences in tourism promotion an empirical research has been done, using the case of Girona's province. The main findings determine that, although interactions among tourism agents can improve destination competitiveness, little cooperation in tourism promotion on Web sites is achieved, as well as a few uses of technological resources in the Web sites to facilitate to tourists a better understanding of tourism resources in the area.

DOI: 10.4018/978-1-60566-134-6.ch009

INTRODUCTION

Each tourism destination can be considered a market in itself. At these destinations tourism suppliers (i.e., accommodations, restaurants, museums, and tourism offices, among others) interact simultaneously with the tourists who consume these products or services. For that reason a market approach is more appropriate than a supply or a demand one.

A tourism destination is the geographical area where a set of tourism agents interact and intervene in tourism activities. These interactions, from a supply point of view, help develop a relational network at the destination. A relational network is the set of economic and personal relationships established among a number of agents who share goals, cooperation systems, knowledge, reputation, and image, among other elements, in common. These elements help the destination network generate collective learning and knowledge, and consequently, achieve greater levels of competitiveness than individual agents would obtain.

In addition, from the demand point of view, these interactions within the destination help minimize the existing gap between perceived and real images. All tourists have a socially constructed image of a destination (Urry, 1990; Galí & Donaire, 2005; Larsen & George, 2006), which conditions their decision-making, and it is important for the tourism agents involved in the network to control the image of a destination.

This control has two simultaneous benefits. The first one, related to the tourism demand, is the potential to influence tourist decision-making. The second one is related to the tourism supply chain: the competitive advantage brought to tourism destinations by projecting a real image.

Internet is a very important channel that helps tourism agents to achieve these two benefits derived from this control in three aspects. First, a number of authors assume the relevance of Internet as a tourism destination image generator (Baloglu & Pekan, 2006; Choi, Lehto & Morrison, 2007; Hashim, Murphy & Muhammad Hashim, 2007); although "research on the Internet as an image formation agent is still at an infancy stage" (Choi, Lehto & Morrison, 2007, p. 118). Second, Internet brings a great number of opportunities to tourism image formation, contributing to destination imagery formation to consumers (Hashim, Murphy & Muhamd Hashim, 2007) and giving to tourism destination an opportunity to improve destination marketing through the use of "Internet's unique features, such as geographical interactivity with audience, low-cost accessibility, world-wide, hyperlinks with other travel suppliers and design flexibility, to attract more tourists and better position their state in the intense competition for visitors" (Lee, Cai & O'Leary, 2006, p. 816). Third, Internet and destination websites, in particular, act as an information tool for tourists, being an influencing element in their decision-making. This article will discuss the attainment of these two benefits and the relevance of Internet in them using the tourism image and social network theories to clarify how supply and demand interact in a tourism destination. A conceptual model will be proposed as part of a theoretical market approach to tourism destinations, which integrates supply and demand, explains interactions between them and highlights the relevance of this scope of analysis to better understand the dynamics of a tourism destination and the possibility of improving its competitive advantage. In addition, the article demonstrates the necessity of using this integrated approach for planning and managing a tourism destination to improve its competitiveness and highlight this theoretical view.

One of the elements that can be planned with an integrated approach, mainly because it helps to establish scale economies in terms of promotion, it is the promotional website content of a destination. Usually different agents take part in this promotion, Destination Marketing Organization's (DMO) local governments, private companies or associations creating different sources of information. Apart from these possible scale economies this will also help in terms of unification of the destination image. If this image is homogeneous gives also an extra value of competitive advantage preventing incoordination.

Finally, a case study is conducted in order to analyze network configuration trough promotional websites in Internet and determine if the tourism product of a destination is integrated and promoted globally or, on the contrary, each tourism agent acts independently. Results show the existence of a reduced network, that means a little integration of the tourism products in the province of Girona.

This article is organized in five main sections. The first explains the process of tourism image formation and how the tourism agents that intervene in this process affect the image that tourists have of tourism destinations, from the perspective of the social construction of tourism images. The second focuses on the network configuration of the destination, taking into account the tourism agents who take part in the tourism system and how networking can generate competitive advantages. The third part presents a theoretical model of an integrated market approach to tourism destinations. The forth section presents a case study, which analyze the network configuration of tourism products in the province of Girona. Finally, the conclusions based on the theoretical model and the case study is drawn, the model's limitations are considered and proposals for future research are made.

HOW TOURISTS PERCEIVE DESTINATIONS

Social Construction of Tourism Destination Images

Images have been used in a number of contexts and disciplines: psychology perceives the image as a visual representation; thought behavior geography emphasizes the association of impressions, knowledge, emotions, values and beliefs; and marketing focus on the relationship between image and behavior of consumers (Jenkins, 1999). The majority of academics from the 1970s to the present day agree that tourism image is "the sum of beliefs, ideas, and impressions that a person has of a destination" (Crompton, 1979).

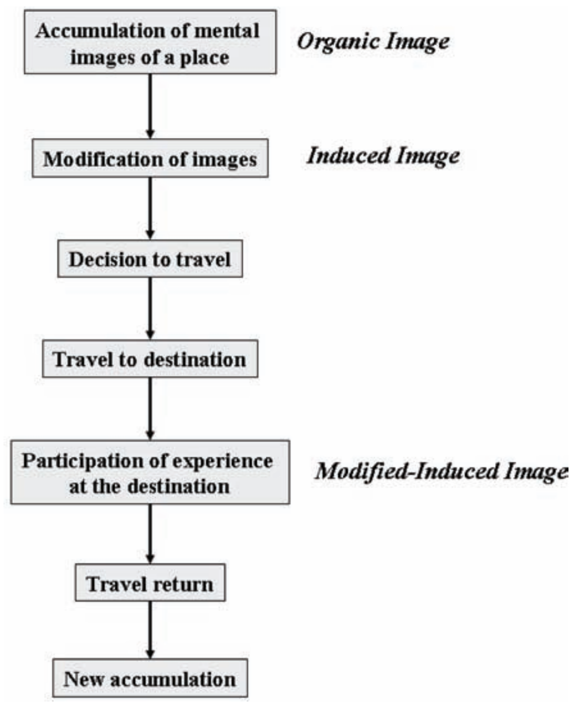
Gunn (1988), in her main academic study mentions "all of us have images of destinations, whether or not we have traveled to them. These images may be sharp or vague, factual or whimsical, but in all cases they are indicative of likes and dislikes." (p. 23) This means that all places have an image, which has not appeared out of nowhere, but they have consciously or unconsciously been created by "somebody". In this sense, one needs to think about how a tourism destination image is constructed.

From the realization of a task and its reiterated repetition by people in a society, this task ends up being institutionalized by this society. The acceptance of this task as habitual makes it "settle" in this society and form part of its traditions, so in that sense, the reality of this society has been constructed collectively (Berger & Luckmann, 1968). Using this approach to tourism image, it could be contemplated that a tourism image is constructed socially in the same way as a task is accepted as a normal way to do something in a society.

Tourism images are full of visual elements and signs that evoke socially constructed images (Urry, 1990), for example a couple of lovers in Paris suggest romantic Paris. In this sense, the image construct implies some overriding impression or stereotype (Mazanec & Schweiger, 1981). However, this tourism image does not always reflect the reality, because "the tourism image is, at the same time, a subjective construction (that varies from person to person) and a social construction, based on the idea of collective imagination" (Galí & Donaire, 2005, p. 778).

Variations in tourism image are complex if one considers how these images are formed, a little bit at a time. As Gallarza, Gil and Calderón (2002) have exposed "image is not static, but changes depending essentially in two variables: time and space". (p. 72) The influence of time on image is demonstrated in a number of studies on tourism image (Gartner, 1986; Gartner & Hunt, 1987;

Figure 1. Seven-stage model of a tourism experience (Source: Gunn, 1972)



Chon, 1991; Selby & Morgan, 1996), especially if one considers its formation as a process (Gunn, 1972). At the same time, the space variable also influences the image of a tourism destination. Some studies in this field show that the distance between potential tourists and the tourism destination affects the perceived image of the place (Miossec, 1977; Talisman-Kosuta, 1989; Gallarza, Gil & Calderón, 2002). Considering the dynamic nature of tourism, image is useful if the effect of marketing actions on time and space variables (Gallarza Gil & Calderón, 2002; Talisman-Kosuta, 1989) is taken into account. In this manner the periodic evaluation of tourism image is relevant (Talisman-Kosuta, 1989).

How is the Image of a Tourism Destination Formed?

Accepting as valid the fact that tourism image is socially constructed (Urry, 1990; Larsen & George, 2004; Galí & Donaire, 2005), some studies point to the existence of factors or components that form part of every tourism image and influence its formation process (Baloglu & McCleary, 1999; Gallarza, Gil & Calderón, 2002; Beerli & Martín, 2004). In this sense, Gartner (1993) mentions that some authors have systematized the elements that influence the process of tourism image formation in different conceptual models. At the same time, it is possible to find a number of authors who focus on the existence of a formation process of the tourism image, which is made up of different stages that contribute to how a tourism image is formed (i.e., Gunn, 1972; Govers & Go, 2004).

One of the most important models that show how a tourism image is formed is the seven-stage process of tourism experience, which has been developed by Gunn (1972). This model shows that images held by potential visitors, nonvisitors, and returned visitors differ (Gunn, 1972). (Figure 1)

At the first stage, potential tourists assimilate general information, such as, newspapers, television documentaries, books, and school lessons. This process generates an organic image of the destination; this is because the mere mention of these places evokes images, which are not necessarily tourism images.

The second stage implies a modification of perceived images based on consulting tourism information (i.e., tourism posters, guides, articles in specialized reviews, etc.). These changes in perceived images are influenced by induced images, which are the result of a conscious effort to develop, promote, and advertise a destination.

When the potential visitor has a perceived image based on the organic and induced images of the place, then they are prepared to make a decision. Other factors such as previous experience or the money available are also taken into consideration.

Travel to the destination may condition the image that a visitor has, but the key factor of a new change in a visitor's perceived image is their personal experience at the destination, as well as their participation in different activities, such as, visiting museums or the use of tourism services such as accommodation. At this stage, visitors have a modified-induced image, which is the result of the balance between the perceived image before visiting the destination and the perceived image after the visit.

Returning home after traveling, visitors evaluate and make reflections about their experience and discuss it with other travelers. At the final stage, tourists accumulate new information if one considers that this is a circular process. In this sense, it is widely recognized in academic literature that experienced tourists will become a "source of information" for other potential visitors, which will be based on their experience at the destination (Balogru & McCleary, 1999; Beerli & Martin, 2004).

As is noted in this model, the creation and modification of tourism images are constant and demonstrate the dynamism of the tourism image. The space variable shows these phenomena through the contact of visitors with tourism destinations.

Image Management as a Competitive Advantage

Academic literature recognizes the need to manage tourism image, as it is one of the most important factors that influences the decision-making process of tourists that choose a destination to spend their holidays (Gartner, 1993; Govers & Go, 2004). Gunn (1972), in her model explains that tourism images are conditioned by the actions of a number of agents that influence the creation of tourism images. Although it is agreed that the tourism image is socially constructed, agents intervene in this process emitting images, which end up being consolidated and accepted as valid in a specific society.

According to Gartner's (1993) agents' classification, there are four types of agents. The first, Overt Induced is a kind of agent who promotes the creation of a specific tourism image of the tourism destination in a conscientious way, to influence a tourist's process of decision-making. Gartner (1993), makes a distinction between these agents, who are of two types. On one hand, Overt Induced I are "the promoters of the destination [that] construct an image of the salient attributes of the destination in the minds of the targeted audience" (Gartner, 1993, p. 197) with the traditional forms of advertising (i.e., television, radio, brochures, etc.). In this case, one could also include tourism businesses of the destinations, such as, accommodation, restaurants, activities, and so on. On the other hand, Overt Induced II are usually "tour-operators, wholesalers or organizations who have a vested interest in the travel decision process, but which are not directly associated with a particular destination area" (Gartner, 1993, p. 199). As Gartner (1993), mentions "destination area promoters do have some control over the images projected through tour operator" (p. 199) because if the tourism image does not conform to the reality of the destination it could create dissatisfactions to both locals and visitors (Govers & Go, 2004).

The second, Covert Induced are agents that apparently emit a tourism image that is not induced. In this case the author also defines two types of Covert Induced agents. The first is called Covert Induced I, who is related to a recognizable spokesperson who recommends a destination to support a higher level of credibility of tourism destination advertising. The second is Covert Induced II, this category corresponds to people or organizations who write articles, reports or stories about a particular place. Often this published information is a result of a familiarization tour for travel writers or special interest media groups. These actions increase credibility and allow destination promoters to project a specific image.

The third kind of agents is called Autonomous. These agents are people or organizations who produce reports, documentaries, movies, and news articles independently without the specific aim of creating a tourism image of a place.

The last group of agents that Gartner (1993), identifies is called Organic, and is related to information and opinions about a place that a person receives from other people, from their previous experience in this place, Unsolicited Organic corresponds to people who give information about a destination where they have been, without having been specifically asked by the other interlocutor, for example, when this is a topic of conversation with friends in colleges. The existence of Solicited Organic agents implies that individuals actively search for information about a destination and somebody informs them using their own experience. Friends or relatives usually constitute these kind of agents, who have a high level of credibility and are an extremely important part of the destination selection process.

When people visit a destination they become an Organic (pure) agent, having the capability to give information in a solicited or unsolicited way.

This model shows that the task of Overt Induced agents, in this context, is undeniable, especially if one considers the sustainable competitive advantage of the destination. Sustainable competitive advantage is generally based on either core competences or unique resources that are superior to those possessed by competitors and are difficult to imitate (Johnson & Scholes, 1999; Aaker, 2001). Govers and Go (2004, p. 169), established that superior resources for a tourism destination "are generally to be found in either its unique and natural environment (climate, wildlife or landscape) or its cultural heritage" and also mention that "competitive advantage might be created through core competences, such as, the host community's existing unique capabilities in attracting visitors from outside." (i.e., destination's ability to stage world class events, festivals or exploit its folklore and prevailing traditions).

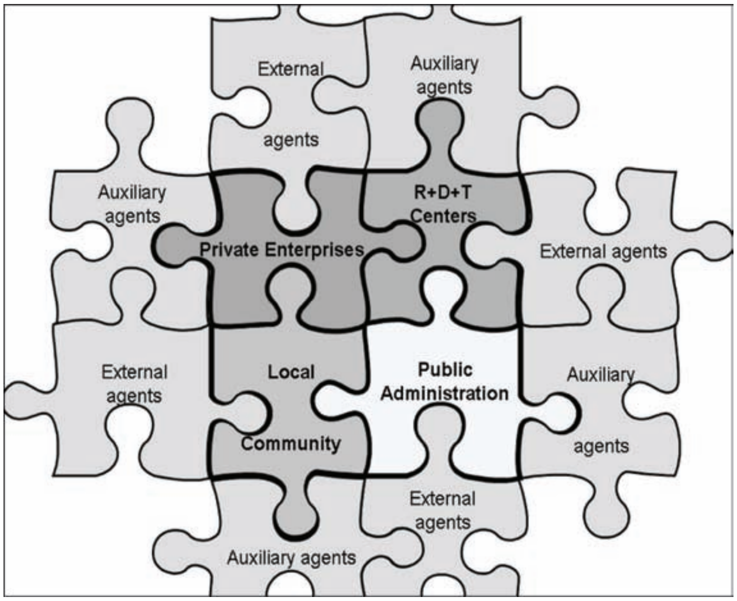
Following these considerations, the management of tourism image is viewed as a management tool (Ritchie, 1993). Govers and Go (2004), propose that it is necessary "to formulate a plan for projecting the 'right' image" (p. 170) as one of the essential parts of tourism development strategy. Gartner (1993), mentions the importance of considering the "image mix", as a continuum of factors that have to be taken into account to decide which agents will intervene in the formation of tourism image, as well as, the amount of money budgeted for image development, characteristics of target market, and demographic characteristics or timing. This task is obviously attributed to promoters of the destination who can select the right mix of image formation agents to maximize their scarce resources (Gartner, 1993).

THE NETWORK CONFIGURATION AND A TOURISM DESTINATION

Tourism Destination Agents

Tourism agents are an essential part of the system and of any destination, therefore they have been

Figure 2. Tourism destination agents



identified. Some definitions that are more applicable to industrial destinations, like innovation systems, clusters, milieux innovateurs, or industrial districts, always consider three main types of agents. Moreover, it is considered that tourism specificity needs another main agent who helps to define the situation of the system and therefore needs to complete it.

Two other types of agents support these main agents at all times and are also necessary to maintain the main set of agents stable. Figure 2 illustrates these agents in detail.

Following this scheme, those organizations that take part directly in generating the tourism experience are private companies. These include basic tourism companies, and also others whose main activity is related to tourism.

Public administrations are those organisms of governmental function that take part in the tourism processes and whose intervention can generate new legislations, give incentives for research, planning, and others.

Research, development, and training centers (R+D+T) are the essential elements capable of generating specialized training and/or research in the scope of tourism, such as, universities, research institutes, or consultants.

These three main agents appear in the academic literature on innovation systems (Lundvall, 1992; Nelson, 1993), as well as, in clusters and industrial districts. In addition, the tourism scheme proposed by Gunn (1997), and later adapted by the OMT (1999) is taken, and the tourism industry as a functional tourism system (Prats & Balagué, 2005) is conceptualized. They demonstrate that the local community also has an essential role in the development of tourism activity, and consequently of the system. The local community is defined as the inhabitants of a territory. These people are individuals or organizations without economic aims, such as, NGO's, civic organizations, or others. The relevance of local community in tourism is emphasized, seeing that civic movements have been able to modify important decisions in city-planning, ecological subjects, or others, restraining or impelling tourism.

After describing the basic elements, the tourism auxiliary agents can be defined as those agents who do not have activities directly related to the tourism industry, but who support the main agents. Looking at the economic theories, the auxiliary agents are some of the receivers of the multiplying effect (McIntosh, Goeldner & Ritchie, 2000). And the external agents are those tourism agents who are part of other destinations, but who interact with one or more internal agents.

The set of agents in a tourism destination is basically located in the same geographical territory. However, a territory by itself does not have enough conditions for their collective coordination, and also the proximity does not generate synergies by itself, but it can contribute to their effectiveness with other dimensions shared between the agents (Zimmermann, 2001). A good example of an agent who belongs to a distant destination could be a specialized tour operator who commercializes destinations, which are geographically distant but relationally close.

Relational Networks

The use of relational networks in the analysis of a company's competitive advantage can be related to several approaches in the fields of economics or sociology, among others (Sorensen, 2004). Therefore, in the most static frame, this analysis has appeared within the network of individual companies who have useful and important connections with other companies, becoming more than just a unit inside an atomized market (Håkansson & Snehota, 1995). In this sense, these companies must be analyzed considering their relationships with other companies outside the network and also the existing relationships among other companies within the network (Holmen, Pedersen & Torvatn, 2005).

In relation to the most dynamic frame, it is observed that it was contemplated not to see the innovation process as a linear and consecutive process, meaning that the result of the initial stage brings up the following one and so on. Innovation is considered an intensive activity in both knowledge (Sundbo, 1998; Roberts, 2001) and learning. It is also totally accepted as a key element in the innovation process. Thus, innovation also arises and takes place through the interactions between companies (Sorensen, 2004), and between these companies and other relevant actors who are important for their activity (Prats & Guia, 2005). These ties must be understood as intense flows of knowledge and, therefore, essential for innovation, and also for competitiveness.

However, Sorensen (2004) presents a definition that considers networks as the set of conscious and accepted business relationships, whether formal or informal, with transmission of resources, immaterial or material, within the company's scope. In any case, it is useful to adopt the perspective of social network analysis, which studies specific relationships between a defined series of elements, like people, groups, organizations, countries or events, among others (Molina, 2001). It is necessary to consider that social network analysis is based on relationships and not on the attributes of elements. Then, a social network can be defined as the group of people, organizations or other social entities connected by a set of significant relations (Wellman, 1997).

Granovetter (1985) and Hite (2003), affirm that the existing relationships within social networks influence economic actions, and Hite (2003), distinguishes seven different types of ties that can take place inside a social network: the main three are business ties, personal ties, and hollow ties, and the other four types are formed as a result of the relationship between the main types.

Porter (1990), with his five forces model and his later approach to clusters, universalized the necessity to maintain the business or commercial ties that had been previously valued by Becattini (1979), and other authors. Other theories such as the industrial districts theory show that personal relationships have to be considered as a value that contributes to empower the agents' ties making them more efficient and trustworthy (Becattini, 1979). Hollow ties appeared only recently in network theories and have become very common, because they represent all those ties that you accept with the mediation of a third person, so your trust in the relationship is not with the agent to whom you are related, but with the agent who did the mediation (Prats, Camprubí & Comas, 2005).

It seems evident then, that a tourism destination can be defined as a relational network. In every single destination exist relationships among its agents (Prats, Guia & Molina, 2007), considering the specificities of the tourism product, and the existence of the different types of ties able to generate an active and beneficial set of agents and relationships. As an example, a tourism package is integrated by different items that are provided by various companies, which are linked though some kind of relationship. However, tourists do not distinguish that these items belong to different companies, although they need a perfect integration of them, in order for the tourism products to be successful. Another example can be the usual sectorial associations, where hotels, travel agencies or other tourism agents are associated in order to gain power in front of the suppliers or in front of the administration.

Networking as a Generator of Competitive Advantage

If different agents interact among themselves, it can be argued that these interactions often allow the agents to have joint benefits from infrastructures, common engineering, and transfer of tacit knowledge. It also makes productive combinations and interactions more difficult to carry out in atomization or individual isolation.

Even with the continuous growth of the on-line travel expenditure, the academic literature focused mainly in how to effectively build and evaluate hospitality and tourism websites (Han & Mills, 2006). Moreover the analysis is focused on individual perspectives comparing different websites as example, but there are not published results related to collective image construction trough websites. “Because destination images influence tourist behavior, a destination must be favorably differentiated from its competition and positively positioned in the minds of consumers” (Hudson & Ritchie, 2006, p. 388), and this can be achieved also by the common internet promotion.

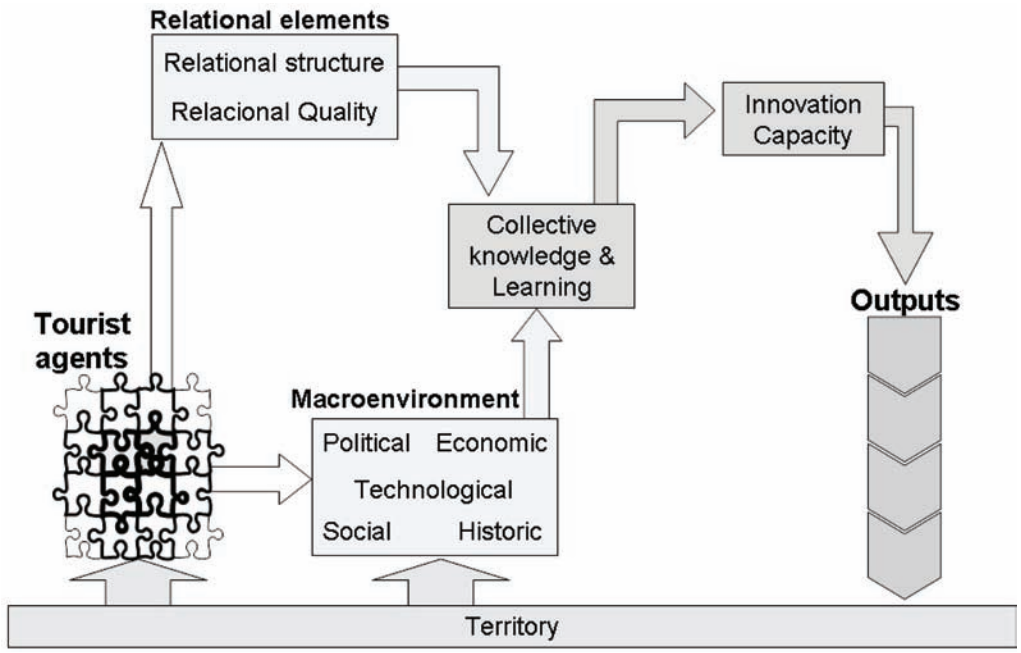
Belonging to a destination or relational network involves interacting with other members, which is usually transformed into routines of the organization. This is what Rallet and Torre (2004) call the belonging logic. This logic and interaction will be easier a priori if there is a common knowledge; this is called logic of similarity.

The interaction of these agents generates a number of factors that determine if a destination or local innovation system is successful or not in all scopes. A first and fundamental factor is the internal and external relationships that take place in the system. These relationships can be very different and they have been summarized into two characteristic groups.

On the one hand, depending on the relational structure that is adopted in a system, the degree of success will vary. In this factor the key element is the degree of connectivity that is obtained, understanding that the better the connectivity between the agents is, the closer it will be to “the ideal” system. It is understood that good internal connectivity will contribute to a more fluid circulation of knowledge between the agents, and this will increase the trust among them. But at the same time, an excess of internal connectivity can make the trust on external agents decrease to such an extent that they are considered intruders (Zimmerman, 2001). The lack of trust between external agents could have serious consequences in the new knowledge generation, because the closure of relational networks in itself could limit information flows that come from outside, blocking the possibility of generating new knowledge and collective learning (Lazerson & Lorenzoni, 1999).

On the other hand, however, it must also be observed that the quality of relationships within

Figure 3. Tourism local innovation system model (Source: Prats & Guia, 2005; Prats, Guia and Molina, 2007)



a system such as this, affects its success. The key element in this factor is trust, as a greater trust between the elements of a system will transmit more relevant information, and greater benefits for the whole destination will increase.

Another determining factor is the macro-environment, which is divided into five elements: (1) political, such as, decisions or political elements that affect the system; (2) economic, for instance economic situations that affect the system; (3) technological, which has two levels: (a) the hard level such as the automation level, and (b) the soft level such as the training level of the population; (4) social, this contributes to the system culture, for example the degree of associationism or the cultural level; and finally, (5) historical macro-environment, which gives perspective and historical experience, such as, political periods or natural disasters.

Using the agents' interactions and macro-environment variables, tourism destinations should be able to generate essential collective knowledge and learning for the evolution of the system. The main purpose of this collective knowledge and learning is being able to generate a constant innovation capacity that will bring dynamism to the system as shown in Figure 3.

This innovation capacity allows the system to obtain four successive outputs, which can be observed in Figure 4. Each stage must be achieved to obtain the desired results. If an "ideal" configuration of the system is obtained, the four outputs will also be obtained, and this will revert again to the tourism destination.

The first unquestionable output of the innovation capacity is innovations in any of their modalities. In the opinions of Prats and Guia (2005) the innovation must allow the system to generate a competitive advantage, allowing the destination to satisfy the needs of the tourists better than the competitor's destinations.

The competitive advantage, consequently,

Figure 4. Tourism local innovation system outputs (Source: Prats and Guia, 2005)

must contribute to the system's collective wealth, which in turn increases the wellbeing of all the agents who join it. Wellbeing is understood to be an improvement of the quality of life of all the elements, which is not based solely on the economic, environmental, or social benefits at an individual level, but is a perfect balance between all of them at a collective level.

This balance allows the system to become sustainable and generates a new and better situation that is a territory improvement, and which also feeds the agents and the macro-environment, varying the behavior of the system constantly, forcing it to reframe itself, and be constantly dynamic.

DESTINATION TOURIST PERCEPTION & NETWORK CONFIGURATION: A THEORETICAL PROPOSAL

The tourism image perceived by tourists and represented in Gunn's model (1972) has a close connection with tourism destinations, and in particular, with agents that interact in the promotion of the destination.

The seven-stages of tourism experience (Gunn, 1972), show interactions between tourists and tourism agents. This materializes, initially, in the process of searching for information, which is done voluntarily by tourists; and later, if tourists travel to the destination the tourism image is again modified by direct contact with tourism agents.

In this context, direct contact and coordination among internal and external agents of the tourism destination are also necessary, so that tourists can perceive a real tourism image of the destination both before and after traveling to the destination. If this situation occurs, it will be easier for tourism destinations to maintain their competitive advantage in a sustainable way.

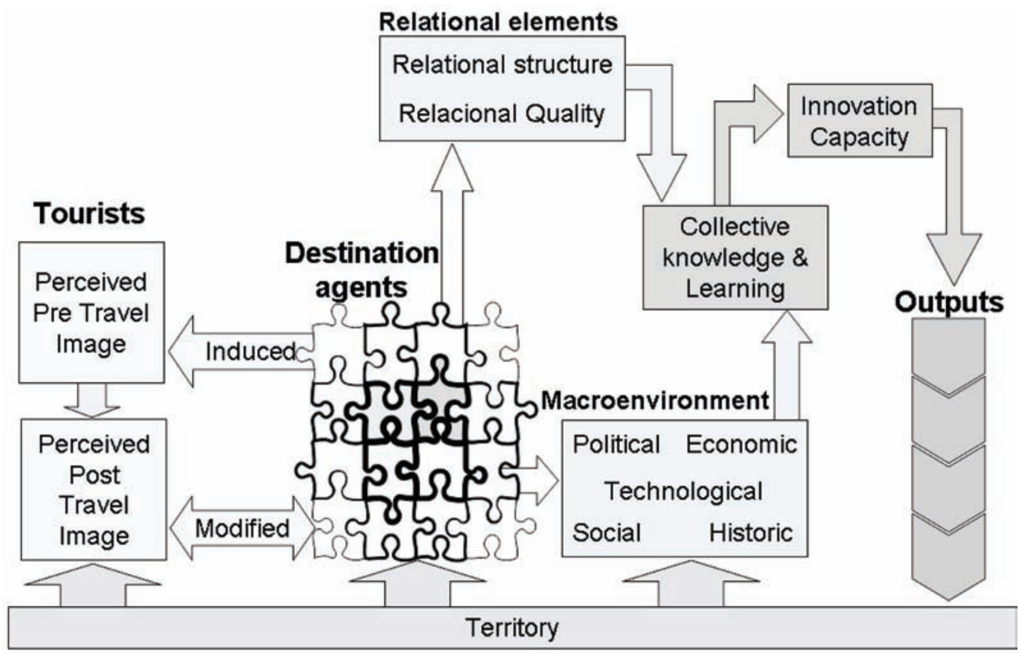
Therefore, tourism destinations, which are established as a network and based on trust among its members, can better guarantee a unique tourism image, which is more coherent with the reality of the tourism destination. This configuration has a close relationship with the structure of network, the quality of ties, and its macro-environment. Moreover, the innovation capacity generated has to be useful to adapt the induced image to the tourism product.

Figure 5 shows a market approach model, which focuses on demand and supply simultaneously and explains the relationship between them, taking into account the multiple factors that affect the behavior of tourism agents. As a consequence the competitive advantage of tourism destinations is explained.

In this model the interaction between tourism agents and tourists is highlighted, giving an overall picture of what happens in a destination. In this context, those tourism agents who interact frequently with tourists tend to be public administrations, private companies, and the local community.

First, public administrations have a direct relationship with tourists through the promotional actions of tourism destinations. These agents act as induced agents of tourism images by acting as

Figure 5. Market approach of tourism destination: Conceptual model



promoters and developers of destinations. In this case, following Gartner's classification (1993), public administrations act as Overt induced agents I, who emit an induced tourism image, which influences both the tourists' perceived image of a destination and the decision-making process at the moment of choosing a destination (Gartner, 1993).

Second, tour operators, as they form part of the private companies of a destination, also act as Overt induced agents II, because they have a clear interest in influencing the decision-making process of tourists at the time of selecting a destination. Public administrations, as well as tour operators, influence the "perceived pre-travel image" of tourists.

When the potential tourist travels to a destination and becomes a real tourist a direct interaction between tourists and tourism companies takes place, and this influences the perceived image of the place that tourists had before going there, creating a new image of the place (Gunn, 1972).

Finally, the local community has a strong relationship with the tourist and usually the tourist's real image is strongly modified by this kind of contact. In the author's opinion a key factor at this level is the perception the local community has of the tourism activity and the benefits that the inhabitants receive from it, because if a local community thinks that the benefits and damages that tourism causes are in perfect balance or in a more beneficial situation for the local community, these inhabitants will contribute to the tourists well-being, otherwise they will behave to the contrary.

Tourism research and training centers play a secondary, but fundamental role, especially as they might condition the induced tourism image through research projects, and simultaneously, they could also influence perceived tourism images indirectly by training the tourism workers who help tourists during their stay.

Therefore, the influence on "perceived post-travel image" comes from tourism companies, as well as the local community and research and training centers.

This model shows that, the relationships between tourists and tourism agents are systematic and necessary throughout the whole process. This means that this interaction is essential both before a tourist travels to a destination as well as during his/her stay.

If tourists do not go to a destination, this might mean that tourism agents cannot control the factors that generate the appropriate knowledge and transmit the right image to convince tourists. However, other uncontrollable factors exist, such as, the travel time needed, the distance to the destination, the money available or to what extent a tourism product fulfils tourist needs.

When tourists are dissatisfied with their visit because the "perceived post-travel image" is extremely different from the "perceived pre-travel image", there is another scenario where the agents have not transmitted the reality of the destination. In this situation tourism agents had the innovation capacity, but they had not used it in the correct way to obtain the desired image outputs. This context shows the importance of communication and coordination among all the tourism agents of a destination, to induce a real and homogenous image.

CASE STUDY

As we mentioned before, tourism products and even more evident tourists doesn't understand political boundaries, but regional and local governments use it to divide the territory. This situation causes management and commercialization problems that don't help to emit a coherent image of the whole destination to the possible tourists. These scenarios can be avoided developing networks for product commercialization that includes all the Overt induced I agents. Is one of the easiest way to start a network, because in that sense they can share marketing costs entering to scale economies.

We analyzed the public institutions' websites of a tourism destination in a local and regional level in order to know if tourism product is configured globally through collaboration between destination's tourism agents; or in a contrary way the tourism product is fragmented. This can help us to understand the level of tourism image coherence that is transmitted to tourist through Internet.

Methodology

In order to analyze how a tourism product is promoted in a local and regional level a website analysis of public institutions from the Girona's province, which has been conducted during the first trimester of the year.

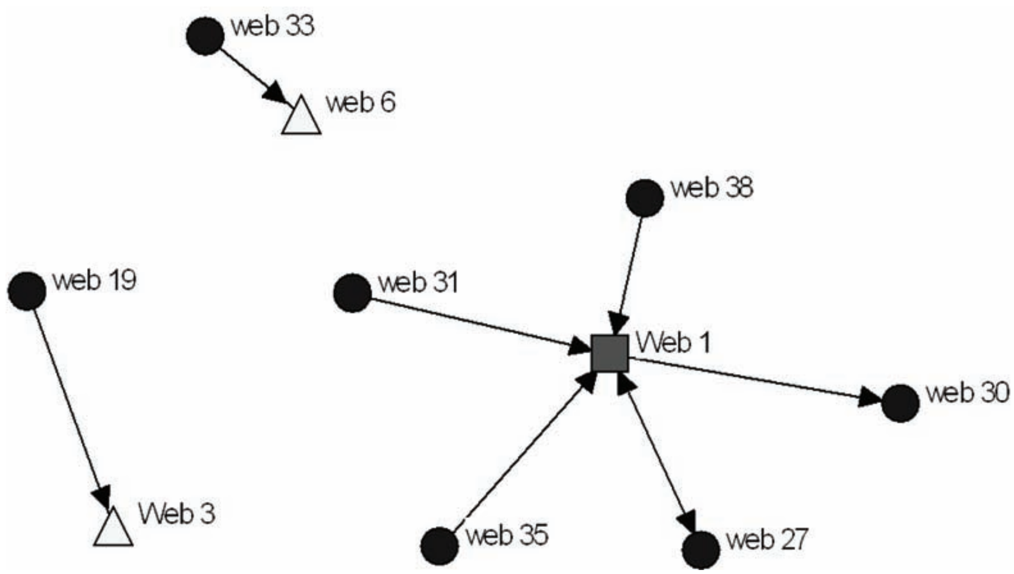
The Girona's province is situated in the north-eastern of Catalonia and their capital is the city of Girona. It is a very rich region in terms of natural resources and heritage; and it has a very privileged placement between the sea and the oriental Pyrenees. Tourism activity is mainly developed in the coastal area, although in Pyrenees tourism is also a relevant activity for their economy. In general terms, Girona's province is integrated by two tourism brands: (a) Costa Brava for the coastal area, and (b) Girona's Pyrenees for the mountain area. These two brands divide their territory and facilitate their promotion.

Websites of 41 public institutions have been analyzed, taking into consideration three labels of public institutions: (a) DMO of the province, (b) regional institutions, and (c) local institutions.

The existence of a section in their website related to tourism activities has been the criteria to select websites. In this context, we have analyzed the role of public agents that act as Overt induced I, following the Gartner's nomenclature (1993) as we mentioned above.

Data collection has been done taken into consideration the level of public institution searching the existing relationship among them. Three kinds of data have been collected. Firstly, it has been detected direct and indirect links that are placed into de website and which is the linked agent. Secondly,

Figure 6. Relational map of direct web links



it has been observed the quality of detected links and it has determined their typology (advertising links, friends' links or partners' links). Finally, it has been also analyzed the e-marketing resources that have been used in the websites.

In order to analyze data, it has been used UCI-NET 6.0 software (Borgatti, Everett & Freeman, 2002), which is a software specialized in social network analysis in a qualitative and quantitative way. In our case it has been used the application to represent graphically the social network.

Data Analysis

Observing the network that can be drowned after the analysis, is possible to assure that don't exist a common commercialization network in the Costa Brava destination. From the 41 analyzed webs only 10 have direct links with other promotion agent. The main problem is that only one of these links is bidirectional. This means that the rest maybe are not well developed links.

Another element to extract from this relational map is the fact that only 2 of 8 regional institutions appeared on it, and always as link receiver and not as a link creator, which means that they promote the tourism elements independently from the municipalities that they represent. (Figure 6)

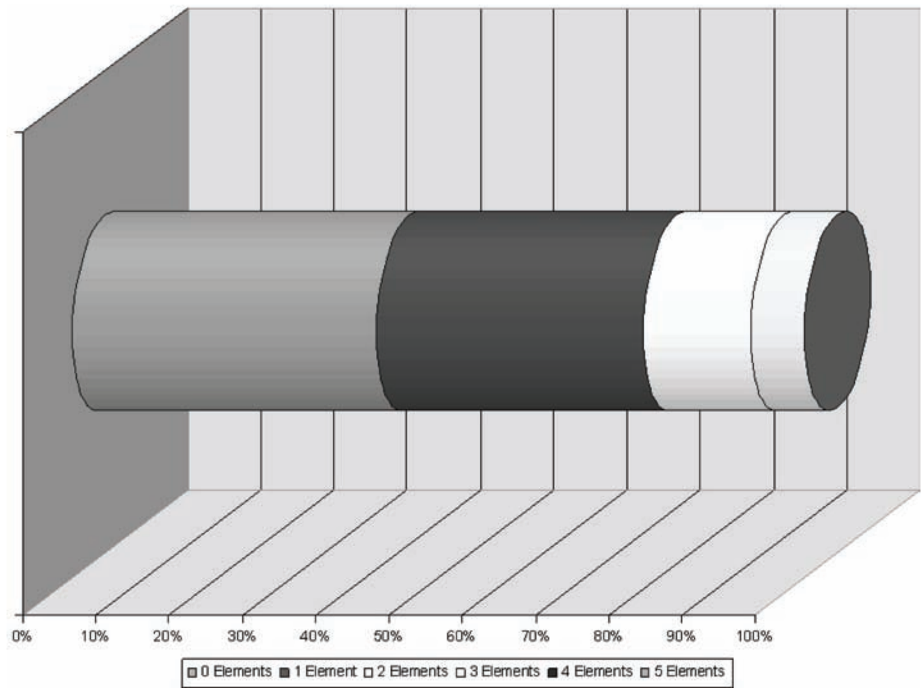
Jumping to the technological elements included into the webs that can help to emit a better image five different types of them were analyzed, the type of compelling web system, the multimedia systems divided into sound, video and photography galleries, and finally the interaction with the tourists represented by interactive maps (Table 1).

What is really shocking is that the main Catalan destination, only have little technological and interactive elements to attract tourists. This is done by the historical tourism tradition of the destination

Table 1. Use of technological and interactive web resources

Type	N
Advanced Compelling	18
Video	2
Sound	1
Photographic galleries	11
Interactive maps	4
Total	36

Figure 7. Percentage of technological and interactive web resources



when an intensive promotion to attract tourists never was needed. Most of the mature tourism destinations have a similar problem.

One third of the tourism Web sites of the destination don't have any of the mentioned elements and nearly half of them only have one. This situation sorts out that the emitted image trough websites and technological elements are really poor. At least one of the webs that have three of these elements is the one of the main DMO maybe the most visited site. The problem is that in this site there is not all the information related with destination as we can imagine for the link system showed before. (Figure 7)

CONCLUSION

Discussion

Initially, it has to be mentioned that it is important to observe demand and supply in an integrated way. This integrated vision gives a greater innovation capacity, in particular, regarding the destinations' tourism agents; who have a broader view of the possibilities of maintaining and improving the sustainable competitive advantage of the destination; allowing to consider the tourists' key role and how the tourism agents' interact with them, as well as when this interaction takes place, and what the basic tools that maintain this relationship are.

In this context it is assumed, as justified earlier, that a tourism destination's image is constructed socially during a complex process in a seven-stage tourism experience. Moreover, the necessity to manage this image is accepted as a method that influences the tourists' process of decision-making and for that reason it is necessary to pay special attention to the agents who take part in this.

Nowadays, Internet is positioned as a relevant tool to contribute to destination imagery formation as well as and information source for tourists that can influence their decision-making. At the same time, it is commonly accepted that city marketing differs in many ways of destination marketing, but if it is analyzed the common projected image trough websites the division line disappears. So we can assume that a big city like Barcelona can have the same number of official emitting image websites than a regional destination like Costa Brava.

In the authors' opinion, social network analysis is a perfect approach for studying tourism destinations. In this sense, they have highlighted the interaction among tourism agents to create a tourism product or service adequate to tourists' needs, as one of the most important factors. One of these interactions can be done in terms of public agents' promotion for the whole destination; this can be materialized though official websites and links those appear in it.

As can be seen, in the case study, the promotion through websites in Girona's province does not help to create a coherent and jointly destination image. This situation can be avoided with a better and wide interaction among the public agents, materializing it through their websites. In addition, probably, the low use of technological and interactive resources in websites is done by the hard sedimentation of tourism industry in this region, and in our opinion these elements will need to be improved if tourism destination goes into a declining process of their life cycle.

Limitations and Future Research

The work that has been presented in this article opens a wide field of future research that takes into consideration all the agents mentioned in the presented model. This empirical analysis will contemplate if having a whole picture of the market will contribute to really improving the planning and management of tourism destinations, as discussed.

In this first conceptualization the autonomous and organic agents from Gartner's model (1993), who generate an uncontrolled tourism image by induced agents, have not been considered. In future research, it would be necessary to revise the conceptual model and include autonomous and organic agents, to have a better proxy, taking into account that tourists also interact with autonomous and organic agents before traveling to the destination.

In the authors' opinion, it could also be interesting to consider the difference between real and potential tourists in future revisions of the model. This reflection could have relevant implications, especially for observing the induced tourism image and searching for explanations for a tourist's reasons to travel to a specific destination and not to others.

REFERENCES

Aaker, D. (2001). *Strategic market management*. (6th ed.). New York, NY: John Wiley and Sons.

Baloglu, S., & McCleary, K. W. (1999). A model of destination image formation. *Annals of Tourism Research*, *26*(4), 808–889. doi:10.1016/S0160-7383(99)00030-4

Baloglu, S., & Pekan, Y. A. (2006). The website design and Internet site marketing practices of upscale and luxury hotels in Turkey. *Tourism Management*, *27*, 171–176. doi:10.1016/j.tourman.2004.07.003

Becattini, G. (1979). Dal sttore industriale al distretto industrialle. Alcune considerazione sull'unità dell'economia industriale. *Revista di Economia e Politica Industriale, 1*.

Beerli, A., & Martín, J. D. (2004). Factors influencing destination image. *Annals of Tourism Research*, *31*(3), 657–681. doi:10.1016/j.annals.2004.01.010

Berger, P. L., & Luckmann, T. (1968). *Social construction of reality*. Amorrortu: Buenos Aires.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet 6 for Windows: Software for social network analysis*. Harvard: Analytic Technologies.

Choi, S., Lehto, X. Y., & Morrison, A. M. (2007). Destination image representation on the web: Content analysis of Macau travel related Web sites. *Tourism Management*, *28*, 118–129. doi:10.1016/j.tourman.2006.03.002

Chon, K. S. (1991). Tourism destination image modification process: Marketing implications. *Tourism Management*, *12*(1), 68–72. doi:10.1016/0261-5177(91)90030-W

Crompton, J. L. (1979). An assessment of the image of Mexico as a vacation destination and the influence of geographical location upon the image. *Journal of Travel Research*, *18*(4), 18–23.

Galí, N., & Donaire, J. A. (2005). The social construction of the image of Girona: A methodological approach. *Tourism Management*, *26*, 777–785. doi:10.1016/j.tourman.2004.04.004

Gallarza, M. G., Gil, I., & Calderón, H. (2002). Destination Image – Towards a conceptual framework. *Annals of Tourism Research*, *29*(1), 56–78. doi:10.1016/S0160-7383(01)00031-7

Gartner, W., & Hunt, J. (1987). An analysis of state image change over a twelve year period (1971-1983). *Journal of Travel Research*, *16*(2), 15–19. doi:10.1177/004728758702600204

Gartner, W. C. (1986). Temporal influence on image change. *Annals of Tourism Research*, *13*, 635–644. doi:10.1016/0160-7383(86)90006-X

Gartner, W. C. (1993). Image formation process. *Journal of Travel & Tourism Marketing*, *2*(2), 191–215. doi:10.1300/J073v02n02_12

Govers, R., & Go, F. M. (2004). Cultural identities constructed, imagined and experienced: A 3-gap tourism destination image model. *Tourism*, *52*(2), 165–182.

Granovetter, M. S. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, *91*, 481–510. doi:10.1086/228311

Guia, J., Prats, L., & Comas, J. (2006). The destination as a local system of innovation: The role of relational networks. In L. Lazzeretti, & C. Petrillo (Eds.), *Tourism local system and networking*. Elseivier: Amsterdam.

Gunn, C. A. (1972). *Vacationscape: Designing tourist regions*. Washington DC: Taylor and Francis/University of Texas.

Gunn, C. A. (1988). *Vacationscape: Designing tourist regions*. New York: Van Nostraud Reinhold.

Gunn, C. A. (1997). *Vacationscape: Developing tourist areas*. Washington DC: Taylor and Francis.

Håkansson, H., & Snehota, I. J. (Eds.). (1995). *Developing relationships in business networks*. London: Routledge.

Han, J. H., & Mills, J. E. (2006). Zero acquaintance benchmarking at travel destination Web sites: What is the first impression that national tourism organizations try to make? *International Journal of Tourism Research*, *8*, 405–430. doi:10.1002/jtr.581

Hashim, H. H., Murphy, J., & Muhamad Hashim, N. (2007). Islam and online imagery on Malaysian tourist destination Web sites. *Journal of Computer-Mediated Communication*, *12*(3), 16. doi:10.1111/j.1083-6101.2007.00364.x

Hite, J. M. (2003). Patterns of multidimensionality among embedded network ties: A typology of relational embeddedness in emerging entrepreneurial firms. *Strategic Organization*, *1*(1), 9–49.

Hjalager, A. M. (2000). Tourism destinations and the concept of industrial districts. *Tourism and Hospitality Research*, *2*(3), 199–213.

Holmen, E., Pedersen, A. C., & Torvatn, T. (2005). Building relationship for technological innovation. *Journal of Business Research*, *58*(9), 1240–1250. doi:10.1016/j.jbusres.2003.10.010

Hudson, S., & Ritchie, J. R. B. (2006). Promoting destinations via film tourism: An empirical identification of supporting marketing initiatives . *Journal of Travel Research*, *44*, 387–396. doi:10.1177/0047287506286720

Jackson, J., & Murphy, P. (2002). Tourism destinations as clusters: Analytical experiences from the World. *Tourism and Hospitality*, *4*(1), 36–52.

Jenkins, O. H. (1999). Understanding and measuring tourist destination images. *International Journal of Tourism Research*, *1*, 1–15. doi:10.1002/(SICI)1522-1970(199901/02)1:1<1::AID-JTR143>3.0.CO;2-L

Johnson, G., & Scholes, K. (1999). *Exploring corporate strategy*. Harlow, UK: Prentice Hall.

Larsen, G., & George, V. (2004, February). The social construction of destination image – A New Zeland firm example. *Working Papers*, 4/01.

Lazerson, M., & Lorenzoni, G. (1999). The firms that feed industrial districts: A return to the Italian source. *Industrial and Corporate Change*, *8*(2), 235–266. doi:10.1093/icc/8.2.235

Lee, G. M., Cai, L. A., & O'Leary, J. T. (2006). An analysis of brand-building elements in the U. S. state tourism Web sites. *Tourism Management*, *27*, 815–828. WWW.Branding.States.US. doi:10.1016/j.tourman.2005.05.016

Lundvall, B. A. (Ed.). (1992). *National systems of innovation. Towards a theory of innovation and interactive learning.* London: Pinter Publishers.

Mazanec, J., & Schweiger, G. (1981). Improved marketing efficiency through multi-product brand names? An empirical investigation of image transfer. *European Research*, *9*(1), 32–44.

McIntosh, R. W., Goeldner, C. R., & Ritchie, J. R. B. (2000). *Turismo: Planeación, administración y perspectivas*. Mexico D.F.: Limusa Willey

Miossec, J. M. (1977). L'image touristique comme introduccion à la géographie du tourisme. *Annales de Geographie*, 55–70. doi:10.3406/geo.1977.17568

Molina, F. X. (2001). European industrial districts: Influence of geographic concentration on performance of the firm. *Journal of International Management*, *7*(4), 277–294. doi:10.1016/S1075-4253(01)00048-5

Nelson, R. (1993). *National innovation systems*. New York, NY: Oxford University Press.

OMT. (1999). *Desarrollo turístico sostenible: Guia para administradores locales*. Madrid: OMT.

Porter, M. E. (1990). *The competitive advantage of nations*. London: The McMillan Press.

Prats, L., & Balagué, J. (2005). Cohesión y sostenibilidad, elementos clave en la competitividad del territorio turístico: El caso de la Costa Brava. *Retos Turísticos*, *3*(3), 8–15.

Prats, L., Camprubí, R., & Comas, J. (2005). Network ties relevance on the destination business relationships. In Universidade do Algarve (Eds.), *Recent developments in tourism research*. Faro: University of Algarve.

Prats, L., & Guia, J. (2005). The destination as a local system of innovation. In C. Petrillo & J. Swarbrooke (Eds.), *Networking and partnership in destinations and development management* (pp. 121-136). Arnhem: ATLAS.

Prats, L., Guia, J., & Molina, F. X. (2007). Tourism local innovation systems or how tourism destinations evolve. In M. Smith & L. Onderwater (Eds.), *Destinations revisited. Perspectives on developing and managing tourist areas* (pp. 35-53). *Arnhem: ATLAS*

Rallet, A., & Torre, A. (2004). Proximité et localisation. *Economie Rurale, 280*.

Ritchie, J. R. (1993). Crafting a destination vision: Putting de concept of residence responsive tourism into practice. *Tourism Management*, *14*, 279–289. doi:10.1016/0261-5177(93)90006-7

Roberts, J. (2001). Challenges facing service enterprises in a global knowledge-based economy: Lessons from the business services sector. *International Journal of Services Technology and Management*, *2*(3), 402–433. doi:10.1504/IJSTM.2001.001612

Selby, M., & Morgan, N. G. (1996). Reconstructing place image: A case study of its role in destination market research. *Tourism Management, 17*, 287–294. doi:10.1016/0261-5177(96)00020-9

Sorensen, F. (2004). *Tourism experience innovation networks*. Doctoral dissertation. Roskilde University, Denmark.

Sundbo, J. (1998). *The theory of innovation: Entrepreneurs, technology and strategy*. Cheltenham, UK: Edward Elgar.

Telisman-Kosuta, N. (1989). Tourism destination image. In S. F. Witt & L. Moutinho (Eds.), *Tourism marketing and management handbook* (pp. 557-561). Cambridge: Prentice Hall.

Urry, J. (1990). *The tourist gaze*. London: Sage: Wellman, B. (1997). An electronic group of virtually a social network. In S. Kiesler (Ed.), *Culture of the internet*. Mahwah: Lawrence Erlbaum.

This work was previously published in Information Communication Technologies and City Marketing: Digital Opportunities for Cities Around the World, edited by M. Gasco-Hernandez; T. Torres-Coronas, pp. 184-202, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 4.11
Successful Web-Based IT Support Services:
Service Provider Perceptions of Stakeholder-Oriented Challenges

Vanessa Cooper
RMIT University, Australia

Sharman Lichtenstein
Deakin University, Australia

Ross Smith
RMIT University, Australia

ABSTRACT

Web-based self-service systems (WSSs) are increasingly leveraged for the delivery of after-sales information technology (IT) support services. Such services are offered by IT service providers to customer firms and increasingly involve business partners. However little is known of the challenges faced by IT service providers as a result of the involvement of the other firms and their employees (end-users). This paper reports related findings from an interpretive study of IT service provider perceptions in six multinational IT service provider firms (Cooper, 2007). The findings highlight that, for IT service providers, (1) it is important to consider and resolve the needs and concerns of other key stakeholders, and (2) significant challenges exist in doing so. The main contribution of the paper is the identification of the key challenges involved. Important implications for theory and practice are discussed.

INTRODUCTION

The continued maturation of the Internet has been accompanied by a corporate shift from the provision of goods to the provision of services, with parallel development of relevant new business models and marketing paradigms (Rust, 2001). Many businesses have developed *E-services*, defined as the provision of services by electronic networks such as the internet (Rust, 2001). Despite the increasing importance of E-services to business success, electronic commerce researchers have been slow to investigate associated issues. As the E-services value chain requires different types of processes and offers greater flexibility in comparison with offline services, there are new research challenges to be explored (Hofacker et al., 2007).

An important new source of value presented by E-services is *supplementary E-services* such as electronic provision of pre- and post-sales customer support for purchased services and products (Hofacker et al., 2007). Experts further suggest that the successful provision of supplementary E-services may be more important strategically to service providers and vendors than the quality of originally-purchased services and products (Otim & Grover, 2006; Piccoli et al., 2004). Marketing of supplementary services (offline and online) can provide differentiation, improve customer service, increase customer retention and lower service costs (Levenburg & Klein, 2006; Reichheld & Schefter, 2000).

This article focuses on the provision by service providers of supplementary E-services to customer firms ("enterprise customers") using the World Wide Web ("Web"). To leverage this market successfully, vendors and service providers aim to improve the implementation and delivery of E-services by employing a systematic approach. One such approach is a Net-based Customer Service System (NCSS) which has been described as "a network-based computerised information system that delivers service to a customer either directly (e.g. via a browser, PDA, or cell phone) or indirectly (via a service representative or agent accessing the system)" (Piccoli et al., 2004 p.424).

This article focuses on the use of a key type of NCSS based on a Web interface – a *Web-based Self-Service System (WSS).* Self-service is gaining importance in contemporary organisations primarily for cost reduction reasons (Doyle, 2007). This article explores the context of *managed information technology (IT) support services.* In this setting IT service providers employ WSSs to provide after-sales IT support to enterprise customers.

Key stakeholders comprise the service provider firm and its employees, business partners and their employees, and enterprise customers and their employees. As this article will show, the involvement of the key stakeholders results in significant challenges for IT service providers aiming to provide successful after-sales support by means of a WSS. These challenges will be explored in the article by examining the IT service provider perspective.

A knowledge transfer lens is used to explore this topic as the transfer of after-sales IT support knowledge (such as IT solutions) from an IT service provider firm to a customer firm is central to the concept of successful after-sales Web-based support services (CSI 2002; Koh et al., 2004).

This article draws on a large study investigating the successful provision of managed after-sales IT support when facilitated by WSSs (Cooper, 2007). The perspectives of six large multinational IT service providers were obtained and analysed. The views of IT service providers are important to understand for improved service provision (Pitt, 1998). Our study focuses on the use of operational IT support services, relating to (1) assembling and operating the core IT environment, and (2) providing key value-adding services such as the Service (Help) Desk (Peppard, 2001).

Five further sections complete this article. Section 2 provides a theoretical background by reviewing representative literature. Section 3 outlines the research design. Section 4 describes the key challenges relating to stakeholders, identified when an IT service provider transfers after-sales IT support-oriented knowledge to enterprise customers when WSSs are used to facilitate service provision. Section 5 discusses the key challenges. Section 6 summarises the main points, draws conclusions, reflects on the limitations of the findings and offers suggestions for future research.

THEORETICAL BACKGROUND

We first situate WSSs within a Customer Relationship Management (CRM) context as the strategic goals of supplementary services - such as after-sales support provision using WSSs - include improving customer service and increasing customer retention (Levenburg & Klein, 2006). We then review the use of WSSs for after-sales IT support provision to enterprise customers. Next a stakeholder-oriented framework of successful Web-based enterprise customer service drawn from earlier findings highlights the importance of stakeholder relationships and related knowledge flows. Finally, the section reviews the knowledge transfer process and the transfer of after-sales support knowledge from an IT support organisation to an enterprise customer.

Customer Relationship Management

In recent years CRM has emerged as a potentially powerful organisational strategy to enable a vendor or service provider to better identify and satisfy customer needs and retain customer loyalty. Enhanced customer relationships may also lead to improved customer-related operational effectiveness and a higher return on investment for the organisation (Barua et al., 2004).

To improve the customer service experience and meet other CRM objectives, the aggregation of data, information and knowledge about the customer is important. Specialised software applications that perform electronic CRM (eCRM) have been developed for this purpose. A common example of an eCRM application is a Web interface, supported by database and data mining tools which record past customer transactions and analyse data to identify customer segments, match products to customer profiles, and better understand target demographics and psychographic characteristics (Brohman et al., 2003). Customer service agents can utilise this “customer intelligence” to potentially up-sell or cross-sell products and services (Brohman et al. 2003). A WSS is an important type of operational eCRM application (Geib et al., 2005; Khalifa & Shen, 2005) and is discussed below in the context of managed IT support services.

WSS and After-Sales IT Support for Enterprise Customers

A WSS is a Web-based information system that enables organisations to move from labour-intensive manual processes towards low-cost automated Web-based self-service (Pujari 2004). WSSs can facilitate the offering of customer support services for pre-sales, sales, and after-sales activities. They are underpinned by complex information systems, complemented by a customer contact centre, and integrated with a multi-channel service strategy (Negash *et al.* 2003).

WSSs offer important advantages to service providers and customers (Geib et al., 2006; Pujari 2004). Such advantages include electronically leveraging the Web interface, customer/service-provider (and customer/customer) interactions, knowledge management (KM) principles and self-service principles in order to capture and provide information and knowledge useful for pre-sales, sales and after-sales support.

In managed IT service environments, WSSs offer informational, transactional, and proactive support services. Informational support includes “break-fix” support which provides customers experiencing technical problems with resolutions to their problems. This type of support includes (1) unassisted support such as answers to Frequently Asked Questions (FAQs) and the download of software patches and (2) assisted support such as peer-to-peer online fora, e-mail and online chat. Other informational support includes the provision of information and knowledge to assist with enquiries and enable customers to access best-practice – for example, by the publication of White Papers. Transactional support includes case tracking, whereby the customer initially documents their IT problem scenario, requests assistance from the support organisation, and is subsequently able to monitor the support organisation's progress in resolving this problem. Proactive support includes the embedding of problem detection support software on customer end-user computers, and personalised messages directing the customer to potential product or service purchases. In the study reported in this article we consider primarily informational and transactional supplementary support services.

WSSs can reduce customer support expenses by empowering employees (Support Agents). These employees provide support in the form of knowledge such as solutions to customers' IT problems. WSSs boost Support Agent productivity by providing access to a comprehensive knowledge base that includes many solutions, delivering convenient and higher quality support, and increasing customer loyalty and retention.

Non-IT businesses outsource IT support services to service providers because IT is not their core competency, service levels are likely to improve, and Total Cost of Ownership should be reduced (CRMInd, 2006; SSPA, 2004). Furthermore, client firms are often receptive to Web-based support delivery (SSPA, 2004). Thus the Web is considered by customer firms as an important channel for IT support delivery.

Despite the potential benefits of a WSS to service providers and enterprise customers, there has been a long-reported dissatisfaction with Web-based self-service (Barnes et al., 2000; Meuter et al., 2003; Ragsdale, 2007). Clearly success from the use of a WSS cannot be assumed and the critical success factors (CSFs) and challenges involved should be identified and addressed.

The study that is reported in this article identified a set of CSFs (Cooper et al, 2005, 2006a, 2006b; Cooper, 2007) and also identified important challenges, as perceived by IT service providers, many of which relate to other key stakeholders. In the next section we develop a stakeholder-oriented framework which situates WSSs in a broader enterprise customer service context and highlights the interactions with stakeholders.

STAKEHOLDER-ORIENTED RELATIONAL WEB-BASED ENTERPRISE CUSTOMER SERVICE

A broad definition of stakeholders is "all those parties who either affect or who are affected by an organisation's actions, behaviours and policies" (Mitroff, 1983). Management concerns about stakeholders arise because stakeholders have varying perspectives of the underlying problems and their ideal solutions may differ. It is important to business success that a firm resolve conflicting stakeholder needs (Hatch, 1997).

Experts report that the needs, roles, responsibilities, relationships and other interactions of stakeholders are especially important to the success of business-to-business (B2B) commerce (Chua et al., 2005; Kandampully, 2003; Pan, 2005; Ritter & Gemunden, 2003; Schultze & Orlikowski, 2004; Singh & Byrne, 2005). In B2B, businesses are increasingly interdependent and the stakeholder issues must be carefully managed (Kumar & van Dissel, 1996). However stakeholder roles and responsibilities along the value chain are complex (Chi & Holsapple, 2005; Ritter & Gemunden, 2003). Chi and Holsapple propose a model of stakeholder collaboration in B2B highlighting three behavioural processes: knowledge sharing, participative decision-making and conflict governance. Other reasons for understanding stakeholder issues in B2B commerce include the need to manage stakeholder expectations (Singh & Byrne, 2005).

Stakeholder relationships in services are different to mere service encounters and are associated with emotions and expectations (Gutek et al., 1999). Such relationships and associated knowledge flows are important in managed IT services environments (Dahlberg & Nyrhinen, 2006; Xu, 2007) as the quality of knowledge exchange influences the quality of outsourcing relationships (Gong et al., 2007). Of interest to this article, the adoption of WSSs can potentially strengthen stakeholder relationships (Bhappu & Schultze, 2006).

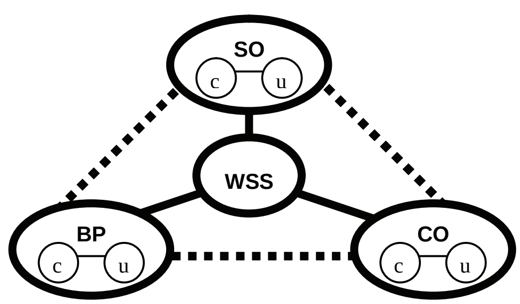
From the evidence above and a review of the literature (Cooper, 2007) we have developed a framework (Cooper et al., 2006a) that conceptualises key stakeholders and relationships in a B2B service context. Such a framework can be helpful in understanding complex knowledge transfer from service providers to customer firms. As mentioned earlier, such knowledge transfer is central to the provision of IT support. The framework in Figure 1 depicts key relationships between the main stakeholder types and a WSS. It shows three key types of stakeholder organisations which may be involved in support provision – a support organisation (previously termed "service provider"), business partner, and customer organisation – and their interaction with one another directly and indirectly via a WSS. At each organisation there are corporate entity representatives (for example, managers) interacting with end-users. The framework clearly highlights the interdependencies found in the multi-stakeholder managed IT support environment.

Knowledge Transfer for IT Support of Enterprise Customers Using WSS

The key knowledge process explored in this article is inter-organisational knowledge transfer. We focus on reviewing staged processual and network models of knowledge transfer for reasons of relevance.

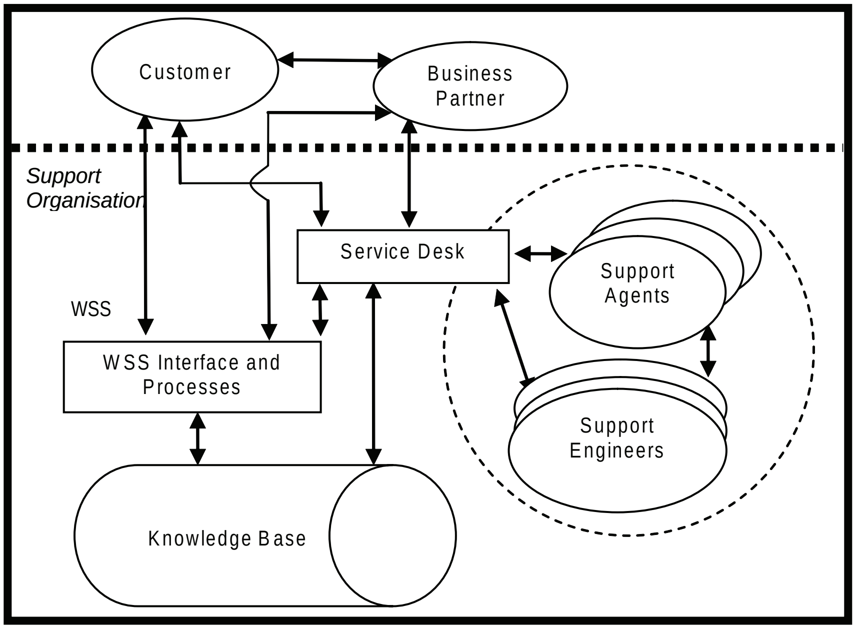
Researchers advocating a staged processual approach to knowledge transfer argue that this can unlock the inner workings of the process and enable a more nuanced identification of barriers and enablers (Szulanski 1996). Staged inter-organisational knowledge transfer models include

Figure 1. A stakeholder-oriented relational framework for web-based enterprise customer service (Cooper et al, 2006a)



LEGEND
SO = Support Organisation
BP = Business Partner Organisation
CO = Customer Organisation
c = Corporate Entity
u = End-User Entity
WSS = Web-based Self-service System
___ = Relationship via WSS
- - - = Direct Relationship

Figure 2. Knowledge transfer in managed after-sales Web-based IT support (Cooper et al, 2006a)



those developed by Cranefield and Yoong (2005) and Nieminen (2005). Cranefield and Yoong's (2005) model comprises six stages: engaging; defining; seeking; articulating; integrating; and disseminating. Among the benefits of this model is its identification of key organisational influences on knowledge transfer during each stage. However the factors identified as most relevant to knowledge transfer in an inter-organisational context are the need for fit between the transferred knowledge and the receiving organisation's current organisational objectives and traditional discipline area, and the need to avoid use of "non-transferable examples" which cannot be readily transferred to other organisational contexts. Nieminen's (2005) model of inter-organisational knowledge transfer focuses on the role of a receiving organisation in enabling knowledge transfer. He writes of the need for a receiving organisation to have a range of competencies among its employees in order to reduce any significant absorptive gap (Lane & Lubatkin, 1998) between the two firms involved in the transfer. Researchers proposing a networking approach to knowledge transfer also adopt a relational approach based on understanding how patterns of connections between individuals and groups facilitate knowledge sharing and knowledge transfer (Kakabadse et al., 2003).

We next describe the knowledge transfer process in the managed IT support context (Figure 2) drawing on literature and empirical data from a case study (Cooper, 2007). IT service providers respond to enterprise customers' after-sales enquiries, incidents and problems regarding core IT products and services by providing and transferring support-oriented knowledge. The knowledge is complex and based on solutions to the enquiries, incidents and problems reported.

Both tacit and explicit knowledge may be transferred. Tacit knowledge is knowledge that resides in the mind and is difficult to articulate. Explicit knowledge is a representation of knowledge. An example of explicit knowledge in the managed after-sales IT support context is an IT solution stored in a knowledge base.

As shown in Figure 2, when a customer firm experiences IT incidents or problems, IT professionals at the firm may use the telephone channel or Web interface to obtain a solution. When the telephone is chosen, first tier Support Agents identify potential solutions by accessing tacit knowledge or by using the WSS to search the solutions knowledge base. More complex problems are escalated to more experienced Support Agents. Tiers of Support Engineers resolve the most difficult problems by drawing on valuable tacit knowledge. New and evolving solutions are captured in the knowledge base as explicit knowledge and organised for reuse. Successive efforts are made by Support Agents to address related customer questions. Such efforts aim to assist the customer firm in institutionalising support-oriented knowledge. Business partners may assist in the knowledge transfer process by providing partial solutions or other support and may have access to the service provider's WSS knowledge base.

RESEARCH DESIGN

Initially a comprehensive literature review was undertaken. Next, in 2005 an in-depth case study was conducted at the Australian headquarters of a large best-in-class multinational IT services organisation, "ServIT" (a pseudonym). ServIT had previously secured prestigious global awards for its WSS. It defined a successful WSS as one that delivers increased customer satisfaction while reducing service provision costs and this definition of success was used throughout the project. Case data was collected from semi-structured interviews with twelve senior customer service managers, IT support managers and support staff, observations of WSS use, document collection, and a focusing workshop (Rockart, 1979) with participation by five senior customer service managers and IT support managers. Case study data were analysed using qualitative content analysis techniques (Mayring 2000) by two researchers working independently to identify the key challenges relating to stakeholders. A set of CSFs for the transfer of after-sales IT solution-oriented knowledge to enterprise customers when a WSS is used, and a set of key challenges, have been identified.

In 2006 a focus group of Australian senior customer service and IT support managers from an additional five large multinational IT service provider organisations, all with successful WSSs, was conducted, seeking to confirm and extend the findings from the ServIT case study. In the focus group discussion it became clear that, while the WSS of each company was successful, various challenges remained. Two researchers working independently confirmed and extended the set of CSFs and key challenges working inductively from the transcript. *Many of the challenges were found to be related to other stakeholders and it is these challenges which are the focus of this article.*

The quality of the findings was assured by appropriate techniques. Achieving validity in qualitative research requires a fair, honest and balanced account from the viewpoint of someone who lives in the social situation (Neuman, 2006). Construct validity was assured in the study by use of multiple sources of evidence. Participants were also given copies of interview summaries and the outcomes from each phase of the research were provided to participants for confirmation. Reliability was increased via a case study protocol (Yin 2003) to document all procedures and problems. Triangulation was achieved by collecting and analysing data from a variety of sources at ServIT and establishing consistency of results. The results of the cross-organisational focus group added another level of triangulation.

FINDINGS

The six participating IT service providers utilised WSS strategies to increase customer satisfaction while reducing the cost of IT support provision. Each firm offered a suite of informational, transactional and proactive support services using WSSs. Twenty-seven CSFs were identified. While the CSFs are not the focus of this article they are summarised for reference in the Appendix. The CSFs were also classified non-orthogonally into six categories:

- Organisational Commitment and Readiness: The organisation must manage the policies, processes and cultural issues which will affect its ability and willingness to embrace Web-based Self-service.
- Manage for Strategic and Operational Benefits: The WSS strategy must assist the organisation in attaining its strategic and operational objectives.
- KM Capabilities and Processes: The organisation must practice the principles of knowledge management and implement associated knowledge management processes, to maximise the benefits received from the WSS strategy.
- IT Infrastructure Capability: The organisation must have an adequate IT infrastructure in place, to enable it to participate in Web-based Self-service.
- Experience Management: The WSS should manage the stakeholder's experience, both at the corporate and end-user level. The stakeholder experience will directly affect satisfaction levels and therefore ongoing use of the WSS.
- Content: The WSS must contain useful, accurate and up-to-date content in order to resolve the end-user's support issue or knowledge requirement.

These six categories were believed by the IT service provider participants in the study to apply to *all* three stakeholder organisations (that is, those shown in Figure 1). This finding is discussed further in the next section, as it presented one of the main challenges for IT service providers. Further details of the CSFs may be found in earlier publications (Cooper et al, 2005, 2006a, 2006b).

SERVICE PROVIDER PERCEPTIONS OF STAKEHOLDER-ORIENTED CHALLENGES

This section summarises the eight major challenges faced by IT service provider firms, pertaining to key stakeholders. Representative quotes from the ServIT case study and cross-organisational focus group illuminate the challenges. Company names are pseudonyms for reasons of anonymity. The reader is reminded that the findings have been derived from the perspective of the IT support firm only. A further comment on this limitation is made at the end of the article.

First, participants strongly believed that <u>all</u> stakeholder viewpoints should be considered by an IT support organisation when planning, implementing and managing a WSS for after-sales IT support provision. Similar findings were made for electronic business settings more generally by Kandampully (2003) and Singh and Byrne (2005). However our study highlighted the potential for different stakeholder types to hold unique perceptions. For example, while a support organisation may find the transfer of IT solutions to a customer firm highly desirable, it was questioned by IT service provider participants whether end-users at a customer enterprise would feel the same way:

I think the provider [support organisation] is interested in transferring knowledge so [that] *they don't have a problem any more and they can manage their costs and help the customers. [However, the customer firm's end-user is thinking] I am interested, not in receiving knowledge… [but] I am interested in my problem being fixed. Don't give me all this stuff [the details of the problem and solution]. Tell me what the problem is so I can fix it. I don't want the transferring of any knowledge.* (Senior IT Architect, I-Systems)

This is an important finding deserving of further research. If an end-user at a customer firm does not wish to learn from a provided IT solution, how can a support organisation ensure such learning? Proposed cost and efficiency benefits to a support organisation stemming from knowledge transfer may have negative implications for customer satisfaction. It should be noted however that end-users at a customer firm have specialised job roles with role-based knowledge needs. Thus, for example, while some end-users may only desire a resumption of IT operations using a supplied solution, other personnel such as Database Administrators are likely to be highly interested in learning about a solution and gaining more general support knowledge which could be useful at a future time. On the other hand if a professional at an enterprise customer has full knowledge of a solution the customer organisation may become increasingly independent of the support organisation. Ultimately the success of the support organisation can be affected. Thus there may be a level of knowledge which a support organisation may wish to retain and not transfer to customer firms.

Second, participants questioned whether the different stakeholder types might interpret the requirements for CSFs (Appendix) differently. For example, for *CSF-9 Ease of use* participants mentioned that regular end-users of the WSS interface would prefer an efficient interface, whereas novice users would prefer easy-to-use interfaces. As a second example, for *CSF-1 Cost-effectiveness* some participants mentioned that cost-effectiveness to a support organisation, when a WSS is used, is not the same as cost-effectiveness from the perspective of an end-user at the customer firm. End-users are concerned with efficiency gains and usefulness of the knowledge gained from the WSS to perform their jobs (Cenfetelli et al., 2005), rather than the financial costs involved.

Cost effectiveness, from both the end-user of the service and for ServIT, are equally important, but I think you will find some subtleties in [how they are] both explained as cost-effectiveness. For an end-user it's 'Can I actually get the results quickly from my perspective?' From ServIT's perspective, it is 'Can we actually reduce the cost of this service for our customers so that we are actually making a profit? (Consulting Services Division Manager, ServIT)

Third, as described earlier, the six CSF categories were found to apply to *all three* stakeholder organisations. This finding supports prior research by Schultze and Bhappu (2005) who note that customers are often partly responsible for a service provider's success. Clearly, a support organisation will have very little control over whether a customer organisation and business partner organisation do, in fact, address the six CSF categories. As a result, performance measurement and management of WSS strategies in a B2B context will be challenging.

Kurnia and Johnston (2002) note the importance of industry capability in internet trading for corporate adoption of an inter-organisational system. Thus if corporate customers and business partners in a managed IT support situation do not manage the CSFs, the support organisation's WSS may not be successful. As electronic commerce environments are increasingly interdependent (Kumar & van Dissel, 1996; Kurnia & Johnston 2002), a support organisation may benefit from providing further assistance to business partners and customer organisations to better manage the key factors.

One promising strategy may be for a support organisation to conduct education and training programs with partner and customer organisations to increase their awareness and understanding of CSFs. Such education and training could be provided by online support, training, newsletters and so on. Indeed participants also identified the importance of education and training in terms of developing relationships with the customer and ultimately meeting key objectives for CRM. Thus education and training programs will not only provide partner and customer firms with the required knowledge to increase their awareness of WSSs but will provide an opportunity for the support organisation to better understand the needs of the customer and business partners in a WSS context. This finding supports current literature highlighting the importance of developing relationships with enterprise customers when providing support services (Peppers & Rogers, 2001; Pujari, 2004).

A second strategy that is more challenging would be to ensure, when developing external service contracts, that customer and business partner organisations are obliged to meet minimum standards relating to CSFs. Any failure to address CSFs at customer firms and business partners would affect the ability of a support organisation to service them effectively. However such a strategy may be incompatible with the development of improved relationships with business partners and customer firms. It is also important to recall that this study only investigated the IT support organisation perspective. While IT support organisations identified the CSFs as important also for partner and customer firms to address, studies of partner and customer perspectives of CSFs may provide quite different results. This question should be explored in future research.

Fourth, the interesting issue was raised of whether relationships are possible—or even enabled—by a WSS. A two-way relationship was posited by several participants, whereby if an end-user trusts the Web-site and the organisation behind it, this trust forms part of the relationship, with the WSS simply providing the connection:

Although it is a piece of software, it is the front end of a company. For instance, if I feel I have a positive relationship with my Netbank Web-site because I trust it, then that is part of the relationship. (IT Customer Consultant, DistSystems)

Others, however, saw Web-based relationships as perhaps one-way only, whereby it was questioned whether there is a relationship between the end-user and the providing organisation, when the end-user uses the WSS anonymously.

Fifth, it was also suggested that stakeholder relationships can be more complex than Figure 1 allows. Participants observed that relationships with stakeholder types may vary.

We can have the relationship with either both the customer and the business partner or we can have it directly with the customer or we could just have it with the business partner…I mean there are just so many different combinations that that relationship can actually take. (IT Customer Consultant, DistSystems)

Consequently in some instances Figure 1 should be modified. For example, for supporting some customer firms a business partner is involved, while for supporting other customers there are no business partners involved.

The complexity of relationships would also increase as stakeholder organisations increase in size. Multiple relationships would be developed at the individual, departmental and corporate levels. This finding is significant as while it has been acknowledged in a growing body of literature that developing relationships with partners is important (e.g. Vlachopoulou & Manthou, 2003), there is very little literature concerning partner-unit-based relationships.

Sixth, intellectual property (IP), security and privacy issues were found problematic for stakeholder-oriented reasons. Inter-organisational firms in collaborative relationships must protect certain knowledge which may be strategic (Solitander 2006). In the present study, service provider firms were concerned about the loss of IP in the leverage of user fora for customer support. In such fora, end-users may provide solutions to reported problems however it can be unclear who owns the IP that is the solution. Scholars have noted that in customer co-production of service, as illustrated by this situation, new models of digital governance and customer-based innovation are needed (Rai & Sambamurthy, 2006).

IT support organisations, business partners and customers frequently operate across international borders. Support organisations must consider issues of security, privacy and IP within the context of off-shore environments (Rai & Sambamurthy, 2006; Tafti, 2005). The IT service providers in this study expressed uncertainty about the ability of national legal systems to deal with such complex issues. For example, customer end-users may be unaware of the location where their personal information is stored and retrieved. Further, while customer organisations may be aware that, in some instances, third parties such as a support organisation's business partners have access to their stored personal information, they may not be aware of the eventual *use* of this information or its accessibility. A perceived potential privacy violation may lead to unforeseen competitor problems – for example, where business partners compete with a customer enterprise. Sensitive information, knowledge and IP may indeed be compromised (Rai & Sambamurthy, 2006).

A really interesting one that is going through the legal battles now and it comes back to privacy. As a main [support] organisation, you have information that could be of a private nature. If you then make the information available to another organisation that does not have the same privacy policy as you, and then [if they] were to use that information in such a way that it violates somebody's privacy, who is ultimately responsible? That is a huge legal question that is being tested now. Does it come back to the original person or the person who let the information go? (Emerging Technology Consultant, OpSys)

Seventh, understanding the complexities of the business partner alliance was an issue. Not all service providers participating in the study agreed with the distinction between a business partner relationship and a customer/supplier relationship. For example, a participant from one organisation claimed that his organisation did not have business partners, while others claimed that their organisations were, in fact, partners of his organisation. It was advocated that the distinction between a business partner and customer/supplier relationship, surrounds whether the relationship is a one-off transaction, or whether it is an ongoing relationship. In an enterprise IT services context, however, there is an apparent blurring of relationships. In some projects, organisations may be considered business partners, while in others they would be considered competitors:

DistSystems, DataCorp, OpSys and I-Systems can say, we have been working together as business partners and as competitors and sometimes even on the same bid…on the one hand you compete, in the next 30 seconds, I might be talking to OpSys about something we are competitors about, and in another thing we are working together on, you switch hats, you switch alliances, it is just the way things are these days…. (Senior Systems Consultant, I-Systems)

Some of the complexities of business partner alliances surround intellectual property and security and privacy, discussed earlier. Lei (1997) argues that, with respect to strategic alliances, regardless of the various types of legal structures which may be put in place, over time companies will absorb and internalise skills, regardless of the amount of formal, legal ownership that is demarcated by the alliance structure. Further, Ferdinand and Simm (2007) suggest a need for increased research on illegal inter-organisational knowledge transfer. Participants in our research study also expressed a need for greater understanding of this complex area. They believed it likely that associated concerns would continue in the foreseeable future.

Eighth, managing customer contributions to service was considered an important challenge. The growing literature on co-production highlights the productivity benefits as well as the managerial challenges that arise when customers become "partial employees" (Benapudi & Leone, 2003). The advantages of co-production include decreased cost of service provision and greater control and autonomy for the customer, while disadvantages include the difficulty of controlling service quality when customers are actively involved in the production process (Schultze & Bhappu, 2005).

In the study IT service providers raised concerns relating to the contribution of knowledge by customers to their own service fulfillment, optimisation and improvement. At ServIT when customers interacted and shared resolution-knowledge in online for a the knowledge was not captured permanently. Further, the online fora were open to a variety of end-users causing new issues of accuracy and liability:

ServIT has ventured into the hosting of forums. Solutions can be provided by non-ServIT people and that's a potential conflict between the reliability of our knowledge and the fact that we are opening up [knowledge] to end-users, which could have a good result, but there is a danger that we are facilitating the [incorrect] solution. (IT Manager, ServIT)

ServIT invests significant resources in its legal team and explicitly states in the terms and conditions of using the WSS and in support contracts that ServIT will not be responsible for any degradation of a customer's systems, if an enterprise customer decides to act upon information derived from the WSS. Another challenge is that while ServIT monitors fora content, it relies more on a "merit" based system whereby forum users are allocated merit points by original posters (of questions) when they provide valuable responses. A system of points aggregation motivates users to share valuable knowledge and thus contribute to E-service provision. However, the researchers found from forum observations that this scheme occasionally led to highly successful users moving on to create their own sites independently of the service provider, such was their fame and following.

Other potentially negative impacts of customer co-contribution include potential defamation in the fora. While the terms and conditions of posting to online fora state that users should not post defamatory statements, such defamation sometimes eventuates.

DISCUSSION

The findings above suggest a need for industry sector improvement. For the IT support industry to learn to do things better, learning at the industry level is needed. Prior studies suggest the importance of business learning in a network context (c.f. Knight, 2002). This would entail new collaborative electronic business projects focused on IT support provision involving IT service providers, business partners and customer organisations. According to Cameron's (2005) review of prior relevant studies, there are four important influences to consider for successful electronic business collaboration: motivation, capability, communication and coordination.

First, organisations must be motivated to participate in collaborative projects. Second, the desire to increase organisational capability (skills and knowledge) can be a powerful motivator. Third, it is important to communicate about the value of such collaborative projects within organisations. Fourth, industry groups can assist in coordinating such projects for successful conclusions.

As many of the key challenges identified in the previous section centre on the potential for stakeholder conflict, the development of individual IT service contracts (including service requirements) with customer firms and business partners requires attention. This process should entail greater clarification and negotiation of stakeholder needs. The objectives of each organisation should be articulated during joint planning, which is an important aspect of business collaboration in supply chains (Holsapple & Jin, 2007). Holsapple and Jin note several other important collaborative decisions typically made in a supply chain. The findings from our study suggest a need for focused studies seeking to identify key problem points in decision-making where stakeholder conflicts are influential so that the overall process can be improved to avoid or manage such conflicts. Chi and Holsapple (2005) propose a model of stakeholder collaboration in B2B highlighting three behavioural processes: knowledge sharing, participative decision-making and conflict governance. Our study provides support for the importance of such processes.

The findings and discussion above also suggest a need for supporting infrastructure. Electronic marketplaces offer a recognised structure which may be useful to support both collaborative IT support projects and specific supply chains (Markus & Christiaanse, 2003). However there are also challenges in using electronic marketplaces successfully. In a recent case study of an electronic marketplace, the key influences affecting its success were the loss of social capital, nature of communication channels used, time taken to reach critical mass, and power imbalances among participants (Driedonks et al., 2005).

The risk of reduced service quality was identified in the practice of customer co-contribution to service. This concern was noted more generally for online service provision by Schultze and Bhappu (2005). There may be ways for IT service firms to better monitor customer contributions via user fora and capture the high quality solutions (which currently are not captured). Related questions of knowledge ownership must be resolved. However regulatory issues at different geographic locations cloud the resolution of such questions.

Indeed there were several areas identified where IT service providers suffer from inadequate regulatory support. The security, privacy and IP issues experienced in the often-offshore managed IT support environment highlight the need for clarification and awareness of relevant regulations and laws at industry, national and international levels. IT support companies also seek to understand how to share knowledge with business competitors within strategic alliances while maintaining a competitive advantage. New theories are sought to underpin such knowledge sharing in an increasingly collaborative global business environment.

The six IT service provider companies in our study uniformly noted that the key objective for use of a WSS for after-sales IT support was to increase customer satisfaction while reducing support costs. This objective was expected to be achieved by relational as well as transactional methods. However as the study showed, the successful accomplishment of both goals can be problematic (also found by Bunduchi, 2005). Research is beginning to appear on the enabling of relationships in service provision when the internet is used to facilitate service provision. In a business to consumer (B2C) context, Sigala (2007) explored online travel service provision and found that the communication aspect of the online service played a key role in relationship development. Where provider-consumer communication was enhanced by use of relevant communication tools, relationships and client satisfaction were improved. Thus IT support organisations may find that the path to satisfying the relational aspect of WSS success is by better leveraging the internet's communication tools and customising communication-oriented content.

Some of the challenges identified in our study relate to the process of knowledge transfer. For example, it was noted that a customer firm end-user may not be interested in institutionalizing transferred IT solutions. Cranefield and Yoong (2005), in identifying key challenges in inter-organisational knowledge transfer, highlighted the need for a fit between the knowledge received and the receiving organisation's objectives. If a customer firm does not prioritise the institutionalization of IT solutions throughout the firm, its employees will not make the effort to learn the solutions transferred from the service provider and will simply apply them to resolve the initial problem. This suggests a need for IT service providers to educate their customers about the importance of institutionalizing IT solutions to their organisation. Nieminen (2005) noted that for knowledge transfer to take place, a receiving organisation must be capable of absorbing shared knowledge. Such capability may be missing from the customer firms which are receiving IT solutions.

CONCLUSION

This article has identified and discussed eight major stakeholder-oriented challenges in the provision of managed after-sales IT support services via WSSs to enterprise customers, from the perspective of multinational IT service providers. Specifically:

- All stakeholder viewpoints should be considered by an IT support organisation when planning, implementing and managing a WSS for after-sales IT support provision;
- Different stakeholder types might interpret the requirements for CSFs differently;
- The six CSF categories identified apply to *all* three stakeholder organisations, however a support organisation will have very little control over whether a customer organisation and business partner organisation do, in fact, address these categories. Thus performance measurement and management of WSS strategies in a B2B context will be difficult;
- It is problematic whether relationships are possible, or even enabled by a WSS;
- Stakeholder relationships can be more complex than Figure 1 allows;
- Intellectual property, security and privacy issues can be problematic for stakeholder-oriented reasons;
- Understanding the complexities of the business partner alliance can be an issue; and
- Managing customer contributions to service is an important challenge.

The findings demonstrate that best-in-class IT service providers face diverse challenges to better understand and resolve potential conflicting stakeholder needs in this context.

Theoretically this article provides numerous insights into the key challenges faced by IT service providers, relating to the different stakeholders, in the provision of B2B after-sales IT support services via WSSs. The article also highlights a need for new theories which integrate WSS strategies across the multiple stakeholders involved. Our findings further suggest that relationship development by WSSs is poorly understood by the companies involved and that further research is needed to develop new understandings.

This article also assists IT service providers by recommending that all stakeholder viewpoints and issues should be considered when planning, implementing and managing WSSs in the managed after-sales IT support context. Addressing diverse stakeholder needs may be particularly challenging in some areas such as security, privacy and IP. Such emerging challenges highlight the complexities of working with business partners which are also considered competitors. Greater collaboration is needed with better supporting infrastructure and regulation.

While the findings from this article are limited by the context (managed after-sales IT support) and scope (the IT support organisation perspective only was studied), they are indicative of possible concerns that other types of service providers may have in offering supplementary E-services using WSSs. Thus the findings provide a foundation for exploration in other settings. Investigating the customer and business partner perspectives would also provide valuable balance to the views expressed and analysed in this article.

ACKNOWLEDGMENT

This article is a significantly extended and updated version of a article published in the *Proceedings of the 20th Bled eConference (Bled07)*, Bled, Slovenia, 4-6 June 2007. The authors are grateful to the anonymous reviewers whose valuable advice helped improve the quality of this article.

REFERENCES

Barnes, J.G., Dunne, P.A. and Glynn, W.J. (2000). Self-service and Technology: Unanticipated and Unintended Effects on Customer Relationships. Handbook of Services Marketing and Management, Swartz, T.A. and Iabocobucci, D. (eds.), Sage Publications, Thousand Oaks, California, 89-102.

Barua, A., Konana, P., Whinston, A.B. and Yin, F. (2004). An Empirical Investigation of Net-enabled Business Value. *MIS Quarterly*. 28(4), December, 585-620.

Benapudi, N. and Leone, R.P. (2003). "Psychological Implications of Customer Participation in Co-production", *Journal of Marketing.* 67(1), January, 14-28.

Bhappu, A.D. and Schultze, U. (2006). The Role of Relational and Operational Performance in Business-to-Business Customers' Adoption of Self-Service Technology. *Journal of Service Research*. 8(4), 372 – 385.

Brohman, M.K., Watson, R.T., Piccoli, G. and Parasuraman, A. (2003). Data Completeness: A Key to Effective Net-based Customer Service Systems, *Communications from the Association for Computing Machinery*. 46(6), 47-51.

Bunduchi, R. (2005). Business Relationships in Internet-based Electronic Markets: The Role of Goodwill Trust and Transaction Costs. *Information Systems Journal*. 15 (4), 321–341.

Cameron, J. (2005). Ten Concepts for an eBusiness Collaborative Project Management Framework. In *Proceedings of eIntegration in Action: 18th Bled eConference*, June 6-8 2005, Bled, Slovenia.

Cenfetelli, R., Benbasat, I. and Al-Natour, S. (2005). Information technology mediated customer service: A functional perspective. In *Proceedings of ICIS 2005*, Las Vegas, USA.

Chi, L. and Holsapple, C. (2005). Understanding Computer-mediated Interorganisational Collaboration: A Model and Framework. *Journal of Knowledge Management*. 9(1), 53-75.

Chua ,C.E.H., Khoo, H.M., Straub, D.W., Kadiyala, S. and Kuechler, D. (2005). The Evolution of E-commerce Research: A Stakeholder Perspective. *Journal of Electronic Commerce Research*. 6(4), 262-280.

Cooper, V.A. (2007). Knowledge Transfer in Enterprise IT Support Provision using Web-based Self-Service, unpublished PhD Thesis, Deakin University, Melbourne, Australia.

Cooper, V.A., Lichtenstein, S. and Smith, R. (2006a). Enabling the Transfer of Information Technology Support Knowledge to Enterprise Customers Using Web-based Self-service systems: Critical Success Factors from the Support Organisation Perspective, in *Proceedings of Seventeenth Australasian Conference on Information Systems (ACIS 2006)*, Adelaide, Australia.

Cooper, V., Lichtenstein, S. & Smith, R. (2006b) Knowledge transfer in enterprise information technology support using web-based self-service systems, *International Journal of Technology Marketing*, 1(2), 145-170.

Cooper, V.A., Lichtenstein, S. & Smith, R. (2005) Emerging Issues in After-sales Enterprise Information Technology Support Using Web-based Self-service Systems, in *Proceedings of Sixteenth Australasian Conference on Information Systems (ACIS 2005)*, University of Technology Sydney, 30 November – 2 December, 2005, Sydney.

Cranefield, J. and Yoong, P. (2005). Organisational Factors Affecting Inter-organisational Knowledge Transfer in the New Zealand State Sector: A Case Study. *The Electronic Journal for Virtual Organisations and Networks.* 7, December.

[CRMInd] CRMIndustry.com (2006). 2006 State of IT Outsourcing. Summary report, <http://www.crmindustry.com/industry_research/outsourcing.htm> (Accessed 03 June 2008).

[CSI] Consortium Service Innovation (2002). Getting Started with KCS. Official Consortium Service Innovation Site, <http://www.thinkhdi.com/files/pdfs/GettingStartedKCS.pdf> (Accessed 03 June 2008).

Dahlberg, T. and Nyrhinen, M. (2006). A New Instrument to Measure the Success of IT Outsourcing. In *Proceedings of the 39th Hawaii International Conference on System Sciences (HICSS'2006)*, IEEE Society Press.

Doyle, S. (2007). Self-service Delivery and the Growing Roles of Channels. *The Journal of Database Marketing & Customer Strategy Management*. 14(2), January , 150-159.

Driedonks, C., Gregor, S. and Wassenaar, A. (2005). Economic and Social analysis of the Adoption of B2B Electronic Marketplaces: A Case Study in the Australian Beef Industry. *International Journal of Electronic Commerce*. 9(3), Spring, 49-72.

Ferdinand, J. and Simm, D. (2007). Re-theorizing External Learning: Insights from Economic and Industrial Espionage. *Management Learning.* 38(3), 297-317.

Geib, M., Kolb, L. and Brenner, W. (2006). Collaborative Customer Management. Financial Services Alliances. In (Fjermestad, J. and Romano, N.C. (eds)), *Electronic Customer Relationship Management*, Armonk, New York.

Gong, H., Tate, M. and Alborz, S. (2007). Managing the Outsourcing Marriage to Achieve Success. In *Proceedings of Hawaii International Conference on the System Sciences (HICSS'07)*, IEEE Society Press.

Gutek, B., Bhappu, A.D., Liao-Troth, M.A. and Cherry, B. (1999). Distinguishing between Service Relationships and Encounters. *Journal of Applied Psychology.* 84(2), 218-233.

Hatch, M.J. (1997). *Organisation Theory: Modern, Symbolic and Postmodern Perspectives.* Oxford:UK, Oxford University Press.

Hofacker, C. F., Goldsmith, R.G., Bridges, E. and Swilley, E. (2007). E-Services: A Synthesis and Research Agenda. *Journal of Value Chain Management*. 1(1/2), 13-44.

Holsapple, C.W. and Jin, H. (2007). Connecting Some Dots: E-commerce, Supply Chains and Collaborative Decision Making. *Decision Line*. 38(5), October, 14-21.

Kakabadse, N.K., Kakabadse, A. and Kouzmin, A. (2003). Reviewing the Knowledge Management Literature: Towards a Taxonomy. *Journal of Knowledge Management Practice*. 7(4), 75-91.

Kandampully, J. (2003). B2B Relationships and Networks in the Internet Age. *Management Decision*. 41(5), 443-451.

Khalifa, M. and Shen, N. (2005). Effects of Electronic Customer Relationship Management on Customer Satisfaction: A Temporal Model. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)*, IEEE Society Press.

Knight, L. (2002). Network Learning: Exploring Learning by Interorganizational Networks. *Human Relations*. 55(4), 427-454.

Koh, C., Ang, S. and Straub, D.W. (2004). IT Outsourcing Success: A Psychological Contract Perspective. *Information Systems Research*. 15(4), 356-373.

Kumar, K. and van Dissel, H.G. (1996). Sustainable Collaboration: Managing Conflict and Cooperation in Interorganisational Systems. *MIS Quarterly*, 20(3), 279-300.

Kurnia, S. and Johnston, R. (2002). A Review of Approaches to EC-enabled IOS Adoption Studies. In *Proceedings of the 35th Annual Hawaii Conference on System Sciences (HICSS'02)*, IEEE Society Press.

Lane, P.J. and Lubatkin, M. (1998). Relative Absorptive Capacity and Inter-organisational Learning, *Strategic Management Journal*. 19(5), 461-477.

Lei, D.T. (1997). Competence-building, Technology Fusion and Competitive Advantage: The Key Roles of Organisational Learning and Strategic Alliances. *International Journal of Technology Management*. 14(2/3/4), 208-37.

Levenburg, N.M. and Klein, H.A. (2006). Delivering customer services online: identifying best practices of medium-sized enterprises. *Information Systems Journal*. 16(2), 135-155.

Markus, M.L. and Christiaanse, E. (2003). Adoption and Impact of Electronic Marketplaces. *Information Systems and E-Business Management*. 1(2), January, 139-155.

Mayring, P. (2000). Qualitative Content Analysis, Forum Quality Social Research, (1)2. <http://www.qualitative-research.net/fqs-texte/2-00/2-00mayring-e.htm> (Accessed 3 June 2008).

Mitroff, I.I. (1983). *Stakeholders of the Organisational Mind*. Jossey-Bass. San Francisco, CA.

Meuter, M.L., Ostrom, A., Bitner, M.J. and Roundtree, R. (2003). The Influence of Technology Anxiety on Consumer Use and Experiences with Self-service Technologies. *Journal of Business Research*. 56(11), 899-906.

Negash, S., Ryan, T. and Igbaria, M. (2003). Quality and Effectiveness in Web-based Customer Support Systems. *Information and Management*. 40(8), September, 757-768.

Neuman, W.L. (2006). *Social Research Methods: Qualitative and Quantitative Approaches*, Sixth Edition. Allwyn & Bacon, Boston, Massachusetts.

Nieminen, H. (2005). Organisational Receptivity – Understanding the Inter-organisational Learning Ability, *Electronic Journal of Knowledge Management*. 3(2), 107-118.

Otim, S. and Grover, V. (2006). An Empirical Study on Web-based Services and Customer Loyalty. *European Journal of Information Systems*. 15(6), 527-541.

Pan, G.S.C. (2005). Information Systems Project Abandonment: A Stakeholder Analysis. *International Journal of Information Management*. 25(2), 173-184.

Peppard, J. (2001). Bridging the Gap Between the IS Organisation and the Rest of the Business: Plotting a Route. *Information Systems Journal*. 11(3), 249-260.

Peppers, D. and Rogers, M. (2001). *One to One B2B: Customer Development Strategies for the Business-to-Business World*, Currency Doubleday, New York.

Piccoli, G., Brohman, M.K., Watson, R.T. and Parasuraman, A. (2004). Net-based Customer Service Systems: Evolution and Revolution in Web-site Functionalities. *Decision Sciences*. 35(3), 423-455.

Pitt, L.F., Berthon, P. and Lane, N. (1998). Gaps Within the IS Department: Barriers to Service Quality. *Journal of Information Technology*. 13(3), 191-200.

Pujari, D. (2004). Self-service with a Smile? Self-service Technology (SST) Encounters Among Canadian Business-to-Business. *International Journal of Service Industry Management*. 15(2), 200-219.

Ragsdale, J. (2007). Self-Service Success Continues to Decline; How Web 2.0 Can Help, Service and Support Professional Association, US.

Rai, A. and Sambamurthy, V. (2006). Editorial Notes – The Growth of Interest in Services Management: Opportunities for Information Systems Scholars. *Information Systems Research*. 17(4), 327 – 331,

Reichheld, F.F. and Schefter, P. (2000). E-Loyalty: Your Secret Weapon on the Web. *Harvard Business Review*. 78(4), July-August, 105-112.

Ritter, T. and Gemunden, H.D. (2003). Interorganisational Relationships and Networks: an Overview. *Journal of Business Research*. 56, 691-697.

Rockart, J.F. (1979). Chief Executives Define their own Data Needs. *Harvard Business Review*. 57(2), March-April, 81-93.

Rust, R.T. (2001). The Rise of E-Service, *Journal of Service Research*, 3(4), May, 283-284.

Schultze, U. and Bhappu, A.D.P. (2005). Incorporating Self-Serve Technology into Co-production Designs. *International Journal of E-Collaboration*. 1(4), 1-23.

Schultze, U. and Orlikowski, W.J. (2004). Practice Perspective on Technology-mediated Network Relations: The Use of Internet-based Self-serve Technologies. *Information Systems Research*. 15(1), 87-106.

Sigala, M. (2007). Investigating the Internet's Impact on Interfirm Relations: Evidence from the Business Travel Management Distribution Chain. *Journal of Enterprise Information Management*. 20(3), 335-355.

Singh, M. and Byrne, J. (2005). Performance Evaluation of e-Business in Australia. *Electronic Journal of Information Systems Evaluation*. 8(2).

Solitander, M. (2006). Balancing the Flows: Managing the Intellectual Capital Flows in Interorganisational Projects. *Electronic Journal of Knowledge Management*. 4(2), 197-206.

SSPA (2004). *2005 Support Demand Research Series***.** Tech Strategy Partners and Service & Support Professionals Association**.**

Szulanski, G. (1996). Exploring Internal Stickiness: Impediments to the Transfer of Best Practice Within the Firm. *Strategic Management Journal*. 17, Special Issue, 27-43.

Tafti, M.H.A. (2005). Risk Factors Associated with Offshore IT Outsourcing. *Industrial Management & Data Systems*. 105(5), 549-560.

Vlachopoulou, M. and Manthou, V. (2003). Partnership Alliances in Virtual Markets. *International Journal of Physical Distribution and Logistics Management*. 33(3), 254-267.

Xu, L. (2007). Outsourcing and Multi-party Business Collaborations Modelling. *Journal of Electronic Commerce in Organisations*. 5(2), 77-96.

Yin, R.K. (2003). *Case Study Research, Design and Methods*, Fourth Edition. Sage Publications, Newbury Park.

APPENDIX.

CSFs for Knowledge Transfer from a Support Organisation to a Customer Organisation using WSSs (A)

Critical Success Factor	Description
CSF-1 Cost Effectiveness	The cost equation for providing/using web-based self-service must be better, or at least not worse, than providing/using non-web-based self-service.
CSF-2: Provision of Additional Services and Cross-Selling Opportunities	Current WSS transactions are used proactively as an opportunity to offer the customer organisation additional advice and services
CSF-3: Critical Mass: Knowledge Content and Knowledge Contributors	A sufficient number of end-users must proactively contribute sufficient knowledge content to the WSS knowledge base, to encourage all parties to initially use, and to continue to use, the WSS as a means of resolving their support issues or information requirements
CSF-4: Usefulness: Provision of Knowledge Which Meets User Requirements	The WSS must provide the functionality and knowledge required to meet the objectives of all stakeholders. For example, for the end-user customer, it should resolve a specific technical or business problem, or provide other required knowledge resources.
CSF-5: Ability to Provide Efficiency	Use of the WSS to resolve a support issue or provide other knowledge resources must be perceived as efficient by all parties. This is inclusive of not only the performance of the WSS tool but the surrounding processes for using the WSS.
CSF-6: Access, Connectivity, Availability and Performance	The providing organisation, relevant business partners and the customer organisation must have sufficient technology infrastructure in place, to enable all parties to participate in web-based self-service.
CSF-7: Effective Information Architecture and Search Engine	The WSS must have an effective Information Architecture and Search Engine such that the information system that organises and retrieves knowledge in the knowledge base is perceived as effective by end-users.
CSF-8: Security, Privacy and Assurance	All stakeholders using WSS must feel secure, private and confident in all aspects of WSS transactions including the stored data components of transactions. Issues surrounding information security and information privacy, and the need to keep confidential related company secrets (intellectual property) must be addressed.
CSF-9: Ease of Use/Usability	An end-user must perceive that use of the WSS does not demand excessive cognitive or ergonomic effort.
CSF-10: Early Positive Experience	The first few end-user experiences with the WSS must result in a positive outcome, where end-user needs are met and they feel valued, in order for the end-user to adopt WSS long term.
CSF-11: Positive Experience	Using the WSS on an ongoing basis must result in a positive outcome, where corporate customer needs and all types of end-users' needs are met and they also feel valued. A positive experience is closely related to customer organisation/end-user satisfaction
CSF-12 Confidence in Solution	The customer organisation/end-user must feel confident that the solution provided by the WSS will resolve their issue and will not result in further issues. They must also have self-confidence in their own ability to apply the offered solution.
CSF-13: Customer Focus: Understand Customer and their Requirements	The support organisation (and relevant business partners) must understand the individual business and technical needs of individual customer organisations and their end-users. With this understanding, WSS must be tailored to meet those individual needs.

APPENDIX.

CSFs for Knowledge Transfer from a Support Organisation to a Customer Organisation using WSSs (B)

Critical Success Factor	Description
CSF-14: Positive Relationship	The relationship between the support organisation, business partners and the customer organisation must be one which supports open communication and trust. This positive relationship should exist at both the corporate and end-user levels.
CSF-15: Provision of Additional Support: Education & Training	Additional assistance, or education and training in respect to how to use the WSS must be provided by the support organisation as requested by end-users.
CSF-16: Employee Focus	Management within the support organisation, business partner and customer organisations must have an understanding of the work processes and conditions which will affect the ability and willingness of employees to adopt the WSS and associated strategies. With this understanding, management must focus on meeting the needs of their employees where possible, in order to maximise employee productivity and the benefits received from the WSS strategy.
CSF-17: Culture	The support organisation should foster an environment that recognises that WSS is part of the way it wants to conduct business. In addition, an open, sharing culture is needed. The culture should extend to customer organisations and business partners.
CSF-18: Marketing and Awareness of Web-based Service	Marketing programs which raise awareness of and support for, the adoption of WSS, must be in place.
CSF-19: Knowledge Creation, Capture and Re-Use	Knowledge capture processes to ensure that valuable knowledge is created and captured into the WSS knowledge base by end-users must be in place. Knowledge reuse processes to ensure that this knowledge is subsequently accessed and re-used by end-users, must also be in place.
CSF-20: Knowledge Validation	Processes must be in place to ensure the accuracy of the knowledge which is captured into the WSS knowledge base and to ensure that once it is captured, it is frequently reviewed and updated for currency.
CSF-21: Knowledge Storage/Retrieval	Processes must be in place to ensure that the structure and format of captured knowledge facilitate findability.
CSF-22 : Presentation of Knowledge	The knowledge must be presented in a form which maximizes the understanding acquired by end-users.
CSF-23: Measurement & Feedback of WSS	Sufficient measurement and feedback methods for assessing the effectiveness of the WSS strategy must be in place.
CSF-24: Alignment and Integration	There must be alignment and integration between the WSS and other channels' support processes, as well as with related business processes, in the context of the business/industry environment.
CSF-25: WSS Override and Recovery	The capability for the end-user or WSS to over-ride transactions initially made via the WSS, must be in place, whereby if an end-user is not finding a satisfactory resolution via the WSS, the transaction is directed to an alternative mode of service delivery (e.g. a chat session or telephone call).
CSF-26: Ease of Re-initiation	A process must be in place whereby an end-user can easily re-initiate a support transaction to re-locate a previously retrieved resolution or other knowledge resource.
CSF-27 Top Management Support	Top management must provide ongoing support and commitment to the WSS and associated strategies.

This work was previously published in International Journal of E-Services and Mobile Applications, Vol. 1, Issue 1, edited by A. Scupola, pp. 1-20, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 4.12
Mailing Lists and Social Semantic Web

Sergio Fernández
Fundación CTIC, Spain

Diego Berrueta
Fundación CTIC, Spain

Lian Shi
Fundación CTIC, Spain

Jose E. Labra
University of Oviedo, Spain

Patricia Ordóñez de Pablos
University of Oviedo, Spain

ABSTRACT

Electronic Mailing lists are a key part of the Internet. They have enabled the development of social communities who share and exchange knowledge in specialized and general domains. In this chapter the auhtors describe methods to capture some of that knowledge which will enable the development of new datasets using Semantic Web technologies. In particular, the authors present the SWAML project, which collects data from mailing lists. They also describe smushing techniques that normalize RDF datasets capturing different resources that identify the same one. They have applied those techniques to identify persons through the mailing lists of open source communities. These techniques have been tested using a dataset automatically extracted from several online open source communities.

DOI: 10.4018/978-1-60566-272-5.ch004

INTRODUCTION

Early forms of electronic mailing lists were invented almost as soon as electronic Mail (e-Mail) and are a cornerstone of Internet, allowing a lot of people to keep up to date on news related with their interests. Besides direct messaging between individuals, mailing lists exist as private or public forums for information exchange in communities with shared interests. Mailing list archives are compilations of the previously posted messages that are often converted into static HTML pages for their publication on the web. They represent a noteworthy portion of the contents that are indexed by web search engines, and they capture an impressive body of knowledge that, however, is difficult to locate and browse.

The reason for this difficulty can be traced back to the translation procedure that run to transform the e-mail messages into static HTML pages. This task is fulfilled by scripts that create static HTML pages for each message in the archive. In addition, some indexes (by date, by author, by thread) are generated and usually split by date ranges to avoid excessive growth.

On the one hand, this fixed structure reduces the flexibility when users explore the mailing list archives using their web browsers. On the other hand, most of the meta-data that were associated to each e-mail message are lost when the message is rendered as HTML for presentational purposes.

We propose to use an ontology and RDF (Resource Description Framework, Klyne 2004) to publish the mailing list archives into the (Semantic) Web, retaining the meta-data that were present in the messages. Additionally, by doing so, the information can be merged and linked to other vocabularies, such as FOAF (Brickley and Miller, 2005).

The rest of the chapter is organized as follows: in section 2 we describe the main developments of Social Semantic Web related with mailing lists. In section 3, we explain several techniques to collect RDF datasets from mailing lists and other social sources. Section 4 contains a description of the SWAML project that collects those RDF datasets from mailing lists. In section 5, we describe several applications that consume that data. In section 6, we discuss some experiments that we have done over those datasets. Finally, in section 7 we present some conclusions and future work.

SOCIAL SEMANTIC WEB

The Semantic Web vision tries to develop new ways to integrate and reuse the information published on the web. To that end, the W3C has developed several technologies, like RDF, which enable to add metadata descriptions that contain meaningful values and global properties to resources. The resulting metadata forms a graph model which can be easily linked with other graphs (Berners-Lee, 2006) incrementing the knowledge represented by the original graph. Those values and properties formalize the knowledge of a particular. In 2004, the W3C consortium developed OWL (Patel-Schneider et al, 2004), a web ontology language which facilitates the definition of those formalizations, called ontologies. Based on description logics, OWL has been adopted as the standard ontology language with several available editors, reasoners and tools. There have been also a number of ontologies developed in OWL for different purposes and with different level of detail, from generic to domain-specific ones.

On the other hand, in the last years, the concept of Web 2.0 has attracted a lot of interest. One of the key aspects of Web 2.0 applications is the social part of the web. Users are not considered as mere consumers of information, but also as producers. People want to share knowledge, establish relationships, and even work together using web environments. It is necessary to develop people-oriented web technologies which can represent people interests and that enable the integration and reuse of people related information in the same way that the semantic web vision advocates. These technologies can be seen as social semantic web and we expect that there will be more and more applications making use of them.

One of the first developments is the FOAF vocabulary, which represents basic properties of people, like their name, homepage, etc. as well as the people they know. FOAF descriptions are very flexible and can be extended to other domains. There are already web portals which export their user profiles in FOAF format and the number of FOAF applications is increasing.

Apart from FOAF, there are other ontologies related to the social semantic web. In particular, SIOC (Semantically-Interlinked Online Communities), provides a vocabulary to interconnect different discussion methods such as blogs, web-based forums and mailing lists (Breslin 2005, Breslin 2006). Although we will apply mainly SIOC to mailing-lists, it has a wider scope than just mailing lists, and generalizes all kinds of online discussion primitives in the more abstract sioc:Forum concept. Each forum represents an online community of people that communicate and share a common interest. The goal of SIOC is to interconnect these online communities.

Other relevant concepts of the ontology are sioc:User and sioc:Post, which model respectively the members of the communities and the content they produce. Instances of these three classes (forums, users and posts) can be linked together using several properties.

The SIOC ontology was designed to express the information contained both explicitly and implicitly in Internet discussion methods. Several software applications, usually deployed as plug-ins, are already available to export SIOC data from some popular blogging platforms and content management systems. The effort, however, is focused on web-based communities (blogs, discussion forums), while little has been done so far to extend the coverage to legacy non-web communities, such as mailing lists and Usenet groups.

SIOC classes and properties are defined in OWL, and their instances can be expressed in RDF. Therefore, they can be easily linked to other ontologies. The obvious choice here is FOAF, which provides powerful means to describe the personal data of the members of a community.

Mailing lists can be easily described by instantiation of the SIOC classes and properties. Each mailing list can be represented by an instance of sioc:Forum (a subclass of Forum might be used instead, although it is not required). Messages sent to the list and their replies become instances of sioc:Post.

Finally, people involved into the list are instances of sioc:User. The SIOC ontology provides a property to link forums and users, namely sioc:has_subscriber. We argue that being subscribed to a mailing list is just one of the roles a user can play with respect to a forum. Moreover, the list of subscribers is often available only to the system administrator for privacy reasons. On the other hand, it is easy to collect the set of people who post to the list, i.e., the people actively involved in the forum. Depending on the settings, the latter may be a subset of the former, in particular in those mailing lists that forbid posting privileges to non-subscribers. Ideally, these two different semantics would be captured using new properties. However, for practical reasons, and to avoid privacy issues, we consider just the already existent sioc:has_subscriber property, and we populate it with the set of active members of a forum. Consequently, inactive members of the forum remain hidden, but this does not represent a problem due to the open world assumption.

Additionally, the Dublin Core (Dublin Core Metadata Element Set, Version 1.1, 2006) and Dublin Core Terms vocabularies are used to capture meta-data such as the message date (dcterms:created) and title (dc:title).

Given the distributed nature of RDF, it is expected that there will be different RDF datasets describing aspects of the same resources. The term *smushing* has been defined as the process of normalizing an RDF dataset in order to unify *a priori* different RDF resources which actually represent the same thing. The application which executes a *data smushing* process is called a *smusher*. The process comprises two stages:

First, redundant resources are identified; then, the dataset is updated to reflect the recently acquired knowledge. The latter is usually achieved by adding new triples to the model to relate the pairs of redundant resources. The OWL property owl:sameAs is often used for this purpose, although other properties without built-in logic interpretations can be used as well (e.g.: ex:hasSimilarName). Redundant resources can be spotted using a number of techniques. In this chapter, we explore two of them: (1) using logic inference and (2) comparing labels.

COLLECTING DATA INTO THE SOCIAL SEMANTIC WEB

Since SIOC is a recent specification, its adoption is still low, and only a few sites export SIOC data. There exist a number of techniques that can be used to bootstrap a network of semantic descriptions from current social web sites. We classify them in two main categories: intrusive and non-intrusive techniques.

On the one hand, methods which require direct access to the underlying database behind the social web site are **intrusive** techniques. The web application acts as the controller and publishes different views of the model in formats such as HTML and RSS. In terms of this pattern, publishing SIOC data is as simple as adding a new view. From a functional point of view, this is the most powerful scenario, because it allows a lossless publication due to the direct access to the back-end database. The SIOC community has contributed a number of plugins for some popular web community-building applications, such as Drupal, WordPress and PhpBB2. Mailing lists are also covered by SWAML, which is described in the next section. There is, however, a major blocker for this approach. All these software components need a deployment in the server side (where the database is). This is a burden for system administrators, who are often unwilling to make a move that would make it more difficult to maintain, keep secure and upgrade their systems. This is particularly true when there is no obvious immediate benefit of exporting SIOC data.

On the other hand, methods which do not require direct access to the database and can operate on resources already published on the web are **non-intrusive**. One technique is the use of cooked HTML views of the information, the same ones that are rendered by web browsers for human consumption. An example could be RSS/Atom feeds, which have become very popular in the recent years. They can be easily translated into SIOC instances using XSLT stylesheets (for XML-based feeds) or SPARQL queries (for RSS 1.0, which is actually RDF). Unfortunately, these feeds often contain just partial descriptions. Another technique is the use of public APIs. The Web 2.0 trend has pushed some social web sites to export (part of) their functionality through APIs in order to enable their consumption by third-party mash-ups and applications. Where available, these APIs offer an excellent opportunity to create RDF views of the data. A shared aspect of these sources is their ubiquitous availability through web protocols and languages, such as HTTP and XML. Therefore, they can be consumed anywhere, and thus system administrators are freed of taking care of any additional deployment. In contrast, they cannot compete with the intrusive approaches in terms of information quality, as their access to the data is not primary.

SWAML PROJECT

SWAML (Fernández et al, 2008) is a Python tool that reads mailing list archives in raw format, typically stored in a “mailbox” (or “mbox”), as defined in RFC 4155 (Hall 2005). It parses mailboxes and outputs RDF descriptions of the messages, mailing lists and users as instances of the SIOC ontology. Internally, it re-constructs the structure of the conversations in a tree structure, and it exploits this structure to produce links between the posts. This script is highly configurable

Figure 1. SIOC post example in RDF/XML

```
<rdf:RDF
  xmlns:dcterms='http://purl.org/dc/terms/'
  xmlns:sioc='http://rdfs.org/sioc/ns#'
  xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
  xmlns:dc='http://purl.org/dc/elements/1.1/'
  xml:base='http://example.org/swaml-devel/'>
  <sioc:Post rdf:about="2006-Sep/post-52">
    <dc:title>Re: [swaml-devel] Changing SWAML ontology</dc:title>
    <sioc:has_creator rdf:resource="subscriber/s10"/>
    <dcterms:created>Wed, 6 Sep 2006 20:14:44 +0200</dcterms:created>
    <sioc:content><!-- ommitted --></sioc:content>
    <sioc:has_reply rdf:resource="2006-Sep/post-69"/>
    <sioc:previous_by_date rdf:resource="2006-Sep/post-51"/>
    <sioc:next_by_date rdf:resource="2006-Sep/post-53"/>
  </sioc:Post>
</rdf:RDF>
```

and non-interactive, and has been designed to be invoked by the system task scheduler. This low-coupling with the software that runs the mailing list eases its portability and deployment.

SWAML could be classified as an intrusive technique because it requires access to the primary data source, even if in this case it is not a relational database but a text file (for instance, the approach followed by mle (Michael Hausenblas at al., 2007) is considered completely non-intrusive). Anyway, it is worth mentioning that some servers publish these text files (mailboxes) through HTTP. Therefore, sometimes it is possible to retrieve the mailbox and build a perfect replica of the primary database in another box. In such cases, SWAML can be used without the participation of the system administration of the original web server.

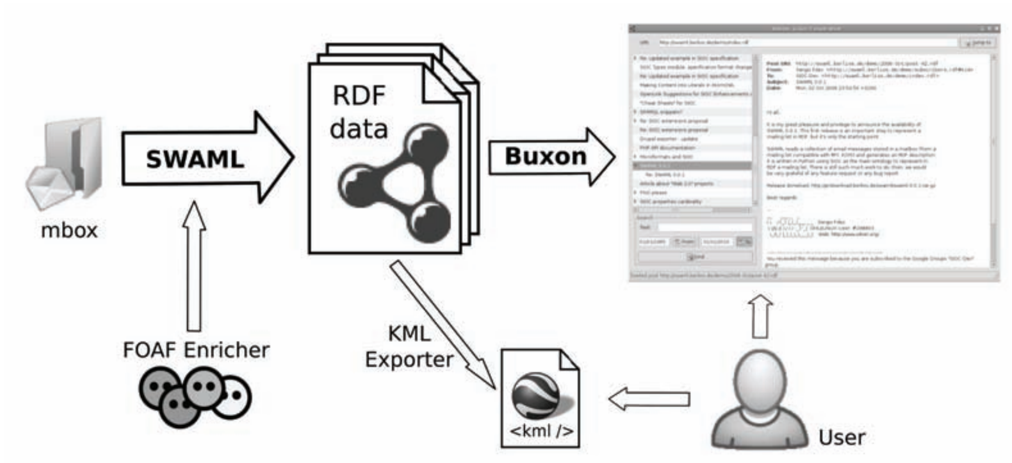
There are many ways in which a mailing list message might be related with other messages. However, we consider just two scenarios. The first one links a post with its replies (sioc:has_reply). Actually, due to sequential layout of the messages in the most widely used format to store mailing list archives (mailbox), it is easier to generate the inverse property (sioc:reply_of). Anyway, the has_reply property can be generated either by a description logics reasoner or by performing two passes over the sequence.

The second link among messages is established between a post and its immediate successor (or predecessor) in chronological order. It is worth to note that this link is not strictly necessary, because the following (or preceding) message can be obtained by sorting by date the sequence of posts. However, this is a rather expensive operation, because the whole set of posts is required in order to perform the sorting. The open world assumption makes this query even more challenging. Therefore, considering that browsing to the previous or next message is a common use case, and the complete set of posts can be very large or even unavailable, we introduced two new properties, next_by_date and prev_by_date. These properties where eventually accepted into the SIOC ontology. An RDF representation of a sample message is shown in Figure 1.

SWAML is essentially a mailbox parser and translator implemented in Python. Its output is a number of SIOC instances (Forum, Posts and Users) in a set of RDF files. SWAML can be invoked by the system task scheduler.

Parsing the mailbox and rebuilding the discussion threads may be sometimes tricky. Although each mail message has a supposedly unique identifier in its header, the Message-ID, defined by RFC 2822 (Resnick, 2001), in practice its unique-

Figure 2. Buxon is an end-user application that consumes sioc:Forum instances, which in turn can be generated from mailboxes using SWAML.



ness cannot be taken for granted. Actually, we have found some messages with repeated identifiers in some mailing lists, probably due to non-RFC compliant or ill-configured mail transport agents. Therefore, SWAML assumes that any reference to a message (such as those created by the In-Reply-To header) is in fact a reference to the most recent message with that ID in the mailbox (obviously, only previous messages are considered). Using this rule of thumb, SWAML builds an in-memory tree representation of the conversation threads, so sioc:Posts can be properly linked.

Actually, SWAML goes further than just a format-translation tool. A dedicated subroutine that runs as part of the batch execution but may be also separately invoked on any sioc:Forum, tries to find a FOAF description for each sioc:User.

One important requirement of the semantic web is to be an extension (and not a replacement) of the current document-based web. Ideally, each user agent must be able to retrieve the information in their format of choice. For instance, current web browsers prefer (X)HTML documents, because they can be rendered and presented to the end user. However, semantic web agents require information to be available in a serialized RDF format, such as RDF/XML or N3. Furthermore, different representations of the same information resource should share a unique URI. Fortunately, the HTTP protocol supports this feature by using "content-negotiation". Clients of the protocol can declare their preferred formats in the headers of an HTTP request using the Accept header. Web servers will deliver the information in the most suited available format, using the Content-type header of the HTTP response to specify the actual format of the returned delivered content. MIME types such as text/html and application/rdf+xml are used as identifiers of the requested and available formats.

Setting up the content negotiation in the server-side usually requires some tuning of the web server configuration. It also depends on some choices made by the publisher of the information, such as the namespace scheme for the URIs or the fragmentation of the information. In (Miles et al, 2006) there is a list of some common scenarios, which are described to great detail, and configuration examples for the Apache web server are provided. The most suitable scenarios (or recipes, as they are called) to publish mailing list metadata are the fifth and sixth, i.e., multiple documents available both in HTML and RDF.

Figure 3. A sample htaccess configuration file for Apache generated by SWAML. These two rules redirect the request to the proper file based on the content negotiation field of the HTTP request. Some lines have been wrapped for readability.

```
RewriteEngine On
RewriteBase /demos/swaml-devel/
AddType application/rdf+xml .rdf
Options -MultiViews

RewriteCond %{HTTP_ACCEPT} text/html [OR]
RewriteCond %{HTTP_ACCEPT} application/xhtml\+xml [OR]
RewriteCond %{HTTP_USER_AGENT} ^Mozilla/.*
RewriteRule ^/([0-9]{4})-([A-Za-z]+)/post-([0-9]+)$
            $1-$2/post-$3.xhtml [R=303]

RewriteCond %{HTTP_ACCEPT} application/rdf\+xml
RewriteRule ^/([0-9]{4})-([A-Za-z]+)/post-([0-9]+)$
            $1-$2/post-$3.rdf [R=303]
```

The fifth scenario is extensively described in the referred source, and it has been implemented in SWAML. At the same time RDF and HTML files are written, SWAML also produces htaccess local configuration files for Apache. One of these configuration file is shown in Figure 3, while a sample request/response negotiation is depicted in Figure 4.

RDF metadata generated by SWAML can grow to a large size for lists with a high traffic and several years of operation, where there are tens of thousands of messages. The partition of the information might be an issue in such cases. On the one hand, information chunks are preferred to be small so any conceivable use case can be satisfied without retrieving a significant overload of unneeded information. However, scattering the metadata across a myriad of small files has some disadvantages. For instance, the number of resources that must be retrieved to fulfill a single query is greatly increased. Therefore, storing the RDF graph in a specialized database is an appealing alternative.

Fortunately, a common protocol to access semantic repositories using SPARQL as the query language is available (Clark 2006) and is gaining support by the RDF databases. This protocol exposes a simple API to execute and retrieve the results of SPARQL queries (at the present moment, SPARQL is a read-only query language, although

Figure 4. An HTTP dialog with content negotiation

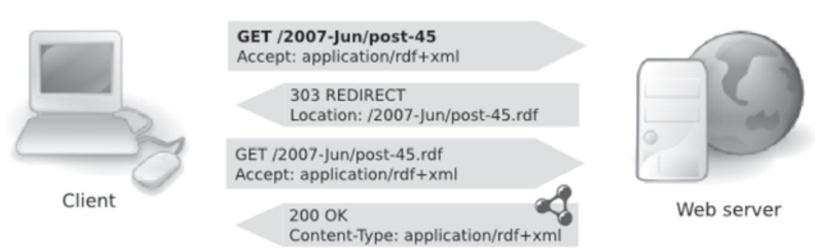


Figure 5. Sample Apache web server rewrite rule to translate HTTP request into SPARQL queries using a Sesame RDF repository. The last line has been wrapped for readability.

```
RewriteEngine On
RewriteBase /lists/archives

RewriteCond %{HTTP_ACCEPT} application/rdf\+xml
RewriteRule ^mylist/(.+)
       http://internal-server/sesame-server/repositories/mylist-rep/
       ?query=CONSTRUCT+{<http://example.org/lists/mylist/$1>+?y+?z}
       +WHERE+{<http://example.org/lists/mylist/$1>+?y+?z}
       &queryLn=sparql [R=303]
```

there are proposals to extend it with full CRUD capabilities such as those of SQL). This abstract query API may be realized by different means, such as SOAP bindings (described by a WSDL 2.0 interface) and HTTP bindings. The former enables interoperability with web service frameworks, while the latter can be exploited without the full-blown web service machinery.

Web service endpoints which implement the SPARQL protocol are sprouting on the web, some of them pouring huge amounts of data into the semantic web. We argue that metadata of large mailing lists can be conveniently exposed as SPARQL endpoints. That means to effectively translate the decision on data selection to the client (Pan 2006), and therefore minimizing the number of requests and the data overload. For instance, the client agent can retrieve all the headers of the messages in a given date range, but skip the body of the messages, saving a considerable amount of bandwidth.

However, non SPARQL-aware agents still need to access the information. This is the scenario of the sixth scenario (recipe) of the above cited document, but unfortunately this one is still being discussed. We propose a simple solution based on URL rewriting of the requests in order to translate conventional HTTP requests for resources into SPARQL queries that dynamically generate an RDF subgraph that contains the requested information about the resource. The rewriting mechanism, the SPARQL query and even the presence of a data repository instead of static files is kept completely hidden to the client. At the same time, by avoiding the undesirable data replication, this technique helps to keep the information consistent. The most representative feature of our proposal is that it does not require any kind of server side script or application to translate the queries, because the data repository can serve the information directly in the format desired by the client.

We have implemented this technique using the Apache web server and Sesame 2.0 RDF repository (Broekstra et al, 2006). Figure 6 reproduces the hand-made htaccess file (as opposed to the ones that are automatically produced by SWAML). Unfortunately, Of course, the rewrite rule must be fired only when RDF data is requested, while requests for HTML must go through it.

We note, however, that our proposal presents some security-related issues. In particular, it is easily vulnerable to SPARQL-injection. Therefore, we strongly discourage the use of this technique in production environments. Nevertheless, some changes in the regular expressions are possible in order to prevent this kind of attack.

There is another different approach to publishing metadata: to embed it into the HTML content. W3C is pushing two complementary technologies, RDFa (Adida & Birbeck, 2007) and GRDDL (Connolly, 2007), which respectively encode into, and extract RDF data from XHTML documents. We have also explored this path. SWAML gener-

Figure 6. A single message rendered as XHTML code with RDFa and GRDDL markup by SWAML.

```
<html xmlns='http://www.w3.org/1999/xhtml'
      xmlns:dcterms='http://purl.org/dc/terms/'
      xmlns:sioc='http://rdfs.org/sioc/ns#'
      xmlns:dc='http://purl.org/dc/elements/1.1/'>
    <head profile='http://www.w3.org/2003/g/data-view'>
    <link href='http://www-sop.inria.fr/acacia/soft/RDFa2RDFXML.xsl'
          rel='transformation' />
    <title>[swaml-devel] CfP: FEWS2007</title>
  </head>
  <body>
    <div about='http://example.org/swaml/post/2007-May/5'
         typeof='sioc:Post'>
      <h1 property='dc:title'>[swaml-devel] CfP: FEWS2007</h1>
      <p>strong>From: </strong>
        <a href='http://example.org/swaml/subscriber/s2'
           rel='sioc:has_creator'>Diego Berrueta</a>
      </p>
      <p><strong>To: </strong>
        <a href='http://example.org/swaml/forum'
           rel='sioc:has_container'>SWAML Devel</a>
       </p>
       <p><strong>Date: </strong>
         <span property='dcterms:created'>
          Tue, 15 May 2007 19:24:49
         </span>
       </p>
       <pre property='sioc:content'><!-- omitted --></pre>
       <p>Previous by Date:
         <a href='http://example.org/swaml/post/2006-Sep/4'
            rel='sioc:previous_by_date'>previous</a>
       </p>
       <p>Next by Date:
         <a href='http://example.org/swaml/post/2007-Mar/6'
            rel='sioc:next_by_date'>next</a>
       </p>
    </div>
  </body>
</html>
```

ates simple XHTML pages for each message to illustrate the usage of both RDFa and GRDDL. We must remark that these pages are just a proof-of-concept of the semantic enrichment, and they lack many of the fancy features and complex templates of the already-existent applications which generate plain HTML.

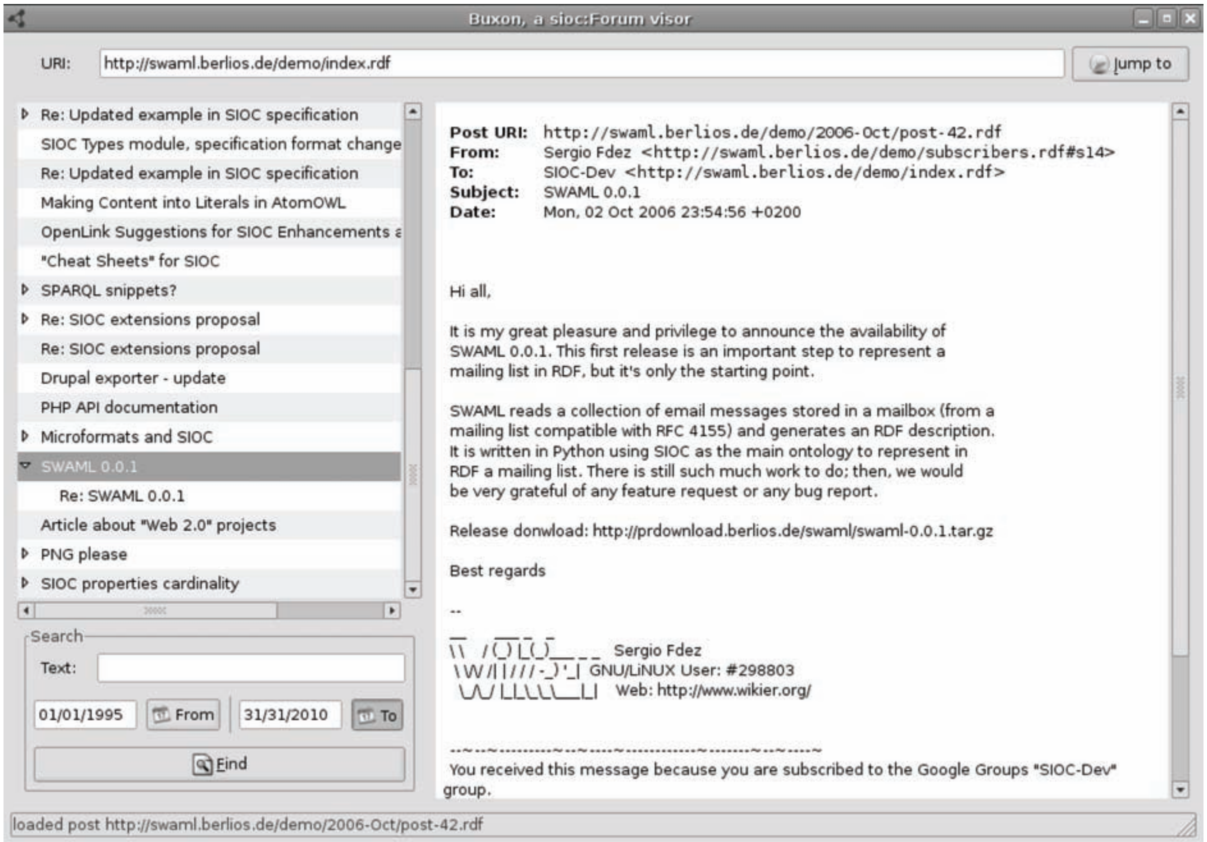
CONSUMING MAILING LIST METADATA

Buxon

Buxon is a multi-platform desktop application written in PyGTK. It allows end users to browse the archives of mailing lists as if they were using their desktop mail application. Buxon takes the URI of a sioc:Forum instance (for example, a mailing list exported by SWAML, although any sioc:Forum instance is accepted) and fetches the data, retrieving additional files if necessary. Then, it rebuilds the conversation structure and displays the familiar message thread list (see Figure 7).

Buxon also gives users the ability to query the messages, searching for terms or filtering the messages in a date range. All these queries are internally translated to SPARQL (Prud'hommeaux & Seaborne, 2007) to be executed over the RDF graph. Newer versions of Buxon can send the sioc:Forum URI to PingTheSemanticWeb.com, a

Figure 7. Buxon browsing SIOC-Dev mailing list.



social web service that tracks semantic web documents. That way, Buxon contributes to establish an infrastructure that lets people easily create, find and publish RDF documents.

Other Browsers and Clients

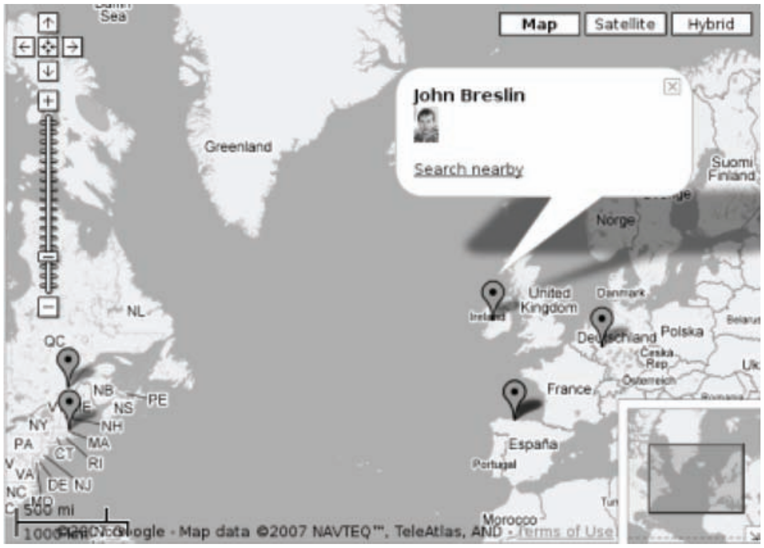
The SIOC RDF data can be explored and queried using any generic RDF browser, such as Tabulator (Berners-Lee et al., 2006). The most interesting applications appear when instances of sioc:User are linked to FOAF descriptions of these users. For instance, it is trivial to write a query to obtain the geographical coordinates of members of a mailing list and to codify them into a KML file (Ricket 2006), provided they describe their location in their FOAF file using the basic geo vocabulary (Brickley 2006). The KML file can be plotted using a map web service such as Google Maps (Figure 8).

It is also possible execute visualize the messages in a time line view using the Timeline DHTML widget by the MIT SIMILE project using a query like the one we propose in Figure 9.

EXPERIMENTATION

A corpus of RDF data with many foaf:Person instances was assembled by crawling and scrapping five online communities. There is a shared topic in these communities, namely open source development; hence we expect them to have a significant number of people in common. We continue the work started in Berrueta et al (2007) to mine online discussion communities, and we

Figure 8. Plotting the geographical coordinates of the members of a mailing list using KML and Google Maps.



extend it to new information sources. More details are described in Berrueta et al We use the following sources:

- *GNOME Desktop mailings lists:* All the authors of messages in four mailing lists (evolution-hackers, gnome-accessibility-devel, gtk-devel and xml) within the date range July 1998 to June 2008 were exported to RDF using SWAML.
- *Debian mailing lists:* All the authors of messages in four mailing lists (debian-devel, debian-gtk-gnome, debian-java and debian-user) during years 2005 and 2006 were scrapped from the HTML versions of the archives with a set of XSLT style sheets to produce RDF triples.
- *Advogato:* This community exports its data as FOAF files. We used an RDF crawler starting at Miguel de Icaza's profile. Although Advogato claims to have +13,000 registered users, only +4,000 were found by the crawler.
- *Ohloh:* The RDFohloh (S. Fernández, 2008) project exposes the information from this directory of open source projects and developers as Linked Data. Due to API usage restrictions, we could only get data about the +12,000 oldest user accounts.
- *Debian packages:* Descriptions of Debian packages maintainers were extracted from apt database of Debian packages in the main section of the unstable distribution.

Instances generated from these data sources were assigned a URI in a different namespace for each source. Some of these data sources do not directly produce instances of foaf:Person, but just instances of sioc:User. An assumption is made that there is a foaf:Person instance for each sioc:User, with the same e-mail address and name. These instances were automatically created when missing. This assumption obviously leads to redundant instances of foaf:Person which will be later detected by the smusher.

The ultimate goal of our experiments is to exercise the smushing processes described previously against a realistic dataset. Two million RDF triples were extracted from the sources described above, and put into OpenLink Virtuoso server

Figure 9. SPARQL query to extract the information required to visualize a time line of the messages posted to any sioc:Forum instance.

```
PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
SELECT ?start ?title ?description ?link
WHERE {
  ?post rdf:type sioc:Post .
  ?post dcterms:created ?start .
  ?post dc:title ?title .
  ?post sioc:link ?link .
  ?post sioc:content ?description
}
```

which provides not only an effective triple store, but also a SPARQL endpoint that was used to execute queries using scripts.

We evaluated two smushers: the first one smushed foaf:Person instances assuming that foaf:mbox_sha1sum is an IFP; the second one smushed the same instances comparing their foaf:name labels for string strict equality, without any normalization. Both smushers were implemented using SPARQL CONSTRUCT rules. The newly created owl:sameAs triples were put in different named graphs. These links were analyzed to find co-occurrences of people in different communities.

Some communities use the e-mail address as their primary key to identify its users. However, other communities use a different primary key, thus allowing users to repeat their e-mail addresses. For instance, a small number of users have registered more than one account in Advogato with the same e-mail (these accounts have been manually reviewed, and they seem to be accounts created for testing purposes).

Our data acquisition process introduces a key difference between how user accounts are interpreted in Debian mailing lists and GNOME mailing lists. The former considers e-mail address as globally unique, i.e., the same e-mail address posting in different Debian mailing lists is assumed to belong to the same user.

On the other hand, a more strict interpretation of how Mailman works is made with respect to the GNOME mailing lists, where identical e-mail address posting in different mailing lists are assumed to belong to a priori different users. In the second case, we rely on the smushing process to merge the identities of these users.

Although they must be handled with extreme care due to the issues afore-mentioned, the combined results of the two smushing processes are consistent with the expected ones. For instance, there is a very high overlap between the Debian developers (maintainers of Debian packages) and the Debian mailing lists. Obviously, Debian developers are a relatively small group at the core of the Debian community, thus they are very active in its mailing lists. Another example is the overlap between Advogato and GNOME mailing lists. Advogato is a reputation-based social web site that blossomed at the same time that the GNOME project was gaining momentum. Advogato was passionately embraced by the GNOME developers, who used Advogato to rate each others' development abilities.

We also studied whether there are some people that are present in many of the communities at the same time. We chose communities which are closely related to each other, consequently, we expected a high number of cross-community subscribers. There are several people who are present in many communities. We can conclude that almost all the most active open source developers in our dataset are core members of the Debian community. Another interesting fact is that only a few people among the top members of the communities consistently use a single e-mail address and just one variant of their names. This fact proves the difficulty of the smushing process, but also its usefulness.

CONCLUSION AND FUTURE WORK

There are a lot of ongoing efforts to translate data already reachable on the web into formats which are semantic web-friendly. Most of that work focuses on relational databases, micro-formats and web services. However, at the time of this writing and to the best of our knowledge, e-mail was almost excluded from the Semantic Web. Our project, in combination with the generic SIOC framework, fills this gap, conveniently providing an ontology and a parser to publish machine-readable versions of the archives of the countless mailing lists that exist on the Internet.

Furthermore, the SWAML project fulfills a much-needed requirement for the Semantic Web: to be able to refer to semantic versions of e-mail messages and their properties using resource URIs. By re-using the SIOC vocabulary for describing online discussions, SWAML allows any semantic web document (in particular, SIOC documents) to refer to e-mail messages from other discussions taking place on forums, blogs, etc., so that distributed conversations can occur across these discussion media. Also, by providing e-mail messages in RDF format, SWAML is providing a rich source of data, namely mailing lists, for use in SIOC applications.

The availability of these data leads to some benefits. In the first place, data can be fetched by user applications to provide handy browsing through the archives of the mailing lists, providing features that exceed what is now offered by static HTML versions of the archives on the web.

Secondly, the crawlers of the web search engines can use the enhanced expressivity of the RDF data to refine search results. For instance, precise semantic descriptions of the messages permit to filter out repeated messages, advance in the fight against spam, or introduce additional filter criteria in the search forms.

Another consequence of no lesser importance is that each e-mail message is assigned a URI that can be resolved to a machine-readable description of the message. This actually makes possible to link a message like any other web resource, and therefore enriches the expressivity of the web.

Integration of the SWAML process with popular HTML-based mailing list archivers, such as Hypermail or Pipermail, would be a giant push to speed up the adoption of SWAML. It is well known that one of the most awkward problems of any new technology is to gain a critical mass of users. The semantic web is not an exception. A good recipe to tackle this problem is to integrate the new technology into old tools, making a smooth transition without requiring any extra effort from users. Merging the SWAML process into the batch flow of tools such as Hypermail would allow users to generate both RDF and production-quality, semantically enriched HTML versions of the archives.

So far, no semantic annotation relative to the meaning of the messages is considered. Obviously, such information can not be automatically derived from a RFC 4155-compliant mailbox. However, it is conceivable that it could be added by other means, such as social tagging using folksonomies, or parsing the metadata added by the authors of the messages using micro-formats or RDFa when posting in XHTML format. The inherent community-based nature of mailing lists can be exploited to build recommendation systems (Celma 2006).

We have also explored smushing techniques to spot redundant RDF instances in large datasets.

We have tested these techniques with more than 36,000 instances of foaf:Person in a dataset automatically extracted from different online open source communities. We have used only public data sources, consequently, these instances lack detailed personal information.

We are aware of the extreme simplicity of our experimentation using label comparison. In our opinion, however, it contributes to show the potential of this smushing technique. We note that it is possible to have more usages for it, for instance, smushing not just by people's names, but also by their publications, their organizations, etc. Surprisingly, the named-based smushing finds a high number of redundant resources even if the comparison strategy for labels (names) is very simplistic (in this case, case-sensitive string equality comparison). More intelligent comparison functions should lead to a higher recall. In this direction, we are evaluating some normalization functions for names. We have also evaluated classical information retrieval comparison functions that take into account the similarity of the strings (e.g., Levenshtein); nevertheless, their applicability to compare people's names is open to discussion.

We believe that the ratio of smushing can be further improved if the dataset is enriched with more detailed descriptions about people. Experiments are being carried out to retrieve additional RDF data from semantic web search engines as a previous step to smushing.

We have implemented a smusher application for persons, and we intend to use it to further investigate the potential for the optimization of the smushing process. The way in which these techniques are translated into actual algorithms is critical to achieve a promising performance of the smushing process, especially for very large datasets. In parallel, increasing the precision of smushing will require to study how to enable different smushing strategies to interrelate and reciprocally collaborate.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Dr. John Breslin and Uldis Bojārs from DERI Galway, whose support and contributions have been a great help to this work. Also thanks to Ignacio Barrientos for his contribution packaging SWAML for Debian GNU/Linux.

REFERENCES

Adida, B., & Birbeck, M. (2008). *RDFa Primer*. Technical Report, W3C Working Draft.

Berners-Lee, T. (2006). *Linked Data Design Issues*. Available at http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., et al. (2006). Tabulator: Exploring and Analyzing linked data on the Semantic Web. *Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI06) workshop*, Athens, Georgia.

Berrueta, D. et al, (2008). *Best practice recipes for publishing RDF vocabularies*. Technical Report, W3C Note.

Berrueta, D., Fernández, S., & Shi, L. (2007). Bootstrapping the Semantic Web of Social Online Communities. *In Proceedings of workshop on Social Web Search and Mining (SWSM2008), co-located with WWW2008*, Beijing, China.

Bojārs, U., & Breslin, J. (2007). *SIOC Core Ontology Specification*. Available at http://rdfs.org/sioc/spec/.

Breslin, J., et al. (2005). Towards Semantically-Interlinked Online Communities. *Proceedings of the 2nd European Semantic Web Conference, ESWC 2005*, Heraklion, Crete, Greece.

Breslin, J. (2006). SIOC: an approach to connect web-based communities. *International Journal of Web Based Communities*, *2*(2), 133–142.

Brickley, D. (2006). *Basic geo (WGS84 lat/long) vocabulary*. Technical report, W3C Informal Note.

Brickley, D., & Miller, L. (2005). *FOAF Vocabulary Specification*. Technical report.

Broekstra, J. (2006). Sesame: A generic architecture for storing and querying RDF and RDF Schema. In *Springer . Lecture Notes in Computer Science*, *2342*, 54–68. doi:10.1007/3-540-48005-6_7

Celma, O. (2006). FOAFing the music: Bridging the semantic gap in music recommendation. *Proceedings of the 5th International Semantic Web Conference*, Athens, USA.

Clark, K. G. (2008). *SPARQL protocol for RDF*. Technical report, W3C Recommendation.

Connolly, D. (2007). *Gleaning Resource Descriptions from Dialects of Languages (GRDDL)*. Technical report, W3C Candidate Recommendation.

Fernández, S., Berrueta, D., & Labra, J. E. (2008). A Semantic Web Approach to Publish and Consume Mailing Lists. *IADIS International Journal on WWW/Internet, 6*, 90-102.

Fernándrez, S. (2008). *RDFohloh, a RDF Wrapper of Ohloh. Proceedings of 1st workshop on Social Data on the Web (SDoW2008), collocated with 7th International Semantic Web Conference*, Karlsruhe, Germany.

Hall, E. (2005). *RFC 4155 - the application/mbox media type*. Technical report, The Internet Society.

Hausenblas, M., & Rehatschek, H. (2007). mle: Enhancing the Exploration of Mailing List Archives Through Making Semantics Explicit. *Semantic Web Challenge 07*, Busan, South Korea.

Klyne, G., & Carroll, J. J. (2004). *Resource Description Framework (RDF): Concepts and abstract syntax*. Technical report, W3C Recommendation.

Pan, Z., et al. (2006). *An investigation into the feasibility of the semantic web*. Technical Report LU-CSE-06-025, Dept. of Computer Science and Engineering, Lehigh University.

Patel-Schneider, P. F., Hayes, P., & Horrocks, I. (2004). *OWL Web Ontology Language: Semantics and Abstract Syntax*. Recommendation, W3C, February.

Prud'hommeaux, E., & Seaborne, A. (2008). *SPARQL Query Language for RDF*. Technical report, W3C recommendation.

Resnick, P. (2001). *RFC 2822 - internet message format*, Technical report, The Internet Society.

Ricket, D. (2006). Google Maps and Google Earth integration using KML. In *American Geophysical Union 2006 Fall Meeting*.

Shi, L., Berrueta, D., Fernández, S., Polo, L., & Fernández, S. (2008). Smushing RDF instances: Are Alice and Bob the same open source developer? *Proceedings of 3rd ExpertFinder workshop on Personal Identification and Collaborations: Knowledge Mediation and Extraction (PICKME 2008), collocated with 7th International Semantic Web Conference*, Karlsruhe, Germany.

This work was previously published in Social Web Evolution: Integrating Semantic Applications and Web 2.0 Technologies, edited by M. D. Lytras; P. Ordonez de Pablos, pp. 42-56, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 4.13
Communicative Networking and Linguistic Mashups on Web 2.0

Mark Pegrum
University of Western Australia, Australia

ABSTRACT

This chapter discusses the application of a range of Web 2.0 technologies to language education. It argues that Web 2.0 is fundamentally about networking, community building, and identity negotiation. Given the textual nature of the Web, all of this is made possible primarily through the medium of language. Consequently, Web 2.0 is ideally suited to the teaching of language and literacy. To be most effective, this requires a broadly social constructivist pedagogical approach as well as a willingness to work with the messy reality of linguistic "mashups," the hybrid uses of languages, codes, and media which inform Web 2.0.

INTRODUCTION

There continues to be widespread confusion and apprehension about the effects of the Internet and new technologies on education. Recent discussions of the web in versions ranging from 1.0 to 3.0 have done little to alleviate this situation, with at least one spurious reference to Web 6.0 (Motteram & Ioannou-Georgiou, 2007) making the point that labels and numbers are not the important thing. However, a glance at Web 1.0 and Web 3.0 can be helpful in an understanding of Web 2.0, the term popularized by Tim O'Reilly through the first Web 2.0 Conference in 2004 (O'Reilly, 2005) and now commonly used to describe the current state of the web.

The retrospective term Web 1.0 refers to the initial *information-oriented web*, authored by a small number of people for a very large number of users. Consisting mainly of static webpages, it offered little room for interactivity. Educational uses largely fell into two categories: information retrieval (as in webquests) or rote training (drill exercises). While there were some clear benefits in terms of student autonomy, use of authentic materials and exposure to multiliteracies, and while problem-based learning and guided discovery approaches to Web 1.0 were not unknown, it was most often used in ways corresponding to traditional transmission or behaviourist models of pedagogy.

Web 3.0, a speculative term describing a possible future version of the web, refers most commonly to the *semantic web*, where software agents will collate and integrate information to give intelligent responses to human operators, and/or the *geospatial web*, where location will be used to index information. These are, however, long-term projections, whose educational implications are impossible to assess at present.

DOI: 10.4018/978-1-60566-190-2.ch002

In between is the presently dominant Web 2.0, also known as the *social web*, which comprises a loose grouping of newer generation social technologies whose users are actively involved in communicating and collaborating with each other as they build connections and communities across the world, negotiating their online identities in the process. What happened, as Davies puts it, was that “society got more technical while software got more social” (2003, p. 5). The 2007 Horizon Report describes Web 2.0’s social networking sites as being “fundamentally about community” (New Media Consortium, 2007, p. 12), while Jimmy Wales (2007), founder of Wikipedia, has linked Web 2.0 to the new digital literacies concerned with “inclusion, collaboration and participation”. In brief, Web 2.0 technologies, from blogs and wikis through social networking sites and folksonomies to podcasting and virtual worlds, are all about communicative networking. Such networking is likely to become increasingly important as a digital native ethos takes over from a digital immigrant one (Prensky, 2001), as more technologies become available to those with little specialist expertise in IT, and as today’s technologies converge to form ever more versatile hybrids.

Web 2.0 has many applications in education, both current and potential, but its greatest impact may well be in subjects which foreground language and communication. After all, given the textual nature of the web, all the connections made online and all the communities established there are enabled primarily through the medium of language. As a result, for language and literacy educators, the advent of Web 2.0 presents great opportunities: to decentralize the role of the classroom (Coleman, 2007), escape the language lab, and engage with the younger generation of digital natives on their own territory. It is a territory whose geography is forged through language and whose key navigation tools are literacies. Teachers can help their students develop greater language competence and additional linguistic tools to navigate Web 2.0, as the students engage in the process of making connections, building communities and shaping their own self-representations online. In this way, language and literacy educators can play a key role in the collaborative enterprise that is Web 2.0. It is important to acknowledge, however, that effective use of Web 2.0 requires a rethinking of approaches to literacy and pedagogy which may have traditionally seemed unproblematic, but which are less than ideally suited to the new online environment — or the wider world in which it is embedded.

This chapter begins by examining recent changes in conceptions of literacy and pedagogy which may enable educators to better frame their use of Web 2.0. It then goes on to discuss common Web 2.0 tools and their applications to language education, focusing firstly on collaborative technologies such as discussion boards, blogs and wikis; secondly on social networking technologies; thirdly on information linking technologies like folksonomies and RSS; and fourthly on cutting-edge technologies such as podcasting, m-learning and virtual worlds. Finally, the chapter explores some of the main limitations of Web 2.0 in education, in a discussion which ranges across pedagogical, social, sociopolitical and philosophical issues. Drawing these threads together, the conclusion offers recommendations for language and literacy educators who wish to use Web 2.0 more extensively in their teaching.

CHANGING LITERACIES AND PEDAGOGIES

It has been clear for some time that traditional print literacy alone is no longer sufficient to allow people to operate effectively in society. Web 2.0 greatly exacerbates the problematical aspects of this situation. As a result, there is an urgent need to pluralize the concept of literacy, as has been claimed in recent work on literac*ies* and *multi*literac*ies* (Barton & Hamilton, 2000; Cope & Kalantzis, 2000; Kist, 2004; Street, 1994;

Unsworth, 2001). It is important to challenge the focus on "formalised, monolingual, monocultural, and rule-governed forms of language" inherent in print literacy pedagogy (New London Group, 2000, p. 9). In their place, multiliteracies should be promoted and developed to facilitate the navigation of "our culturally and linguistically diverse and increasingly globalised societies" as well as "the burgeoning variety of text forms associated with information and multimedia technologies" (ibid.). The multiliteracies paradigm can thus refer to multiple cultural and linguistic codes on the one hand, and to multiple media on the other. Both aspects, but particularly the latter, are reflected in the rapidly multiplying treatises on computer, electronic and hypertext literacies (Dudfield, 1999; Kern, 2006; Selber, 2004; Warschauer, 1999, 2003; Wray, 2004).

In short: in the Web 2.0 environment, there is a dynamic fusion of media and a rich blend of cultures, languages and, within languages, evolving codes and registers. While English may be the default lingua franca, it is less a single international English and more a loosely concatenated assemblage of World Englishes. And generational differences ensure that, even among speakers of single varieties of English, there is a bewildering mixture of modes of self-expression. Indeed, the multilingualism and multiliteracies which underpin Web 2.0 parallel the increasingly productive mixing of pre-existing video, graphics, music and text commonly referred to as *mashups* (a term derived from the hip hop practice of mixing songs to create new hybrids). "Linguistic mashups," then, would seem to be in the nature of international socialization and online networking: the emphasis is on communication, which involves sophisticated aggregations of multiple media drawing on increasingly porous cultural and linguistic codes. Web 2.0 is not about neat definitions or clear borders. Rather, its users must find ways to work with the global cacophony of voices which make up its textual fabric.

Fortunately, there is a range of appropriate pedagogical tools at hand. While Web 1.0 lent itself to transmission pedagogies and behaviourist drills, working effectively with Web 2.0 demands a more constructivist orientation. Social constructivist pedagogy, with its roots in the work of Vygotsky and carrying influences from Dewey and progressivism, views social interaction as the source of all learning. Acknowledging and valuing students' pre-existing knowledges and multiple perspectives, it helps students deconstruct and reconstruct these as they engage actively and collaboratively in building new understandings through scaffolded learning experiences (Dalgarno, 2001; Finger, Russell, Jamieson-Proctor & Russell, 2007, p. 119; Jonassen, 1992). As Hoppe, Joiner, Milrad and Sharples (2003) state, "there is an imperative to move from a view of e- and m-learning as solely delivery mechanisms for content"—the transmission approach typical of Web 1.0—and to embrace contemporary pedagogy with its "high valuation of *active, productive, creative and collaborative learning methods* [which go] much beyond the 'absorption' of codified information" (p. 255; italics in original). It might be argued that a constructivist approach is becoming ever more relevant in a world where, as Warschauer (2007) indicates, "[t]he ability to draw on rote answers is inadequate" because "yesterday's answers are outdated faster than ever" (p. 42). What is relevant in such a world is the ability to seek out information through networks of contacts, and to collaboratively build understanding with others engaged in similar pursuits. The social networking, dialogue building and collaborative knowledge construction tools of Web 2.0 are uniquely suited to preparing students for this world.

Another useful perspective is provided by the communities of practice paradigm, where learning is conceived of as "social participation," meaning that people engage in the "process of being active participants in the *practices* of social communities and constructing *identities* in relation to these communities" (Wenger, 1998, p. 4; italics in original). Communities of practice have, in fact, been defined as "networked learning systems" which connect "all participants and learning system components across multiple levels of practice and inquiry" (Quinton, 2006, p. 563). This is precisely the kind of educational networking that can be fostered by Web 2.0 applications. As students begin to use these tools, they are not only gaining important future skills but may well find themselves entering, as legitimate peripheral participants, the very communities of practice in which they will eventually become full participants. It is implicitly a community of practice orientation that Holmes, Tangney, FitzGibbon, Savage and Mehan (2001) ascribe to when they express the hope that, in a "communal constructivist" approach to new technologies, "students will not simply pass through a course like water through a sieve but instead leave their own imprint in the development of the course, their school or university, and ideally the discipline" (p. 1).

In language teaching itself, the last decade of the twentieth century witnessed a move away from the ideals of the communicative approach — which, having dealt with some of the key limitations of preceding approaches, came to create its own problems — and towards a conception of intercultural communicative competence. While continuing to recognize the importance of the communicative element, the intercultural communicative competence movement has rejected any insistence on the imitation of native speaker models along with the accompanying goal of integration into a target culture. Rather, the language learner is encouraged to move into a "third place" (Kramsch, 1993) between cultures; from here, he or she will be able to explore his/her own culture as well as other cultures, which are not seen as static entities into which full integration might be possible, but rather as multiple, contradictory and in flux (Byram, 1997; Corbett, 2003; Kramsch, 1998; Phipps & Gonzalez, 2004). Intercultural competence is thus very much about negotiating communication in "the messy real world of cultural flows and mixes" (Pegrum, forthcoming 2008a) — one whose messiness is exponentially increased by the technological affordances and communicative possibilities of Web 2.0.

In the new millennium, the notion of identity has also emerged as a major focus of research in language pedagogy, thanks in large part to the work of Norton (2000), who observes that "an investment in the target language is also an investment in a learner's own identity" (p. 11). Pavlenko and Blackledge (2004) foreground the questions of power and em*power*ment which underpin identity concerns:

individuals are agentive beings who are constantly in search of new social and linguistic resources which allow them to resist identities that position them in undesirable ways, produce new identities, and assign alternative meanings to the links between identities and linguistic varieties. (p. 27)

Ricento (2005) goes even further in describing:

the central role of language in the negotiation of a person's sense of self at different points in time and in different contexts, and in allowing a person access (or lack thereof) to powerful social networks that give learners the opportunity to speak. (p. 898)

Web 2.0 places an even greater premium on such issues for language teachers and learners: it elevates to the level of a constituting principle the notion that identity is constructed through language.

In sum, if the limitations of a "single-mode, single-language, single-culture literacy" (Pegrum, forthcoming 2008a) were always apparent to some, they are all the more obvious in our shrinking world, where members of the net generation are simultaneously bound together and yet differentiated from each other through their use of Web 2.0

tools. What Pennycook (2007) has recently written in regard to the rapidly globalizing culture of hip hop—original source of the mashup—applies equally to students' desire for linguistic and cultural self-realisation on Web 2.0:

If we believe that education needs to proceed by taking student knowledge, identity and desire into account, we need to engage with multiple ways of speaking, being and learning, with multilayered modes of identity at global, regional, national and local levels.

Unless we get in touch with this as educators, the flow will pass us by. … Languages will flow and change around us, new combinations of languages and cultures will be put together, texts will be sampled and mixed in ever new juxtapositions. Students are in the flow; pedagogy needs to go with the flow. (p. 158)

Of course, it is not only about multiple Englishes, but multiple languages. It is not only about multiple texts, but multiple textualities. It is time, as Canagarajah (2003) has suggested, to begin teaching the "fluid literacies" (p. xi) essential for navigation and negotiation in this new hybrid world:

Rather than developing mastery in a 'target language,' we should strive for competence in a repertoire of codes and discourses. Rather than simply joining a speech community, we should teach students to shuttle between communities. Not satisfied with teaching students to be context-sensitive, we should teach them to be context-transforming. (p. xiii)

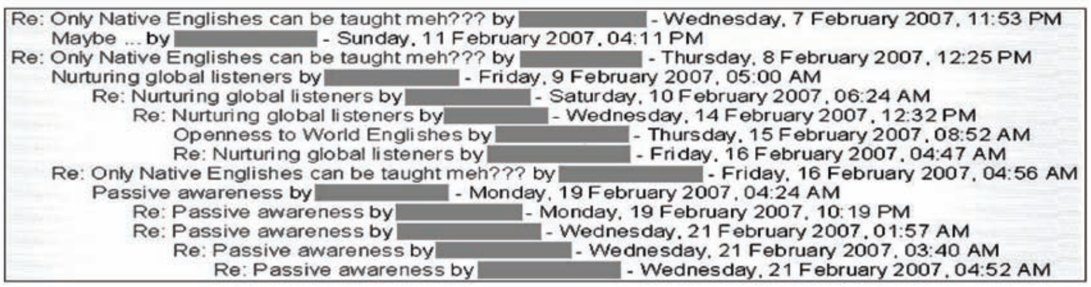
Few can doubt that students are part of this world already, on the web and beyond it. But that does not mean they are fully accomplished navigators, have all the language and literacy skills they need, or always exercise appropriate critical judgement. Most students, Hubbard (2004) notes, can "profit from some formal, sustained training in how to take *operational* competence in a given computer application and transfer that into *learning* competence" (p. 51). More than this, students need to learn critical literacy skills to sort through, evaluate and prioritize the masses of data with which they are confronted, turning information into understanding (McFarlane, Roche & Triggs, 2007; Pesce, 2007). They also need a grasp of the powerful linguistic and media options at their disposal for shaping their identities and engaging with others online. It is a fallacy to think that educators in this new virtual world are no more than facilitators. As has been widely argued in the literature about online learning, and in line with social constructivist pedagogical models, teachers must be prepared to play a central organizing, guiding and mentoring role (Garrison & Anderson, 2003; Pegrum, 2007; Warschauer, 2007).

In doing so, they have a golden opportunity to engage with their students. They can support the latter's online self-presentations and endorse their community building by helping to enhance their language and literacy skills. At the same time, teachers should be open to learning from their students about their digital lifestyles — and in the process, teachers may well find their own language and literacy skills enhanced in unexpected ways. Collaboration which brings together teachers' pedagogical and critical expertise and students' technological and practical expertise is the only way to unlock the full educational potential of Web 2.0.

THINKING COLLABORATIVELY

Much of Web 2.0 is devoted to fostering communities of interest or practice which nurture collaborative thinking. As such, it effectively illustrates the potential, noted by Kaye in the early days of computer-mediated communication, for the "weaving together of ideas and information

Figure 1. Sample discussion thread replies from Third Space Trial 1, Feb. 2007



from many peoples' [sic] minds" (1989, p. 3). This principle underpins asynchronous discussion boards (DBs), in some ways a spiritual precursor of Web 2.0, along with the more multifaceted blogs, wikis and hybrid blikis (or blokis), all of which may contain in-built discussion or comments features.

Being text-based, asynchronous DBs are natural vehicles for the development of writing skills, while there is some limited evidence they may also support the development of oral skills (Burgmer, 2006, p. 96; Levy & Stockwell, 2006, p. 182). It has been widely observed that writing on the Internet, because of its conversational nature, often takes the form of a hybrid code, mixing together features of speech and writing with its own peculiar elements (Crystal, 2001a, 2001b; cf. Al-Sa'Di & Hamdan, 2005, on synchronous chat). It is worth bearing in mind, then, that DBs may not only help students learn about standard spoken and written language, but about hybridized language uses of the kind with which they need to be familiar in order to enter fully into many online environments.

If structured carefully, asynchronous DBs can promote the formation of learning communities where students, reacting to and building on each other's ideas in branching discussion threads, collaboratively construct their understandings of the subject matter at hand — all through the medium of written language, which is probably more conducive to reflective educational discussion than newer voice alternatives (whether synchronous VoIP or asynchronous voiceboards). See Figure 1 for an example of threaded postings in an international Master's forum for language teachers. Used in conjunction with face-to-face classes, DBs may help cater to differing learning styles and needs. For example, they allow more time to be spent on composition of contributions by less extroverted or non-native students; the time-independence of DBs may thus "mitigate the effects of certain inequalities" (Locke, 2007, p. 188). It has also been widely claimed that DB exchanges typically display a high level of cognitive sophistication (e.g., Garrison & Anderson, 2003, p.26; Heckman & Annabi, 2005; Hiltz & Goldman, 2005, p.6). This may be because "[t]he historical divide between speech and writing has been overcome with the interactional and reflective aspects of language merged in a single medium" (Warschauer, 1999, p.6). This particular aspect of online hybridity would certainly seem to have major advantages.

When they involve multilingual or multicultural cohorts of students, DBs may equally promote the development of intercultural competence. In the ongoing *Third Space in Online Discussion* research project, which involves language teachers enrolled in Master's courses at the University of Western Australia and Canterbury Christ Church University, UK, discussion forums (like that seen in Figure 1) are being analyzed as educational "third spaces" which exist in the interstices between students' cultural and educational experiences, and where there is ample space for the deconstruction and reconstruction of pedagogical, linguistic and cultural knowledge and understanding (Pegrum & Bax, 2007). It is apparent that, as Zieghahn (2001) realized some years ago, "the online environment offers a unique medium through which to reflect upon individual cultural position and on intercultural communication" (p. 144). While most educational DBs necessarily operate in a single lingua franca, multilingual forums are possible in some language learning situations. Linguistically as well as culturally, then, DBs can help educators respond to Canagarajah's aforementioned plea to teach students to shuttle between communities. In the process, their sense of their online — and perhaps offline — selves may be shaped through their interactions with peers.

While blogs — described by Doctorow (2002) as "outboard brain[s]" — can function as reflective diaries, they can also be conversational centrepieces: readers may leave comments for a blog's author and each other, thereby forging connections and community around topics of mutual interest. Students can certainly join the conversations on others' blogs, but they can equally set up their own. Receiving feedback on blog entries from peers and teachers can facilitate knowledge construction as well as perspective shifts as they go about developing their online personas. Indeed, with fully public blogs, students can potentially receive feedback from anyone on the entire Internet and may, as a result, invest themselves more fully in writing and publishing tasks.

Because blogs can be multilingual (allowing some mixing of the mother tongue with the target language), multimodal (allowing pictures, video and audio to support written text), and carefully designed (drawing on technical knowledge and artistic flair), students at even the lowest levels of linguistic proficiency need not feel the work they are creating fails to capture or express important aspects of their identities or beliefs. At higher levels, as students' linguistic competence develops, they can present more nuanced versions of themselves. As Kazan indicates, "[w]ithin cyberspace, writers have flexibility in how they construct a self and the more strategies they acquire, the more flexibility they have" (2007, p. 264). The task for teachers is to help students make more "informed rhetorical decisions" (ibid.), which will allow them not only to shape their online identities as they wish, but also to "develop a public voice about issues they care about" and so come to understand "their literacies as citizenship skills as well as avenues to entertainment" (Rheingold, 2007).

Wikis are even more strongly oriented towards collaboration than blogs since they are effectively co-operatively authored websites. They turn the element of collective intelligence implicit in blogging communities into a structural principle. Students are able to engage in a form of process writing in which they draft and redraft work collaboratively, each contributor adding to and modifying the work of peers. With a private wiki, feedback can be received from the class teacher and peers, or, with a public wiki, from the entire Internet. As Mitchell (2005) notes, it has even been suggested that wikis are an example of "the tried and trusted system of peer review taken to a new level" (p.120).

One option is for students to contribute to preexisting wikis such as Wikipedia or, for learners, Simple English Wikipedia, thereby entering into established communities of practice. Alternatively, dedicated class wikis can be set up on subjects of relevance or interest, and in time new communities of practice may form around these. Even a course constrained by a tight, exam-oriented syllabus can exploit wiki technology: under the guidance of the teacher, each individual's or group's research could feed into a network of student-constructed documents reviewing material to be covered in the exam. This might include vocabulary accompanied by definitions and examples; grammar points accompanied by explanations and illustrations; or set literature accompanied by summaries and quotations. Once again, there is ample opportunity for multilingual, multimodal, technically sophisticated and artistically creative presentation. The more sophisticated the wiki, the greater the students' facility with multiliteracies will need to be — or become.

SOCIAL NETWORKING

Social networking technologies also promote collaborative thinking, many of them effectively harnessing the power of collective intelligence, but the accent is on the networking aspect. It has been suggested that Facebook, for example, "puts the social community first, with content — including, but not limited to, educational content—being the medium of exchange" (Downes, 2007). Some observers claim that virtual networks are replacing the gradually disappearing or increasingly inaccessible public spaces in which young people formerly gathered (boyd, 2006). These networks are intimately bound up with selfhood; the sense of empowerment that comes from the crafting of personal identity on social sites (Coghlan, 2007) goes hand in hand with negotiating membership of the groups of friends and acquaintances who congregate there. The potential effects on language education are an extension of the paradigm shift neatly captured at the start of the millennium by Kramsch, A'Ness & Lam (2000, p. 97) in their comments on language learning through participation in informal online interaction:

The kind of language experience … in which rules are learned first and then put to use in conversation, has given way to a learning by doing, and learning to meet the demands of doing in specific contexts, to solve immediate problems together in the small culture of communities of practice (Holliday, 1999; Lave & Wenger, 1991; Uber Grosse & Leto, 1999; Wenger, 1999). Rather than an object of reverence or study in itself, language is viewed as a tool which brings people together and creates intimacy (Harmon, 1999). What is important is how you relate, emotionally, and physically, to that world.

Social networking sites, with MySpace and Facebook being by far the most popular, allow each user to set up an online identity, or profile, and to keep in touch with friends and acquaintances by constantly updating this profile while regularly viewing others' profiles; new contacts can be established through mutual acquaintances or shared interests. Since 2006, Facebook has used a news feed system to keep users updated on changes to the profiles of their contacts. Typically, social networking sites integrate a range of other communication channels, which may include email, instant messaging (IM) and even blogs, with facilities for sharing photos, videos and audio files. There is a fine line separating these sites from social sharing services, such as Flickr for photos or YouTube for videos. Facebook allows the integration of Flickr photos as well as del.icio.us tags (see below) into profiles, while it is now also possible for users to assemble friends and acquaintances from the virtual world Second Life alongside their other contacts.

Social networking sites are perhaps the most maligned feature of Web 2.0, mainly due to fears of Internet predation but also because of concerns over time spent online, as well as the possible degeneration of literacy skills as the digital natives communicate ever more rapidly in ever more truncated "netspeak." Yet the reality is that students are already using social networking sites and educators have the choice to work with or against them. The advantage of the former strategy is that it is possible to openly address concerns over Internet safety or time spent online, attempting to provide guidance in such areas. This might be extended to include a focus on what Barney (2007, p. 279) refers to as "critical technological literacy": asking questions about the presuppositions and blind spots, the benefits and drawbacks, in short, the "affordances and … denials" of different technologies. Helping students adopt a critical distance to all technologies will do them a much greater service in the long run than simply closing down all discussion in the classroom, leaving them to conduct their explorations, unguided, in their own time.

At the same time, there are many educationally beneficial aspects of social sites which can be more fully exploited. According to recent US statistics, some 59% of 9-17 year olds say they talk about topics broadly related to education on social networking sites, while 50% claim to discuss schoolwork (National School Boards Association, 2007, p.1). Thus, whatever educators may think, students have already appropriated social networking as a constructivist learning tool. However, educators could certainly do more to encourage the use of this tool for groupwork outside the classroom. The potential for language learning partnerships is undoubtedly great. Lakshimi (2007), discussing her English language students' use of the social networking site Orkut, comments: "Students who have been incommunicado in the classroom are so interactive on Orkut that it leaves me wondering if Orkut would be a better teacher than I am in helping students learn to use English to be socially interactive" (n.p.). Interaction, of course, is precisely the motivation: the wish to communicate and participate, with language being an essential tool.

Social sharing sites offer the additional possibility of posting individual or collaborative work to the web, with students viewing each other's materials and, for example, commenting on their peers' photographed posters (Flickr), PowerPoint slides (Slideshare), presentations (YouTube) or short films posted to blogs (such as the *English Advertising Class*). As Coghlan (2007) observes with regard to student-created advertisements on the last of these sites, some examples may involve little traditional language use, but there is a lot of learning potential in the areas of "multiliteracy, digital literacy and e-literacy."

The communication on social networking and social sharing sites offers, finally, a unique opportunity to explore with students the nature and uses of netspeak, when and where it is appropriate, and how to codeswitch between netspeak and more standardised language forms. One of the main reasons for the widely criticized spread of netspeak into more traditional domains of literacy may well be students' ignorance of codeswitching or their inability to carry it out appropriately. Teachers' failure to explicitly address this area with students can only limit the latter's repertoire of literacies and constrain their ability to access and move between linguistic communities.

INFORMATION LINKING

Folksonomies are a step beyond social sharing. Relying very much on the principle of collective intelligence, they are a way of indexing distributed knowledge, which is then typically presented in the semi-organic form of a tag cloud, as seen in Figure 2. In essence, they allow information linking with a social element, because people (the "folk") have a central organizing role, which gives rise to rich "person-mediated serendipity" (Lambe, 2006, n.p.). After all, people who use the same tags are likely to have similar interests; and a folksonomy allows tags to be traced to users, and those users' other tags to be explored. The potential for "collaborative information discovery" (Alexander, 2006, p. 36) may be exploited by students working together to create class folksonomies dependent on criteria negotiated and evaluated by the students themselves. This could even involve the tagging of the students' own material posted on wikis or social sharing sites. Given the usefulness of well-constructed folksonomies, they might also be consulted by members of wider communities of practice on the Internet and could provide a means of entry into such communities; as Wenger (1998) reminds readers, learning communities should not be isolated but should "use the world around them as a learning resource and be a learning resource for the world" (p. 275). In all

Figure 2. Extract from E-language Tag Cloud (http://e-language.wikispaces.com/e-learning-tagcloud)

cases, tagging, like indexing of any kind, requires a high level of facility with the language being used for classification. With sufficient scaffolding, folksonomy building can function as a literacy enhancement exercise.

RSS (Really Simple Syndication) feeds provide automatic updates of syndicated content — ranging from blog entries to podcasts—from sites to which a user subscribes. Many homepage, blog and wiki services now make it very easy to include selected RSS feeds on webpages. Drawing in feeds from other sites in this way amounts to the incorporation of others' views and perspectives, leading to the co-construction of knowledge within a new frame. At the same time, as Anderson (2006) notes with respect to blog feeds, distribution of content by RSS allows "public review, argument and resolution of topic issues by students globally—in the process creating outstanding international learning opportunities" (p. 146).

Incoming feeds naturally entail a constant stream of information flowing into a desktop aggregator or webpage. The language could be that of native speakers; thus, learners could conceivably subscribe to media or blog feeds in languages they know or are learning, and would be exposed to extensive authentic input. There is also an argument, however, for subscribing to non-native language feeds. For example, TESOL students working in a World Englishes paradigm might find it beneficial to subscribe to feeds from Kachru's outer or expanding circles. Incorporating both native and non-native feeds would lead to a rich patchwork of first and additional language usage, approximating in some ways the multi-dialectal reality of today's world. Awareness of multiliteracies can be enhanced through feeds which distribute audio or video content in addition to or in place of written text.

MASHUP FRONTIERS

Some of the greatest educational promise is to be found in the areas of podcasting, vodcasting, m-learning and virtual worlds, all of which offer considerable language learning opportunities, especially for those prepared to work with multiple literacies and language mashups.

M-learning refers to education involving mobile technology. The best-known example is podcasting, where syndicated audio files, potentially with accompanying text or image files, are downloaded from the web and transferred to a portable device such as an iPod or MP3 player, thus facilitating "time and place shifting to access the content" (Molina & 2006 EDUCAUSE Evolving Technologies Committee, 2006, p. 122). Listening to podcasts is widely perceived as advantageous for learning foreign languages or even brushing up on grammar, vocabulary or style in one's first language. Surveying a selection of national iTunes stores on the randomly selected date of 17 October, 2007, for instance, it was found that the majority of the 25 most popular educational podcasts in each country were related to foreign language learning or first language improvement: 24 in Spain, 22 in Germany and Switzerland, 21 in Australia, New Zealand and the UK, 20 in Canada and Ireland, 19 in France, Sweden and the US, and 17 in Italy.

M-learning can also involve regularly sending students digestible chunks of information via mobile phones, as has been done, for example, with Italian vocabulary accompanied by quizzes at Griffith University in Australia (Levy & Kennedy, 2005). However, there is the potential for greater levels of interactivity than this, as suggested in a recently proposed definition of m-learning as "the processes of coming to know through conversations across multiple contexts among people and personal interactive technologies" (Sharples, Taylor & Vavoula, 2007, p.225). For example, students can work individually or, better still, collaboratively to create podcasts or even vodcasts – as video podcasts are usually known – for publication to the web. Moblogging, or mobile blogging, allows students to use devices like mobile phones to post text, audio or video files to blogs. Peers and teachers can then respond to these postings in traditional text or mixed-media formats, addressing the communicative intent while possibly also critiquing features of language or composition. In many cases, spoken language will be foregrounded, thus helping to balance out the heavy emphasis on written text still typical of the web, including Web 2.0. Sometimes there may be room for multiple linguistic codes and registers if not multiple dialects or languages. More sophisticated versions of m-learning involve participants interacting with real-world environments and each other with the aid of GPS-enabled phones and other portable devices, which may provide instructions and information as well as a variety of communication channels; salient examples range from the MOBIlearn Uffizi Gallery trial in Florence, Italy (Sharples, Taylor & Vavoula, 2007, pp.236-242) to the *Handheld Augmented Reality Project*, or *HARP*, conducted at Harvard University in the USA (Harvard University, n.d.).

Virtual worlds are perhaps the most striking realization of the possibilities of Web 2.0. The avatars which inhabit them are certainly Web 2.0's clearest example of the potential for identity creation, shaping and development. These worlds are very much about networking. Within them, avatars' understandings of their new environment are constructed largely through their engagement — their sharing and building of knowledge — with other avatars. Externally, virtual worlds are supported by and increasingly integrated with blogs, wikis, and social networking sites. Operating around and through these sites are distributed knowledge systems where, as in the gaming communities discussed by Williamson and Facer (2004), the key information is found "in the interconnections between the 'nodes' (the people, texts, tools and technologies) in the network, rather than with isolated individuals" (p. 266, with reference to Gee). In a comment which captures something of the richness of the virtual/non-virtual interface, the best-known of these worlds, Second Life (SL), has been described as "a playground [and] a crucible for ideas about how people can augment their interaction through constructive, and constructivist, play/work/whatever" (Stevens, 2007, n.p.).

Since the rollout of voice technology to SL in mid-2007, in-world avatar-to-avatar interactions can involve a mixture of spoken and written language not unlike that found in the real world. This creates valuable opportunities for students to try out new language, building up confidence and fluency before embarking on real world encounters. Language teachers have been quick to pick up on this potential, with the inaugural SLanguages Colloquium taking place on 23 June, 2007, and bringing together around 50 educators from across the globe; a snapshot of the opening talk by Gavin Dudeney is shown in Figure 3. Language teaching is already underway in SL, with English classes on offer, for example, through The English Village and Languagelab.com. SL also offers immersive

Figure 3. Inaugural SLanguages Colloquium on EduNation in Second Life, 23 June 2007. Reproduced by kind permission of Gavin Dudeney, EduNation.

linguistic experiences outside formal classes, a point emphasised at the inaugural in-world Festival of European Languages in 2007, which promoted the idea of learners seeking out target language areas of SL in which to practise their skills.

A certain degree of linguistic versatility is advantageous for anyone wishing to develop a fuller SL presence, since language is the glue which holds together any community which establishes itself there. Community, in fact, has been described as the killer app of SL (Yowell, cited in Panganiban, 2007). Different languages, and certainly different dialects and registers, are necessary for effective participation in a range of contexts and communities, with an increased linguistic repertoire being a concomitant of increased community involvement and wider social networking. This is not unlike the real world, except it is now possible to cross linguistic and cultural boundaries without leaving one's desk.

Anecdotal evidence suggests multilingual interactions in SL are becoming more common. A striking example of a four-person, five-language (Catalan, English, French, Portuguese and Spanish) conversation has been described by Gavin Dudeney, who writes of "the ease with which some of us switched between the languages we knew, and typed furiously to reformulate things we thought one of the others wouldn't understand into a language they would," resulting in "a very rich evening" (personal communication, 11 Oct. 2007). Vance Stevens (2007) quotes a comment about SL which hints at intriguing language education possibilities: "Yesterday a cheerful Italian gave me a Babbler translator so we started teaching each other Italian and Hungarian using English as the common language, which was real fun, especially that we were figure ice-skating meanwhile" (n.p.). Participation in such conversations — and teaching scenarios — requires a willingness to engage with the unruliness of linguistic globalization as reflected through virtual world encounters. It demands a capacity to codeswitch and a facility with intercultural communicative competence skills: in short, the agility to shuttle between linguistic and cultural communities. This, in turn, reads like a set of lesson aims compiled from recent thinking on language pedagogy. While a single target language will necessarily remain the focus of most language lessons — and can be supported with SL immersion experiences — there is no reason why students should not occasionally be exposed to multilingual, multicultural interactions, especially as these are likely to become ever more central not only to the SL microcosm, but to the wider web, and indeed the world which lies beyond it.

LIMITATIONS OF WEB 2.0 IN EDUCATION

This chapter has discussed the potential of Web 2.0 for education generally and language education specifically. However, it will take time for current practices to become more widespread and for the potential of Web 2.0 to be fully realized. This requires further "normalization" of computing, so that the majority of educators eventually come to regard it with neither fear nor awe, but see it as simply providing a set of tools which may be used in the service of particular pedagogical goals (Bax, 2003; Chambers & Bax, 2006). Teacher training has a major role to play in demystifying computing. Specifically, this entails providing teachers with appropriate pedagogical frameworks for e-learning; an overview of the range of tools available; and adequate technological skills so that they do not feel intimidated by their students' know-how and, moreover, have the confidence to draw on the latter's technological expertise to complement their own pedagogical expertise. In addition, Web 2.0 provides very serviceable tools for building social constructivist professional development forums, and it is possible to imagine that in time "Web 2.0 may well become the biggest training institution in the world" (Consultants-E, 2007). This point will only be reached, however, if intensive preparatory work is carried out by today's teacher training institutions.

While learning about the advantages of Web 2.0, teachers must equally come to understand that e-learning is not, in and of itself, automatically constructivist or pedagogically progressive (Pegrum, forthcoming 2008b), and demands for speed, flexibility and cost saving can easily lead to impoverished content delivery systems. As suggested earlier, some creativity is needed to work within the constraints of rigid syllabi or assessments. As rewarding as it may be, well-designed online learning will normally require a heavy investment of time and energy by both staff and students. There is also a danger that, in their current state of "continuous partial attention" (Stone, 2006), technology users will lose the ability to focus clearly as well as the will to occasionally power down their multifarious communication channels and make time for reflection—a crucial part of education (Pegrum, 2005). And, even while acknowledging the benefits of constructivism, it might be asked whether it is possible or desirable to teach everything in a constructivist manner all of the time. It is important to maintain balance in all of the above areas.

If students are already spending a lot of time online, added educational demands should have a clear value. The identity issues permeating online presence are complex and delicate, and educators should beware of aggravating narcissistic tendencies which may be nourished by social networking (Ryan, 2007). Teachers must also face the fact they may not be welcome to approach students on some sites and through some channels; sensitivity is needed in negotiating educational uses with students.

Collaborative work raises questions of authorship and ownership, while non-participation is often not an option, as Conrad (2002) has noted with regard to virtual learning environments: "you cannot run *and* you cannot hide. Online life is a fishbowl existence" (p.208; italics in original). There is some cause for concern over privacy on social networking sites like Facebook (boyd, 2008). What is more, a lot of online material is preserved indefinitely so that, as Friedman (2006) warns in a more general context, "whatever you do, whatever mistakes you make, will be searchable one day" (p. 185).

Of course, the continued presence of a digital divide—or, more accurately perhaps, a digital spectrum (Haythornthwaite, 2007)—means that not everyone around the world, or within any given society, has equal access to the Internet. While the rapid spread of mobile technologies partially alleviates this situation, it is not the end of the issue: in recent years, the digital divide has come to be seen less in terms of access to technology and more in terms of skills and patterns of use (*ibid.*; Warschauer, 2003) or, in short, digital literacy. It should also be remembered that "global communication technologies are cultural artifacts that are produced by and productive of socio-historically located subjects" (Belz & Thorne 2006, p. xviii), and that they carry the Anglo-centric and, more broadly, Western values of their creators (Ess, 2007; Goodfellow, 2003; Reeder, Macfadyen, Chase & Roche, 2004). Students from varying linguistic, cultural, ethnic, religious, social and educational backgrounds may have their reasons for not wishing to participate in some or all online activities—reasons whose legitimacy is often eclipsed in Western secular education. Compromises must be sought with students who, for example, may struggle with the radically egalitarian nature of social networking technologies or who, as Sabre (2007) notes, might be uncomfortable with virtual worlds because of religious prohibitions on graphic representations of humans.

But perhaps the greatest single issue for would-be Web 2.0 educators may be an inability to step outside traditional philosophical and sociopolitical frames of reference. This could mean an inability to see outside the frame of Enlightenment rationalism and objectivism and to grasp the socially constructed nature of knowledge and learning, a fundamental flaw in Keen's timely if hyperbolic critique, *The Cult of the Amateur* (2007). It might mean an inability to value collaboration and community on their own terms outside of a capitalist paradigm of competition, as seen in Tapscott and Williams' otherwise informative *Wikinomics* (2006). It could mean an inability to perceive that, for the net generation, the notion of a prophylactic divide between "virtual" and "real" life makes little sense: like the radio or the telephone for older generations, the virtual is just another part of the real (cf. Davies, 2003; Thorne, 2006, p. 20). In fact, the connections between them are becoming ubiquitous, as seen in services such as Vodafone's InsideOut, which allows calls between the physical world and Second Life.

For this reason, despite initial evidence which points to a lowering of cognitive performance and efficiency through multitasking (Baron, 2008; Wallis, 2006), it is possible that students who monitor multiple IM channels while writing assignments or who send text messages during lectures are engaging in what, for them, is "a natural way to interact and construct their own learning" (Reddekopp, 2006). Through practice, they may have adapted to such behaviour (Baron, 2008). What if, moreover, such a melding of learning, networking and identity building could give rise to lateral connections and a more holistic mode of education? In the absence of empirical evidence, these reflections are necessarily speculative. However, it is important not to close off new possibilities before they are fully apparent, thereby perhaps losing valuable educational opportunities—and losing students' allegiance along the way. While the digital natives have much to learn about language and literacy from an older generation of teachers, the teaching profession as a whole has much to learn from its digital native students, especially here at the technological and social frontier of Web 2.0.

CONCLUSION

The technologies covered in this chapter — discussion boards, blogs, wikis, social networking, social sharing, folksonomies, RSS, podcasting, vodcasting, m-learning and virtual worlds — comprise a representative Web 2.0 list, but one which is both incomplete and unstable. New technologies and applications are constantly appearing, while there is an overall tendency towards functional convergence. Yet, however this list might look a few years from now, it is likely to still be informed by the fundamental features this chapter has described as underpinning Web 2.0: communicative networking, community building and identity negotiation, performed through hybrid codes, multiple media and linguacultural mashups.

Writing of Web 2.0, McIntosh (2006) suggests that "[t]he reason these social technologies work is because they are social. But they are also changing the way that we socialise" (p. 72). As has been seen, socializing and networking on Web 2.0 are very much dependent on language. Web 2.0 is, after all, "a means whereby just about anyone can contribute to an ongoing 'conversation' in which knowledge is both discovered and constructed as it goes on" (Freedman, 2006, p.13), and there can be no conversation without language. It is little wonder, then, that Crystal (2001a, 2001b) has called the Internet a linguistic revolution; that Macfadyen & Doff (2005, following Cicognani) have claimed that cyberspace must be viewed in linguistic terms; or that some observers feel Web 2.0 comes close to realizing Tim Berners-Lee's original idea of the web as a "read-write medium" (Lee & Berry, 2006, p. 20).

It has been suggested in this chapter that language and literacy educators are in an ideal position to exploit the linguistic nature of Web 2.0. This requires a conception of literacy – indeed, of multiliteracies – which is appropriate to Web 2.0 and the increasingly interconnected world of which it is both a symbol and a product. It requires a suitable pedagogical base for e-learning, drawing on social constructivism, communities of practice, intercultural communicative competence and identity studies. It requires a familiarity with the advantages and drawbacks of each Web 2.0 tool, coupled with an ability to tailor such tools to particular cultural contexts. It requires some reflection on how to address pedagogical, social, sociopolitical and philosophical limitations on the use of Web 2.0 in education. In all of the above, teacher training has an important role to play.

Beyond this, if language and literacy educators are to fully exploit the potential of Web 2.0 as a platform to enhance language teaching and to help their students become more sophisticated users of language(s) within — and beyond — the digital environment, they need to adopt an open, exploratory and flexible attitude. They need to appreciate and work with the social orientation of Web 2.0. They need to become comfortable with linguistic and media mashups and actively foster the codeswitching and shuttling skills demanded by the untidy realities of globalization, on- and offline. And, while continuing to provide the same level of educational input and guidance as good teachers have always done, they need to trust the digital natives to help them map what, for education, is still largely uncharted territory.

REFERENCES

Al-Sa'Di, R. A., & Hamdan, J. M. (2005). Synchronous online chat English: Computer-mediated communication. *World Englishes*, *24*(4), 409–424.

Alexander, B. (2006). Web 2.0: A new wave of innovation for teaching and learning? *EDUCAUSE Review*, *41*(2), 32-44. Retrieved November 29, 2007, from http://www.educause.edu/ir/library/pdf/erm0621.pdf

Anderson, T. (2006). Interaction in learning and teaching on the educational Semantic Web. In C. Juwah (Ed.), *Interactions in online education: Implications for theory and practice* (pp.141-155). London: Routledge.

Barney, D. (2007). The question of education in technological society. In J. Lockard & M. Pegrum (Eds.), *Brave new classrooms: Democratic education and the Internet* (pp. 271-284). New York: Peter Lang.

Baron, N. (2008). Adjusting the volume: Technology and multitasking in discourse control. In J.E. Katz (Ed.), *Handbook of mobile communication studies* (pp.177-194). Cambridge, MA: MIT Press.

Barton, D., & Hamilton, M. (2000). Literacy practices. In D. Barton, M. Hamilton & R. Ivanič (Eds.), *Situated literacies: Reading and writing in context* (pp. 7-15). London: Routledge.

Bax, S. (2003). CALL – Past, present and future. *System*, *31*, 13–28. doi:10.1016/S0346-251X(02)00071-4

Belz, J. A., & Thorne, S. L. (2006). Introduction: Internet-mediated intercultural foreign language education and the intercultural speaker. In J.A. Belz & S.L. Thorne (Eds.), *Internet-mediated intercultural foreign language education* (pp. viii-xxv). Boston: Heinle & Heinle. boyd, d. (2006). *Identity production in a networked culture: Why youth heart MySpace*. Paper presented to American Association for the Advancement of Science, 19 February. Retrieved November 29, 2007, from http://www.danah.org/papers/AAAS2006.html
boyd, d. (2008). Facebook's privacy trainwreck: Exposure, invasion, and social convergence. *Convergence, 14*(1), 13-20.

Burgmer, C. (2006). *Computer-mediated communication (CMC): A beneficial tool for modern language learning?* Unpublished M. Phil. dissertation, University of Dundee, UK.

Byram, M. (1997). *Teaching and assessing intercultural communicative competence*. Clevedon: Multilingual Matters.

Canagarajah, S. (2003). Foreword. In G. Smitherman & V. Villanueva (Eds.), *Language diversity in the classroom: From intention to practice* (pp. ix-xiv). Carbondale, IL: Southern Illinois University Press.

Chambers, A., & Bax, S. (2006). Making CALL work: Towards normalisation. *System*, *34*, 465–479. doi:10.1016/j.system.2006.08.001

Coghlan, M. (2007). *Language learning in a connected world.* Paper presented at New and Emerging Technologies in ELT, Loyola College, Chennai, India, 3-5 Aug.

Coleman, J. (2007). *Eeyore and the pixel drop-out: What's wrong with technology-enhanced learning?* Paper presented at New and Emerging Technologies in ELT, Loyola College, Chennai, India, 3-5 Aug.

Conrad, D. (2002). Inhibition, integrity and etiquette among online learners: The art of niceness. *Distance Education*, *23*(2), 197–212. doi:10.1080/0158791022000009204

Cope, B., & Kalantzis, M. (Eds.). (2000). *Multiliteracies: Literacy learning and the design of social futures*. London: Routledge.

Corbett, J. (2003). *An intercultural approach to English language teaching*. Clevedon: Multilingual Matters.

Crystal, D. (2001a). *Language and the Internet.* Cambridge: Cambridge University Press.

Crystal, D. (2001b). A linguistic revolution? *Education Communication and Information, 1*(2), 93–97.

Dalgarno, B. (2001). Interpretations of constructivism and consequences for computer assisted learning. *British Journal of Educational Technology*, *32*(2), 183–194. doi:10.1111/1467-8535.00189

Davies, W. (2003). *You don't know me but… Social capital and social software.* London: The Work Foundation. Retrieved November 29, 2007, from http://www.theworkfoundation.com/Assets/PDFs/you_dontknowme.pdf

Doctorow, C. (2002). My blog, my outboard brain. *O'Reilly WebDev Center*, 31 May. Retrieved November 29, 2007, from http://www.oreillynet.com/lpt/a/javascript/2002/01/01/cory.html

Downes, S. (2007). Places to go: Facebook. *Innovate, 4*(1). Retrieved November 29, 2007, from http://www.innovateonline.info/index.php?view=article&id=517

Dudfield, A. (1999). Literacy and cyberculture. *Reading Online*. Retrieved November 29, 2007, from http://www.readingonline.org/articles/dudfield/main.html

Ess, C. (2007). Liberal arts and distance education: Can Socratic virtue and Confucius' exemplary person be taught online? In J. Lockard & M. Pegrum (Eds.), *Brave new classrooms: Democratic education and the Internet* (pp. 189-212). New York: Peter Lang.

Finger, G., Russell, G., Jamieson-Proctor, R., & Russell, N. (2007). *Transforming learning with ICT: Making it happen*. Frenchs Forest: Pearson.

Freedman, T. (2006). Introduction. In T. Freedman (Ed.), *Coming of age: An introduction to the new world wide web* (pp.13-14). Ilford: Terry Freedman. Retrieved November 29, 2007, from http://fullmeasure.co.uk/Coming_of_age_v1-2.pdf

Friedman, T. L. (2006). *The world is flat: The globalized world in the twenty-first century* (2nd ed.). London: Penguin.

Garrison, D. R., & Anderson, T. (2003). *E-learning in the 21st century: A framework for research and practice*. London: Routledge Falmer.

Goodfellow, R. (2003). *Literacies, technologies, and learning communities: Speaking and writing in the virtual classroom*. Paper presented at EuroCALL 03, Limerick, Ireland, 5 Sept. Retrieved November 29, 2007, from http://iet.open.ac.uk/pp/r.goodfellow/Euroc03/talk.htm

Harvard University. (n.d.). *Handheld augmented reality project (HARP)*. Retrieved November 29, 2007, from http://isites.harvard.edu/icb/icb.do?keyword=harp

Haythornthwaite, C. (2007). Digital divide and e-learning. In R. Andrews & C. Haythornthwaite (Eds.), *The Sage handbook of e-learning research* (pp.97-118). Los Angeles: Sage.

Heckman, R., & Annabi, H. (2005). A content analytic comparison of learning processes in online and face-to-face case study discussions. *Journal of Computer-Mediated Communication, 10*(2). Retrieved November 29, 2007, from http://jcmc.indiana.edu/vol10/issue2/heckman.html

Hiltz, S. R., & Goldman, R. (2005). What are asynchronous learning networks? In S.R. Hiltz & R. Goldman (Eds.), *Learning together online: Research on asynchronous learning networks* (pp.3-18). Mahwah, NJ: Lawrence Erlbaum.

Holmes, B., Tangney, B., FitzGibbon, A., Savage, T., & Mehan, S. (2001). *Communal constructivism: Students constructing learning for as well as with others*. Dublin: University of Dublin Department of Computer Science. Retrieved November 29, 2007, from https://www.cs.tcd.ie/publications/tech-reports/reports.01/TCD-CS-2001-04.pdf

Hoppe, H. U., Joiner, R., Milrad, M., & Sharples, M. (2003). Guest editorial: Wireless and mobile technologies in education. *Journal of Computer Assisted Learning*, *19*(3), 255–259. doi:10.1046/j.0266-4909.2003.00027.x

Hubbard, P. (2004). Learner training for effective use of CALL. In S. Fotos & C. Browne (Eds.), *New perspectives on CALL for second language classrooms* (pp. 45-68). Mahwah, NJ: Lawrence Erlbaum.

Jonassen, D. H. (1992). Evaluating constructivistic learning. In T. M. Duffy & D. H. Jonassen (Eds.), *Constructivism and the technology of instruction: A conversation* (pp. 137-148). Hillsdale, NJ: Lawrence Erlbaum.

Kaye, A. (1989). Computer-mediated communication and distance education. In R. Mason & A.R. Kaye (Eds.), *Mindweave: Communication, computers and distance education* (pp. 3-21). Oxford: Pergamon.

Kazan, T. S. (2007). Braving the body: Embodiment and (cyber-)texts. In J. Lockard & M. Pegrum (Eds.), *Brave new classrooms: Democratic education and the Internet* (pp. 251-269). New York: Peter Lang.

Keen, A. (2007). *The cult of the amateur: How today's Internet is killing our culture.* London: Nicholas Brealey.

Kern, R. (2006). Perspectives on technology in learning and teaching languages. *TESOL Quarterly*, *40*(1), 183–210.

Kist, W. (2004). *New literacies in action: Teaching and learning in multiple media*. New York: Teachers' College Press.

Kramsch, C. (1993). *Context and culture in language teaching*. Oxford: Oxford University Press.

Kramsch, C. (1998). The privilege of the intercultural speaker. In M. Byram & M. Fleming (Eds.), *Language learning in intercultural perspective: Approaches through drama and ethnography* (pp. 16-31). Cambridge: Cambridge University Press.

Kramsch, C., A'Ness, F., & Lam, W. S. E. (2000). Authenticity and authorship in the computer-mediated acquisition of L2 literacy. *Language Learning & Technology*, *4*(2). Retrieved 29 Nov., 2007 from http://llt.msu.edu/vol4num2/kramsch/default.html

Lakshimi, S. K. C. (2007). *The role of teachers and the modern tool of online community networks in facilitating social interaction.* Paper presented at New and Emerging Technologies in ELT, Loyola College, Chennai, India, 3-5 Aug.

Lambe, P. (2006). Folksonomies and rich serendipity. *Green Chameleon*, 20 Oct. Retrieved November 29, 2007, from http://www.greenchameleon.com/gc/blog_detail/folksonomies_and_rich_serendipity/

Lee, S., & Berry, M. (2006). Effective e-learning through collaboration. In T. Freedman (Ed.), *Coming of age: An introduction to the new world wide web* (pp. 19-28). Ilford: Terry Freedman. Retrieved November 29, 2007, from http://fullmeasure.co.uk/Coming_of_age_v1-2.pdf

Levy, M., & Kennedy, C. (2005). Learning Italian via mobile SMS. In A. Kukulska-Hulme & J. Traxler (Eds.), *Mobile learning: A handbook for educators and trainers* (pp. 76-83). London: Routledge.

Levy, M., & Stockwell, G. (2006). *CALL dimensions: Options and issues in computer assisted language learning.* Mahwah, NJ: Lawrence Erlbaum.

Locke, T. (2007). E-learning and the reshaping of rhetorical space. In R. Andrews & C. Haythornthwaite (Eds.), *The Sage handbook of e-learning research* (pp. 179-201). London: Sage.

Macfadyen, L. P., & Doff, S. (2005). The language of cyberspace: Text, discourse, cultural tool. In C. Ghaoui (Ed.), *Encyclopedia of human-computer interaction*. Hershey, PA: The Idea Group. Retrieved November 1, 2006, from http://homepage.mac.com/leahmac/LM/Docs/C65MacfadyenDoff.pdf

McFarlane, A., Roche, N., & Triggs, P. (2007). *Mobile learning: Research findings. Report to BECTA.* Coventry: BECTA. Retrieved January 14, 2008, from http://partners.becta.org.uk/upload-dir/downloads/page_documents/research/mobile_learning_july07.pdf

McIntosh, E. (2006). Podcasting and wikis. In T. Freedman (Ed.), *Coming of age: An introduction to the new World Wide Web* (pp.71-75). Ilford: Terry Freedman. Retrieved November 29, 2007, from http://fullmeasure.co.uk/Coming_of_age_v1-2.pdf

Mitchell, P. (2005). Wikis in education. In J. Klobas et al., *Wikis: Tools for information, work and collaboration* (pp. 119-147). Oxford: Chandos.

Molina, P. G. 2006 EDUCAUSE Evolving Technologies Committee. (2006). Pioneering new territory and technologies. *EDUCAUSE Review, 41*(5), 112-134. Retrieved January 14, 2008, from http://www.educause.edu/ir/library/pdf/erm0659.pdf

Motteram, G., & Ioannou-Georgiou, S. (2007). *Are teachers fit for Web 6.0?* Paper presented at New and Emerging Technologies in ELT, Loyola College, Chennai, India, 3-5 Aug.

National School Boards Association [USA]. (2007). *Creating and connecting: Research and guidelines on online social—and educational —networking*. Alexandria, VA: National School Boards Association. Retrieved November 29, 2007, from http://www.nsba.org/site/docs/41400/41340.pdf

Norton, B. (2000). *Identity and language learning: Gender, ethnicity and educational change.* Harlow: Longman.

O'Reilly, T. (2005). What is Web 2.0: Design patterns and business models for the next generation of software. *O'Reilly,* 30 Sept. Retrieved January 14, 2008, from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html?page=1

Panganiban, R. (2007). SLCC Day 1: 'Second Life has the killer app, which is community'. *The Click Heard Round the World,* 25 Aug. Retrieved November 29, 2007, from http://www.rikomatic.com/blog/2007/08/slcc-day-1-seco.html

Pavlenko, A., & Blackledge, A. (2004). Introduction: New theoretical approaches to the study of negotiation of identities in multilingual contexts. In A. Pavlenko & A. Blackledge (Eds.), *Negotiation of identities in multilingual contexts* (pp. 1-33). Clevedon: Multilingual Matters.

Pegrum, M. (2005). Speed kills: Slowing down online language teacher training. In B. Beaven (Ed.), *IATEFL 2005: Cardiff Conference selections* (pp. 156-158). Canterbury: IATEFL.

Pegrum, M. (2007). Socrates and Plato meet neoliberalism in the virtual agora: Online dialog and the development of oppositional pedagogies. In J. Lockard & M. Pegrum (Eds.), *Brave new classrooms: Democratic education and the Internet* (pp. 13-34). New York: Peter Lang.

Pegrum, M. (2008a). Film, culture and identity: Critical intercultural literacies for the language classroom. *Language and Intercultural Communication*, *8*(2), 136–154. doi:10.1080/14708470802271073

Pegrum, M. (forthcoming, 2008b). From hype to hope: 10 myths of e-learning. In S. Ioannou-Georgiou & P. Pavlou (Eds.), *Learning technologies in the language classroom: Research and reflections.* Canterbury: IATEFL.

Pegrum, M., & Bax, S. (2007). *Catering to diversity through asynchronous online discussion: Linking teachers across continents.* Paper presented at Diversity: A Catalyst for Innovation — The 20th English Australia Conference, Sydney, Australia, 13-15 Sept.

Pennycook, A. (2007). *Global Englishes and transcultural flows.* London: Routledge.

Pesce, M. (2007). *Challenges and opportunities: Peer-produced knowledge and Australian education.* Education.au. Retrieved November 29, 2007, from http://blogs.educationau.edu.au/gputland/wp-content/uploads/2007/08/challenges-and-opportunities.doc

Phipps, A., & Gonzalez, M. (2004). *Modern languages: Learning and teaching in an intercultural field.* London: Sage.

Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon, 9*(5). Retrieved November 29, 2007, from http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf

Quinton, S. (2006). A brief critique on the future of learning: Assessing the potential for research. In D.L. Fisher & M.S. Khine (Eds.), *Contemporary approaches to research on learning environments: Worldviews* (pp. 543-578). Singapore: World Scientific Publishing.

Reddekopp, C. (2006). Multitasking and student learning: Possible explanations for a growing phenomenon. *Occasional Papers in Educational Technology.* Saskatoon: University of Saskatchewan. Retrieved November 29, 2007, from http://www.usask.ca/education/coursework/802papers/redekopp/index.htm

Reeder, K., Macfadyen, L. P., Chase, M., & Roche, J. (2004). Falling through the (cultural) gaps? Intercultural communication challenges in cyberspace. In F. Sudweeks & C. Ess (Eds.), *Fourth International Conference on Cultural Attitudes towards Technology and Communication* (pp.123-134). Perth: Murdoch University.

Rheingold, H. (2007). *Vision of the future*. Paper presented to Education.au, 2 October. Retrieved November 29, 2007, from http://www.educationau.edu.au/jahia/Jahia/pid/521

Ricento, T. (2005). Considerations of identity in L2 learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 895-910). Mahwah, NJ: Lawrence Erlbaum.

Ryan, Y. (2007). *Do you YouTube? Wanna come to MySpace? Musings on narcissism in the 21st century student.* Paper presented at First Year in Higher Education Conference, QUT, Brisbane, Australia. Retrieved November, 29, 2007, from http://www.fyhe.qut.edu.au/past_papers/papers07/final_papers/Do%20you%20YouTube%20_2_.pdf

Sabre, G. (2007). SL education vs personal beliefs. *The Metaverse Journal,* 10 Sept. Retrieved January 14, 2008, from http://www.metaversejournal.com/2007/09/10/sl-education-vs-personal-beliefs/

Selber, S. A. (2004). *Multiliteracies for a digital age*. Carbondale, IL: Southern Illinois University Press.

Sharples, M., Taylor, J., & Vavoula, G. (2007). A theory of learning for the mobile age. In R. Andrews & C. Haythornthwaite (Eds.), *The Sage handbook of E-Learning research* (pp. 221-247). London: Sage.

Stevens, V. (2007). *Second Life and online collaboration through peer-to-peer distributed learning networks*. Paper presented at METSMaC Conference, Abu Dhabi, UAE, 17-19 Mar. Retrieved November 29, 2007, from http://www.homestead.com/prosites-vstevens/files/efi/papers/metsmac/metsmac_secondlife.htm

Stone, L. (2006). *Attention: The real aphrodisiac*. Paper presented at ETech, San Diego, 6-9 Mar. Retrieved November 29, 2007, from http://radar.oreilly.com/archives/2006/03/etech_linda_stone_1.html

Street, B. V. (1994). Struggles over the meaning(s) of literacy. In M. Hamilton, D. Barton & R. Ivanič (Eds.), *Worlds of literacy* (pp.15-20). Clevedon: Multilingual Matters.

Tapscott, D., & Williams, A. D. (2006). *Wikinomics: How mass collaboration changes everything.* New York: Portfolio.

The Consultants-E. (2007, Nov.). A potted history of technology in teaching – present. *The Consultants-E Newsletter*. Retrieved January 14, 2008, from http://www.theconsultants-e.com/newsletter/The%20Consultants-E%20Newsletter%20-%20November%202007.pdf

The New London Group. (2000). A pedagogy of multiliteracies: Designing social futures. In B. Cope & M. Kalantzis (Eds.), *Multiliteracies: Literacy learning and the design of social futures* (pp. 9-37). London: Routledge.

The New Media Consortium/The EDUCAUSE Learning Initiative. (2007). *The horizon report: 2007 edition.* Austin, TX: The New Media Consortium. Retrieved November 29, 2007, from http://www.nmc.org/pdf/2007_Horizon_Report.pdf

Thorne, S. L. (2006). Pedagogical and praxiological lessons from Internet-mediated intercultural foreign language education research. In J.A. Belz & S.L. Thorne (Eds.), *Internet-mediated intercultural foreign language education* (pp.2-30). Boston: Heinle & Heinle.

Unsworth, L. (2001). *Teaching multiliteracies across the curriculum: Changing contexts of text and image in classroom practice*. Buckingham: Open University Press.

Wales, J. (2007). *Challenging how knowledge is created.* Education.au. Retrieved November 29, 2007, from http://www.educationau.edu.au/jahia/Jahia/home/challenging

Wallis, C. (2006). The multitasking generation. *Time*, *10*(April), 46–53.

Warschauer, M. (1999). *Electronic literacies: Language, culture and power in online education*. Mahwah, NJ: Lawrence Erlbaum.

Warschauer, M. (2003). *Technology and social inclusion: Rethinking the digital divide.* Cambridge, MA: MIT Press.

Warschauer, M. (2007). The paradoxical future of digital learning. *Learning Inquiry*, *1*(1), 41–49. doi:10.1007/s11519-007-0001-5

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity.* Cambridge: Cambridge University Press.

Williamson, B., & Facer, K. (2004). More than 'just a game': The implications for schools of children's computer games communities. *Education Communication and Information*, *4*(2/3), 253–268.

Wray, D. (Ed.). (2004). *Literacy: Major themes in education.* Vol. 4: New literacies: The impact of technology. London: Routledge Falmer.

Zieghahn, L. (2001). 'Talk' about culture online: The potential for transformation. *Distance Education*, *22*(1), 144–150. doi:10.1080/0158791010220109

KEY TERMS

Codeswitching: This term refers to the use of more than one language or language variety in a given context, for example to aid communication or to signal aspects of identity.

Continuous Partial Attention: According to Linda Stone, citizens of Western societies increasingly live in a state of *continuous partial attention*, as they continuously monitor multiple communication and information channels in an attempt not to miss anything. She argues that this is a post-multitasking behavior motivated less by the need to save time or be efficient than by the desire to always be connected to the network.

Folksonomy: An index produced in a bottom-up manner by adding user-generated tags to webpages of interest through a service such as del.icio.us. The resulting list of tags is known as a *folksonomy* and may be displayed in the form of a *tag cloud*, in which more prominent tags are shown in larger and darker type.

Mashup: This term, which stems from the hip hop practice of mixing music and/or lyrics from different songs to create new hybrids, can refer to web applications which combine data from different sources or, more commonly, to digital files which mix together pre-existing video, graphics, music, text, etc, in new combinations.

Social Constructivism: *Social constructivism* is a theory of learning which draws heavily on the work of the Soviet psychologist Lev Vygotsky (1896-1934). It suggests that learners add to and reshape their mental models of reality through social collaboration, building new understandings as they actively engage in learning experiences. Scaffolding, or guidance, is provided by teachers or more experienced peers in the learner's zone of proximal development, that is, the zone between what a learner can achieve independently and what s/he may achieve with support.

Third Place: This term is used by Claire Kramsch to refer to the space between cultures which language learners may reach as they develop intercultural (communicative) competence.

Web 1.0: A retrospective term which emerged after the advent of Web 2.0, *Web 1.0* refers to the original, information-oriented version of the World Wide Web. Created by Tim Berners-Lee in 1989/1990, it consisted of largely static web-pages developed by a small number of authors for consumption by a large audience.

This work was previously published in Handbook of Research on Web 2.0 and Second Language Learning, edited by M. Thomas, pp. 20-41, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 4.14
Adoption of Web Services in Digital Libraries: An Exploratory Study

Fatih Oguz
Valdosta State University, USA

ABSTRACT

This chapter describes a research study with an objective to explore and describe decision factors related to technology adoption. The study utilized theories of diffusion of innovations and communities of practice as frameworks and a case study of Web services (WS) technology in the digital library (DL) environment to develop an understanding of the decision-making process. A qualitative case study approach was used to investigate the research problems and data was collected through semistructured interviews, documentary evidence (e.g., meeting minutes), and a comprehensive member check. Face-to-face and phone interviews were conducted with respondents from five different DL programs in the U.S., selected based on distinctive characteristics (e.g., size of the DL program). Findings of the research suggest that the decision-making process is a complex procedure in which a number of factors are considered when making WS adoption decisions. These factors are categorized as organizational, individual, and technology-specific factors.

DOI: 10.4018/978-1-60566-042-4.ch014

INTRODUCTION

With the advent of the Internet and specifically the World Wide Web (WWW) application, means of accessing data and information have changed forever. The Internet brought great opportunities for libraries as well as dilemmas and problems, such as technology choice and readiness.

Digital libraries (DL) were envisioned as network-accessible repositories in the 1990s. Now, DLs extend the classical brick-and-mortar library concept, bring value to society, and transform information landscape by improving and changing the means of knowledge access, creation, use, and discovery across disciplines, regardless of temporal and geographical barriers (Larsen & Watctlar, 2003; Reddy & Wladawsky-Berger, 2001).

The speed of technological advances in information technologies (IT) in the last 10 years has enabled DLs to provide innovative resources and services to people. The information landscape is changing as a result of the revolutionary developments in IT, incompleteness of content on Internet, ever increasing digital content along with the evolution of networked technologies and applications, lack of standards, ineffective information retrieval mechanisms, and minimal cataloging. These factors present challenges to the future of DL development efforts (Borgman, 1999; Reddy & Wladawsky-Berger, 2001).

The concept of Web services (WS) has emerged as the next generation of Web-based technology for exchanging information. This effort began with the submission of the SOAP 1.1 to the World Wide Web Consortium (W3C) (Barefoot, 2002). WS are self-contained applications that can be described, published, invoked, and located over the Internet (or any network). Once a Web service is deployed, other applications can discover and invoke the service. WS provide a programmable interface for other applications without requiring custom programming and proprietary solutions regardless of the operating systems and programming languages to share information as opposed to providing users with a graphical user interface (Boss, 2004).

According to the W3C, a Web service is defined as a software system designed to support interoperable machine-to-machine interaction over a network by using XML for sending and receiving messages (Booth, Haas, McCabe, Newcomer, Champion, Ferris, et al., 2004). Simplicity and flexibility of XML made it a definitive standard for data transmission and storage. XML is an open standard and can be accessed and processed by any tool capable of reading and writing American standard code for information interchange (ASCII) text. By definition, the only requirement for a Web service is to use XML.

The basic WS platform is composed of XML and a transport protocol. HTTP is the commonly used transport protocol on the Internet (Hickey, 2003). XML, simple object access protocol (SOAP), and Web services description language (WSDL) are tools to create WS. A Web service provides the framework for creating the next generation of distributed systems by which organizations can encapsulate existing business processes, publish them as services, search for and subscribe to other services, and exchange information throughout and beyond the enterprise (Adams, Gisolfi, Snell, & Varadan, 2002). Besides recognizing heterogeneity of networked resources and applications as a fundamental ingredient, WS are independent of platform and the development environment can be packaged and published on the Internet. Also, WS enable just-in-time integration and interoperability with legacy applications (Oguz & Moen, 2006).

The development and widespread deployment of more intelligent knowledge environments that not only support scholarly inquiry and communication but also that are open, accessible to all, and transparent in their operation remains as a fundamental challenge for DL practitioners and researchers.

DL applications need to have some room to accommodate future technological innovations regardless how they are built, using off-the-shelf software vs. custom-built, and thus decision makers who include managers, coordinators, designers, and developers need to make important decisions at some point in time to adopt or reject an innovation, including a specific technology, application, framework or idea related with DLs. Decision makers who need information about an innovation may seek this information through both informal and formal communication channels while making such critical decisions.

In the context of DLs, roles and influence of informal communication channels on the decision-making process to adopt or reject WS technology has not been investigated before. The adoption of a new technology, WS, which is its early stages of adoption in the DL environment, may provide a significant opportunity to investigate decision factors. The goal of this study is to shed a light on the decision-making process to adopt or reject a new technology, WS, in the context of DLs.

As technologies rapidly change and the information landscape is transformed, DLs find themselves dealing with the issues of technology adoption decisions to exploit this dynamically changing technology environment to meet their users' needs and expectations. Therefore, understanding the decision-making process regarding adoption of WS technologies in the context of DLs is important.

BACKGROUND

This study used the diffusion of innovations (DOI) and communities of practice (CoPs) as theoretical frameworks and a case study of WS technologies in the DL environment to develop an understanding of the decision-making process.

Diffusion of Innovations and Communities of Practice as Theoretical Frameworks

The DOI research methodology provides required instruments, both quantitative and qualitative, to assess the rate and pattern of diffusion of an innovation and identifies various factors that facilitate or hinder its adoption and implementation (Fichman, 1992). These major factors include properties of the innovation, characteristics of adopters, and the means leading to adoption.

An innovation can be an idea, behavior, practice, or object perceived as new by the adopter (e.g., organization, individual). The concept of newness may be determined by the human reaction to it as well as the time passed since its discovery or first use. If the idea seems new to an individual or organization, it is considered an innovation (Daft, 1978; Rogers, 1995).

DOI researchers study the characteristics of the innovation to explain the rate of adoption of an innovation. Rogers (1995) classifies characteristics of innovations into five general categories: relative advantage, compatibility, triability, observability, and complexity. Innovations with greater relative advantage, compatibility, triability, observability, and less complexity are more likely to be adopted faster than others that lack these characteristics (Rogers, 1995). However, there are structural factors (e.g., formalization and centralization) as well as other innovation characteristics (e.g., cost, profitability, social approval) influencing adoption of an innovation, and therefore Rogers' DOI theory needs to be extended to accommodate such factors, specifically in organizational settings (Daft, 1978; Damanpour, 1991). In addition, Tornatzky and Klein (1982) found that relative advantage, compatibility, and complexity have the most consistent relationships with the adoption of innovations across a wide range of industries.

Rogers (1995, p. 23) defines a social system as "a set of interrelated units that are engaged in joint problem-solving to accomplish a common goal" and the members or units of a social system may be composed of individuals, organizations, and informal groups. Patterned social relationships (e.g., hierarchical positions) among the members of a social system define the social structure of a system which, in return, can facilitate or delay the diffusion of an innovation and lays out a framework for making predictions about the human behavior in a system since such structure provides regularity and stability to human behavior (Rogers, 1995).

Established behavior patterns called norms are the ruling principles of a social system, which may also influence diffusion (Rogers, 1995). In other words, norms serve as a guide or a standard for the members against which they can assess their own behavior. Norms may slow the diffusion process when an innovation does not comply with the norms of a social system even if the adoption of an innovation offers important benefits for the system (Raghavan & Chand, 1989).

The innovation-decision can be made by an individual member of a system as well as by the entire system. The decision can be made collectively by reaching a consensus among the members of a social system or by a relatively few individuals who possess status, power, or technical expertise (Rogers, 1995). A decision made by an individual to adopt or reject a new idea independently from other members of a system is called an optional-innovation decision (Rogers, 1995). An adoption decision may be influenced by the norms of the system and informal communication channels. In this case, the decision is made by an individual member of the system rather than the entire social system, and the individual member is fully responsible for the consequences of the decision. Collective-innovation decisions are made by members of a system through a consensus to adopt or reject a new idea. All the units within the social system are expected to comply with the decision. However, reaching a collective decision may be a time-consuming process because it is made by a consensus among the members. Authority-innovation decisions are made by a select set of members of a social system who have authority and higher status in the organizational chart; in this decision-making process an individual member has little or no influence on the decision. In organizational settings, collective and authority-innovation decisions are more common than the optional-innovation decisions, and authority-innovation decisions result in higher rate of adoption than others (Rogers, 1995).

Diffusion of an innovation is a social process that is influenced by various factors such as characteristics of the innovation (e.g., relative advantage) and the decision-making unit (e.g., individual characteristics), depending on the level of adoption (individual vs. organizational). The information about the innovation is communicated through formal (e.g., mass media) and informal in the course of the innovation-decision process. Rogers (1995) suggests that having some exposure to mass media and informal communication channels such as interpersonal networks increases a potential adopter's chance of knowing about an innovation earlier than others. This chapter specifically focuses on CoPs which serve as an informal communication channel.

CoPs are composed of people who share a concern, common problems, or a passion about the domain, and who want to gain more knowledge and expertise pertaining to the domain by interacting regularly (Wenger, McDermott, & Snyder, 2002). CoPs provide a learning environment through social participation, where participation refers to being active participants in the practice and building a sense of identity associated with the CoP to which they belong.

CoPs embody individuals with diverse backgrounds and social structures (e.g., other CoPs, organizations), which in turn, reduce the learning curve and rework, and promote innovation by enabling them to share and disseminate both tacit and explicit knowledge (Lesser & Storck, 2001). Sharing tacit knowledge requires personal interaction and CoPs provide such an informal learning platform through conversation and apprenticeship, for example. Members of the community become aware of their peers' expertise, knowledge, and skills through creating a venue for them to interact with each other. Thus, they are able to compare, verify, and benchmark their professionally developed expertise in the field against their colleagues' knowledge. When these benefits of CoPs are considered, their contribution to DL development efforts is vital in making informed technology, specifically WS, adoption decisions. The literature (e.g., Borgman, 1999; Marchionini, 1998) and nature of DL development efforts (e.g., open source) suggest the existence of informal structures such as CoPs.

As the organizations, specifically commercial ones, expand in size, geographical coverage, and complexity, knowledge has become the key to improving organizational performance and the formation of informal social groups like CoPs become a natural part of organizational life (Lesser & Storck, 2001; Wenger et al., 2002). CoPs make knowledge an integral part of their ongoing activities and interactions. Interpersonal interactions play an important role, especially in sharing tacit knowledge; the learning tools utilized by CoPs such as storytelling, conversation, and apprenticeship increase the efficient use of knowledge. CoPs act as a "living repository" for collective knowledge through creating a value for both the members and the organizations supporting and sponsoring these social structures (Wenger et al., 2002).

The DL conferences, funding agencies, workshops, and professional societies (e.g., Association for Computing Machinery) play important roles both in building and cultivating the CoPs in the DL field, and such meetings serve as a breeding ground for future collaboration in DL development efforts (Borgman, 1999). In addition, the experts in the field reached a consensus that "efforts associated with development of digital libraries are primarily collaborative" in a Delphi study conducted by Kochtanek and Hein (1999, p. 253).

Web Services in Digital Libraries

In general, DLs enable far broader range of users than traditional physical and organizational arrangements (e.g., libraries) to access information. Gathering, organizing, sharing, and maintaining such information resources require a flexible, scalable, and interoperable infrastructure (Larsen & Watctlar, 2003). Interoperability is an important issue where various system architectures, operating systems, and programming languages are required to communicate with each other. In addition, DL development efforts are closely related with the progress in general purpose technologies such as high-speed networking, security, and interoperability (Marchionini, 1998). However, the size, heterogeneity, and complexity of the today's information resources become critical factors when building DL systems because such factors may create immense challenges for interoperability, or the ability to ensure seamless information exchange across multiple DLs and information resources (Akscyn & Witten, 1998; Gonçalves et al., 2002; Marchionini, 1998). Marchionini (1998) addresses interoperability in two levels. The first level is the efforts to create standards for data storage and transmission, for query representation, and for vocabulary control; DLs adopt such standards and modify their content and services at the local level. However, standards development is a complex social process and requires consensus among stakeholders (Moen, 1997). The second level encourages individual DLs to create standards-based services that can be easily accessible and used by other DLs.

A vision set forth for the DLs by the President's Information Technology Advisory Committee (PITAC) Panel on Digital Libraries is that of providing the means of searching and accessing all human knowledge anytime and anywhere via Internet for all citizens (Reddy & Wladawsky-Berger, 2001). One of the key issues in accomplishing this vision is improving the ability to store and retrieve digital content across disparate and independent systems and collections by improving interoperability among diverse DL implementations (Reddy & Wladawsky-Berger, 2001). Thus, interoperability is an important factor to consider in the DL environment when making decisions to adopt WS technologies.

Important decisions have been made in the past as to adopt or reject a new technology for various reasons including the pursuit of this vision, delivering content in more efficient and advanced manner, and social status (e.g., being a pioneer in offering new DL services) (Pasquinelli, 2002). Some of the key technologies and standards related with interoperability that have been adopted in the past in DL environments such as the ANSI/NISO Z39.50 protocol, open archives initiative protocol for metadata harvesting (OAI-PMH), and open URL.

Hickey (2003) lists various ways of using WS technology in DLs from registering different types of objects and search services to navigating hierarchies and decomposing objects into simpler objects. The search/retrieve Web service (SRW) is a standardized Web service built on the 20 years of experience of the Z39.50 information retrieval protocol. SRW provides an easy way to implement the protocol with the power of older and more complex Z39.50 (Sanderson, 2004). Even some libraries are replacing Z39.50 with WS technologies as the protocol of choice between library portals and online electronic resources (Boss, 2004). WS facilitate access to electronic databases, and digital libraries providing access to such resources benefit from this technology (Boss, 2004).

The flexible and extensible digital object and repository architecture (Fedora) system, designed by the Cornell University Information Science and The University of Virginia Library's Digital Library Research and Development Group (DLR&D), is a promising open source digital library software initiative. Fedora was originally implemented based on CORBA architecture; however, the next release (Fedora 2.0) has adopted a service-oriented approach (SOA) based on WS ("Tutorial 1: Introduction," 2005). DSpace is another open source system, developed by Hewlett-Packard and MIT Libraries, to store the digital research and educational material produced by an organization or institution as a repository. Greenstone is yet another open source digital library software from New Zealand Digital Library Project at the University of Waikato that has a focus on publishing (Don, Buchanan, & Witten, 2005). The DELOS network pays close attention and contributes to the use of WS technologies in digital libraries. EBSCO publishing, a provider of a broad range of full-text and bibliographic databases, has introduced its WS interface to EBSCOhost, an electronic journal service for academic and corporate subscribers, forming a basis of real-time communications among library systems, portals, and all other systems in the future (Boss, 2004).

The major strength of WS is its reliance on XML. Given the characteristics of WS technologies and current use in DLs and e-commerce, WS are poised to play an important role as a technology providing interoperable standards-based access to DLs.

MAIN THRUST OF THE CHAPTER

This chapter attempts to explore and describe factors, activities, processes, and forces involved in the decision-making process related to adoption of WS technologies in DLs.

Research Problems and Methodology

The research strategy consisted of two components: a qualitative methodology and a case study. This strategy provided a framework of methods and data that would yield answers to the two research questions: (1) What are the key decision factors that lead decision makers to adopt or reject WS in the DL environment? and (2) What are the activities and entities that influence the decision regarding adoption of WS technologies in the DL environment?

The exploratory and descriptive nature of the study justified the use of a qualitative research approach that allows discovery and description of the social processes involved in decision making. Although quantitative methods have been predominant in information technology (IT) adoption research (Choudrie & Dwivedi, 2005), this chapter aims to develop a better understanding of decision factors influencing adoption of WS technologies in the context of DLs.

In-depth information about this complex social process involving decision makers was acquired through semistructured interviews and documentary evidence (e.g., meeting minutes and reports). The interview respondents and academic libraries that they are associated with were selected based on characteristics of DL programs identified by Greenstein and Thorin (2002). These characteristics included age of the program, staff size, and organization and orientation of the program. Seven respondents with different responsibilities (administrative vs. technical) were interviewed from five different DL programs in the US. These DL programs included big (i.e., staff size) programs such as California Digital Library and University of Texas at Austin and relatively smaller ones such as University of North Texas and University of Texas at Dallas.

Following Patton's (2002) guidelines, purposeful sampling, specifically maximum variation sampling, was employed when selecting the respondents who had the best knowledge, expertise, and overview about the topic of the research. The maximum variation sampling aimed at "capturing and describing the central themes that cut cross great deal of variation" (Patton, 2002, p. 234).

The respondents were from DL programs at the California Digital Library, University of North Texas, University of Texas at Dallas, University of Texas at Austin, and a university in the American Southeast. Some of the participating libraries are members of various influential professional societies and organizations in the DL field, including Digital Library Federation (DLF), Association of Research Libraries (ARL), and Coalition for Networked Information (CNI). Seven interviews were conducted with administrators and technical personnel who were involved in the decision-making process at these five academic libraries' DL programs.

Patton (2002) sets no rules for the sample size in qualitative inquiry by arguing that "the validity, meaningfulness, and the insights generated from qualitative inquiry have more to do with the information richness of the cases selected and the observational/analytical capabilities of the researcher than with sample size" (p. 245). The researcher stopped conducting interviews when data saturation was reached to meet the research goal, that is, to understand and describe decision factors related to WS adoption. Data saturation is defined as the point in a data collection process where new information becomes redundant (Bogdan & Biklen, 1992). Romney, Batchelder, and Weller (1986) conclude that samples as small as four participants could be enough to meet research objectives where purposeful sampling is carefully carried out to include information-rich respondents.

Documentary evidence provided additional and clarifying information supplemental to the data collected through interviews. Further, a comprehensive member check was conducted which allowed to obtain additional information from respondents and to have study findings reviewed by them. This final verification process allowed respondents to evaluate the researcher's interpretation of findings and analysis of data from their perspectives (Lincoln & Guba, 1985; Patton, 2002).

Results and Findings

Data revealed a number of factors that influenced and informed the decision-making process in WS adoption. These factors are categorized at organizational, individual, and technical levels.

Characteristics of DL programs that appeared to influence the decision-making process and categorized as organizational level factors included: organizational culture, program's relationships with surrounding academic units and external partners, management style and work structure, focus and direction of a program, formalization (e.g., flexibility in hierarchal order), functional differentiation in a program, size and age of a program, administrative attitude toward change, financial resources, technology readiness (e.g., expertise, technology infrastructure), and program's expectations (e.g., user needs). These organizational level factors appeared to play a critical role, especially in influencing members' information-seeking and communication behaviors. Individual level factors included members' connectedness with their colleagues, skill-set (e.g., competence), participation in CoPs, perception of organizational culture and goals, and openness to new ideas. Technical level factors included: interoperability, modularity, scalability, flexibility, addressability, rapid deployment of services, subscription service, and open-standards base of WS.

Some of the organizational level factors include management style, focus and direction of the program, size and age of the program, and organizational culture. Organizational level factors were closely associated with the organization itself and indirectly impacted by a DL program's staff, for example. Administrative personnel had an influence on some of organizational level factors (e.g., management style). Other factors could be regarded as more individual characteristics of DL staff members in terms of their information-seeking and communication behavior. Impact of individual level factors on decision-making vary from one technology to another depending on role of the technology (i.e., mission critical vs. non-mission critical) in the DL program.

Respondents identified financial concerns as a critical factor in guiding technology adoption decisions, and these concerns included: initial cost, ongoing cost, payoff, budgetary restrictions, and funding requirements. However, the extent of influence of these factors on WS adoption decisions appeared to vary from one DL program to another depending on the DL program's expectations from the technology and needs, focus, and direction of the program. These expectations and needs were closely related with size and age of the program. Data suggested that as programs grew in size over time, so did their collections, responsibilities, and user expectations. For example, although respondents formed positive opinions regarding open source software, they were aware that the lack of necessary skills in the program would be an important factor when getting a project initiated and providing technical support if they chose to use open source software. In addition, acquiring necessary technology skills through hiring new staff members and additional training were also factors impacting cost. Lack of technical expertise as a decision factor appeared to reflect the importance of Davis' (1989) "ease of use," Tornatzky and Klein's (1982) "ease of operation," and Rogers' (1995) "complexity" as innovation characteristics since adopters' technical background and skills were closely associated with perception of these characteristics.

Technology readiness of the DL program was another organizational factor that appeared to have an influence in the decision-making process. Technology readiness had two aspects: a human aspect (e.g., expertise, staffing), and the technological compatibility of WS technologies with existing technical infrastructure (i.e., hardware, software, and standards). In addition, technology readiness was also closely associated with availability of financial resources in case a hardware or software upgrade was needed. Respondents noted that compatibility of WS technologies with their existing technological infrastructure was an important factor that informed the decision-making. Tornatzky and Klein (1982) found compatibility as one of the most addressed innovation attributes. Compatibility also refers to consistency of an innovation with existing values and norms of the DL program. Furthermore, technology readiness was also an important factor for triability purposes. Small scale experiments were generally conducted in DL programs prior to making an adoption decision.

Individual level adoption decisions could be made especially for the use of WS technologies in nonmission critical applications. A personal positive experience with WS, existing skill-set, potential benefits for the work (i.e., Davis' [1985] perceived usefulness), and having easy access to experienced-based knowledge through CoPs appeared to influence an individual's perception and lowered the individual's learning curve. Technology and specific factors such as interoperability, modularity, flexibility, and WS subscription service were also decision factors in this case. In DL programs where WS had already been adopted, adoption decisions were made collectively and WS technologies were used for major and mission-critical applications. Both organizational and technology-specific factors were taken into consideration.

Another important decision factor was technology-specific benefits offered by WS, including interoperability, modularity, and open standards. WS provides an interoperable platform and is built on open standards (e.g., XML) where programs written in different programming languages and running on different operating systems are able to communicate with each other based on open standards and protocols. Data suggested that interoperability was an important factor since WS would not require major changes in existing technical infrastructure. In addition to interoperability, respondents identified additional technology-specific factors. For example:

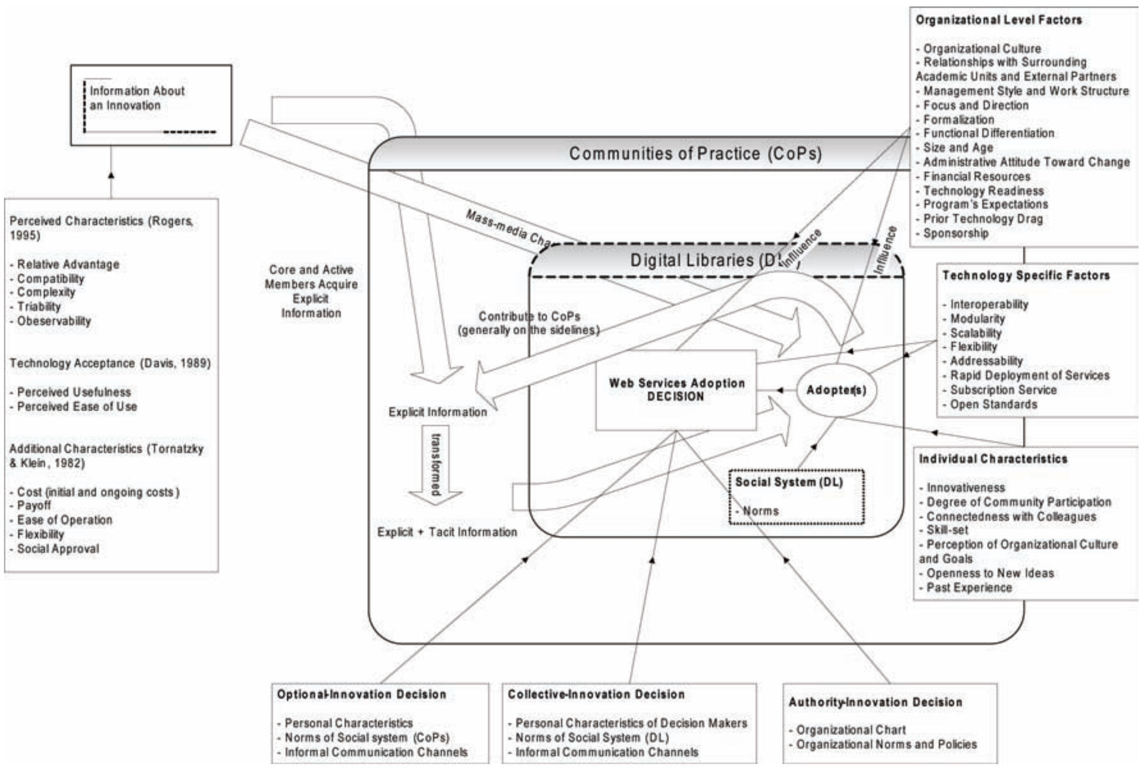
- **Modularity and flexibility:** Ability to act as building blocks to create distributed applications through reuse of existing applications that can be published and accessed over the internet or intranets.
- **Rapid deployment of Services:** Development time for new applications or services is greatly reduced through use of standard interfaces (e.g., WSDL) and open standards.
- Scalability: Ability to handle a growing amount of usage loads (e.g., Web caching, load balancing).
- **WS subscription service (UDDI):** A registry services for WS and allows other WS applications to automatically discover services and use them.

Small DL programs often looked to older and bigger DL programs when it came to adopting new technologies and standards. Their limited financial resources, staff size, and skill set were important barriers preventing them from taking initiatives that might be considered risky. These initiatives included technologies or standards that have not been tried or are in early stages of adoption in other DL programs. They often chose to rely on experiences of other DL programs so that they would be less likely to fail and run into unexpected problems. As for bigger programs, it appeared that they sometimes wanted to be the first or early adopters of some technologies in the DL environment to set an example for other programs. Though setting an example for other DL programs or wanting to be an early adopter was not a key decision factor, it was one of the factors occasionally taken into consideration in the decision-making processes.

Data suggested that availability of financial resources, focus and direction, size and age, collection size, users' and programs' expectations, and technology readiness were important factors influencing decision makers. Especially when making an optional-innovation decision, a potential adopter's existing technical skill-set and connectedness with the adopter's colleagues were key factors. At the technical level, interoperability, modularity, scalability, flexibility, addressability, rapid deployment of services, subscription services, and the open standards base of WS were key decision factors leading decision makers to adopt or reject WS in the DL environment.

Figure 1 presents a conceptual framework for the study informed by the theoretical frameworks in light of results and findings discussed earlier. Organizational and technical (i.e., technology specific) level factors have an influence both on adopters' information seeking behaviors and on the adoption decision itself depending on the type of innovation decision (i.e., optional, collective, and authority) made. Individual level factors (i.e., individual characteristics) guide adopters' information seeking activities (e.g., participation in CoPs) and influence their perception of organizational values which in turn inform adopters' contribution to the decision-making process. As shown in Figure 1, potential adopters may acquire information about an innovation through formal (e.g., mass media) and informal (e.g., interpersonal) communication channels. The information acquired through these channels includes perceived characteristics of an innovation that may play a key role as decision factors in the decision-making process. Data suggested that use of information collected through

Figure 1. Conceptual framework



these channels in the decision-making process vary due to adopters' degree of participation in CoPs, characteristics of knowledge (explicit vs. tacit) acquired, and factors at organizational, technical, and individual levels.

On the other hand, there were number of activities that members of DL programs participated in, entities that provided them with guidance, processes that helped them develop an understanding of WS, motivations that encouraged or discouraged them towards WS technologies, and forces that informed and guided their information seeking and communication behaviors. These activities, entities, processes, and forces were in play when making a decision regarding adoption of WS technologies in the DL environment.

DL programs had good ties with surrounding academic departments and information services, including other library departments, IT department, library and information school, and faculty. DL programs benefited from such connections not only by accessing their expertise but also by acquiring their content and collections.

DL staff members' interactions with others appeared to be maintained informally and the organizational structure of DL programs encouraged informal communication. Informality in communicating with others is one of the key characteristics of CoPs. Respondents interacted with their colleagues who were part of their own DL programs, as well as people from libraries, other university units, or external institutions and organizations to advance and share their knowledge and contribute to the field. Informal communication was also cited as an important part of the technology assessment process.

Some of the participating academic libraries were members of various professional associations and organizations which have an influence on DL-related issues, including use of DL technologies, digital preservation, standards, and DL development activities. Participation in these organizations (e.g., DLF, CNI, ARL) provided venues for DL programs to share their work and connect with other DL programs. The DLF promotes work on DL structures, standards, preservation, and use, CNI is interested in various areas critical to present and future of DLs, and ARL is one of the sponsor organizations of CNI and its member institutions are very active in the field. Further, these organizations engage in collaborative activities with each other in pursuit of their missions and goals. Collegial activities that were made possible through these relationships with these external entities appeared to play a central role in formation and continuation of informal communities that can be characterized as CoPs. Entities in the DL environment generally included CoPs, surrounding academic units and external partners, funding agencies, and the program itself.

In addition, attending national and international conferences were the most commonly used venues to obtain new information and served as a breeding ground for building personal contacts with colleagues. These collegial activities were regarded as communal activities. Preexisting personal contacts and the connections established in various venues with other institutions, organizations, and DL initiatives appeared to be very important for information access and sharing purposes.

CoPs attracted individuals with diverse backgrounds and skills from all around the world regardless of their geographical locations and provided an informal learning platform for their members. These CoPs were generally built and maintained in an online environment and occasionally supported with face-to-face interactions. In addition to distributed virtual CoPs, there were other CoPs, which may be subgroups of a broader CoP, that were locally networked and physically located. Participation of members in discussions in CoPs enabled online communities to cultivate and nurture knowledge acquired thorough experience, print, or other online resources and, in turn, these discussions enhanced and improved members' understanding of the technology. In other words, this mediating process gave rise to cross-fertilization of ideas and appeared to improve credibility of the knowledge generated and housed in CoPs. CoPs provided a living repository for the knowledge generated within the community while they were also perceived as places where up-to-date and quality information could be acquired. CoPs were also used to verify information acquired from different sources.

The conceptual framework (see Figure 1) helped to structure the data-found analysis in a format which may help the reader see this very complex landscape and understand this complex social process. Further, this chapter provided evidence that Rogers' DOI model needs to be complemented with organizational level factors identified by other researchers such as Daft (1978), Davis (1989), and Tornatzky and Klein (1992) to understand and describe diffusion of innovations in organizational settings.

CONCLUSION

This research was an exploratory and descriptive study to shed a light on the decision-making process to adopt or reject a new technology, WS, in the context of DLs and the unit of analysis was the decision to adopt or reject a new technology. The information landscape is transformed as technologies rapidly change and DLs often find themselves in a critical position to make a decision whether to adopt or reject emerging technologies such as WS.

Since the study employed a qualitative case study approach that supported the exploratory and descriptive nature of the research, results and findings of the study are not intended to be statistically generalizable to other technology adoption cases. However, detailed description in the narrative may assist the reader of this case study research to determine applicability of these findings to other technology adoption decisions in the reader's own setting.

This chapter provides evidence that CoPs as informal communication channels practice play an important role in enabling staff members of a DL program to access up-to-date and experienced-based knowledge, providing a distributed problem-solving and learning platform, facilitating informal communication and collaborative activities among DL programs, and informing the decision-making process. Technical characteristics (e.g., interoperability, open standards), compatibility with existing technical infrastructure, applicability to existing DL projects, total cost of ownership (e.g., licensing, maintenance cost), technical expertise in the DL program (e.g., staffing, training, learning curve), and success of a pilot project are cited as key decision factors influencing adoption Web services technologies in the DL environment. This chapter provides an adequate foundation for further research on the impact of organizational, individual, and technology-specific factors on decision-making processes in the DL environment.

The complexity of the decision-making process and the variety of factors that informed and influenced this process are reflected. A review of the relevant literature suggests that this is a complex process, and the findings inform and provide details about the complexity, as presented in Figure 1. The theoretical frameworks selected for this chapter proved useful to achieve the goal of the study. The chapter attempts to provide a complete account of decision factors related with adoption of WS technologies in the DL environment. An outcome of this study suggests that an exploratory and descriptive study such as this is an important step towards understanding the decision-making process as technologies rapidly change in the DL environment.

ACKNOWLEDGMENT

The author thanks Corrie Marsh for her valuable feedback in writing of this chapter and to Dr. William Moen for his guidance and support for this research.

REFERENCES

Adams, H., Gisolfi, D., Snell, J., & Varadan, R. (2002). Best practices for Web services: Part 1 - formation of a semantic framework. *IBM developerWorks*. Retrieved March 5, 2004, from http://www-106.ibm.com/developerworks/Webservices/library/ws-best1/

Akscyn, R. M., & Witten, I. H. (1998). *Report of First Summit on international cooperation on digital libraries.* Retrieved May 3, 2005, from http://Web.archive.org/Web/20010529104724/ks.com/idla-wp-oct98/

Barefoot, D. (2002). *Web services primer*. CapeScience.

Bogdan, R. C., & Biklen, S. R. (1992). *Qualitative research for education: An introduction to theory and methods* (2nd ed.). Boston: Allyn and Bacon.

Booth, D., Haas, H., McCabe, F., Newcomer, E., Champion, M., Ferris, C., et al. (2004, February). *Web services architecture* (W3C Working Group Note 11). The World Wide Web Consortium (W3C).

Borgman, C. L. (1999). What are digital libraries? Competing visions. *Information Processing & Management*, *35*(3), 227–243. doi:10.1016/S0306-4573(98)00059-4

Boss, R. W. (2004). *Web services.* Retrieved November 1, 2005, from http://www.ala.org/ala/pla/plapubs/technotes/Webservices.htm

Choudrie, J., & Dwivedi, Y. K. (2005). Investigating the research approaches for examining technology adoption issues. *Journal of Research Practice*, *1*(1).

Daft, R. L. (1978). A dual-core model of organizational innovation. *Academy of Management Journal*, *21*(2), 193. doi:10.2307/255754

Damanpour, F. (1991). Organizational inertia and momentum: A dynamic model of strategic change. *Academy of Management Journal*, *34*(3), 555–591. doi:10.2307/256406

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 318. doi:10.2307/249008

Don, K., Buchanan, G., & Witten, I. H. (2005). *Greenstone 3: A modular digital library.* Retrieved November 1, 2005, from http://www.sadl.uleth.ca/greenstone3/manual.pdf

Fichman, R. G. (1992, December). *Information technology diffusion: A review of empirical research.* Paper presented at the Thirteenth International Conference on Information Systems (ICIS), Dallas.

Gonçalves, M. A., Fox, E. A., Watsom, L. T., & Kipp, N. A. (2002). Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. [TOIS]. *ACM Transactions on Information Systems*, *22*(2), 270–312. doi:10.1145/984321.984325

Greenstein, D., & Thorin, S. E. (2002). *The digital library: A biography*. Digital Library Federation.

Hickey, T. B. (2003). *Web services for digital libraries*. Paper presented at the Cross Language Applications and The Web 27th Library Systems Seminar, Bern, Switzerland.

Kochtanek, T. R., & Hein, K. K. (1999). Delphi study of digital libraries. *Information Processing & Management*, *35*(3), 245–254. doi:10.1016/S0306-4573(98)00060-0

Larsen, R. L., & Watctlar, H. D. (2003). *Knowledge lost in information.* Report of the NSF Workshop on Research Directions for Digital Libraries.

Lesser, E. L., & Storck, J. (2001). Communities of practice and organizational performance. *IBM Systems Journal*, *40*(4), 831–931.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage Publications, Inc.

Marchionini, G. (1998). Research and development in digital libraries. In A. Kent (Ed.), *Encyclopedia of library and information science* (Vol. 63, pp. 259-279). New York: Marcel Dekker.

Moen, W. E. (1997). *The Development of ANSI/NISO Z39.50: A Case Study in Standards Evolution*. Unpublished Dissertation, Syracuse University.

Oguz, F., & Moen, W. E. (2006). *Texas library directory Web services application: The potential for Web services to enhance information access to legacy data.* Paper presented at the International Conference on Next Generation Web Services Practices (NWeSP'06), Korea.

Pasquinelli, A. (2002). *Digital library technology trends*. Retrieved December, 9, 2005, from http://www.sun.com/products-n-solutions/edu/whitepapers/pdf/digital_library_trends.pdf

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Raghavan, S. A., & Chand, D. R. (1989). Diffusing Software-Engineering Methods. *IEEE Software*, *6*(4), 81–90. doi:10.1109/52.31655

Reddy, R., & Wladawsky-Berger, I. (2001). *Digital libraries: Universal access to human knowledge: A report to the President*. President's Information Technology Advisory Committee (PITAC), Panel on Digital Libraries.

Rogers, E. M. (1995). *Diffusion of Innovations* (4th ed.). New York: The Free Press.

Romney, A., Batchelder, W., & Weller, S. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*, 313–338. doi:10.1525/aa.1986.88.2.02a00020

Sanderson, R. (2004). *SRW: Search/retrieve Webservice version 1.1*. Retrieved February 2, 2005, from http://srw.cheshire3.org/SRW-1.1.pdf

Tornatzky, L. G., & Klein, K. J. (1982). Innovation characteristics and innovation adoption-implementation: A meta-analysis of findings. *IEEE Transactions on Engineering Management*, *29*(1), 28–45.

Tutorial 1: Introduction. (2005). Retrieved August 15, 2005, from http://www.fedora.info/download/2.0/userdocs/tutorials/tutorial1.pdf

Wenger, E., McDermott, R., & Snyder, W. M. (2002). *A guide to managing knowledge: Cultivating communities of practice*. Boston: Harvard Business School Press.

This work was previously published in Managing Web Service Quality: Measuring Outcomes and Effectiveness, edited by K. Khan, pp. 307-320, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 4.15
A Context-Based Approach to Web 2.0 and Language Education

Gary Motteram
University of Manchester, UK

Susan Brown
University of Manchester, UK

ABSTRACT

Web 2.0 offers potentially powerful tools for the field of language education. As language teacher tutors exploring Web 2.0 with participants on an MA in Educational Technology and TESOL at the University of Manchester, UK, we see that the potential of Web 2.0 is intimately linked with teachers' perceptions of their teaching contexts. This chapter will describe a "context-based" approach to the exploration of Web 2.0 on a module focusing on the potential role of distributed courseware in language education. It will begin by giving an overall picture of where and how the exploration of Web 2.0 tools fits into the MA program. It will then describe the main aims and aspects of the module and discuss in some detail our context-based approach in relation to participants as well as Web 2.0 in existing literature. The chapter will conclude with two case studies concerning how teachers incorporate Web 2.0 technologies in courseware for their contexts.

DOI: 10.4018/978-1-60566-190-2.ch007

INTRODUCTION

This chapter explores the way that participants on a module run as part of an MA in Educational Technology and TESOL learn about, make use of and evaluate Web 2.0 technologies. This module is a new departure for the course and represents the ongoing need for the MA to refresh itself and to bring new and developing technologies into its domain.

Web 2.0 has its advocates and its detractors; however, it has become a *de facto* part of today's Internet landscape. The very nature of Web 2.0, its emphasis on such features as collaboration, interactivity and user-generated content, seems to make it an obvious choice for a focus of discussion when it comes to looking at current trends in the use of technology in language education. These trends, as shall be later discussed, reflect a focus on learner centered, collaborative tasks which, in Second Language Acquisition terms, allow channels for authentic language input and output (Chapelle, 1998).

However, it is important to realize that for many language teachers Web 2.0 may simply appear to be another technological innovation that will pass them by along with the many others that they have seen during their career, despite the slowly increasing range of references to the uses and benefits of key Web 2.0 technologies (e.g. blogs, podcasts and wikis) in language education.

As people and communities in various parts of the world increasingly embrace Web 2.0, some educational institutions are inevitably responding to those societal trends and trying to harness Web 2.0 in their learning programs. Others, although they are in societies where technology is more normalized (Bax, 2003a) have, for various reasons, not taken those technologies on board. Perhaps now, more than at any other time, language teachers may need to negotiate these changes as they impact, or not, on their institutions, and consider the implications of ever greater technology use for their language teaching. They may be inspired or effectively obliged to engage with the nature of Web 2.0 and analyze its affordances for language education. Other teachers, even if they are aware that it is being used in the wider world may currently see no application for it in language education.

Web 2.0 is described as relatively easy to use and therefore accessible to anybody with access to the Internet. This is in contrast to its Web 1.0 predecessor which is seen to require at least some familiarity with HTML as a minimum. Setting up and contributing to a blog for example may seem comparatively uncomplicated. However, once a blog has been set up the user may be confronted with concepts and technicalities that may be more difficult to get to grips with, RSS, by way of example. Teachers struggling to understand the concept behind RSS, and the different technologies that support it, are unlikely to be able to stand back and evaluate its uses in language learning terms.

Such issues notwithstanding Web 2.0 tools do appear to offer a lot of what language teachers would want in order to support learners language development: they can potentially distribute the learning and enable students to be in regular touch with a world-wide community of learners; they appear to enable an easier connection to be made between the classroom and the “real” world; they might enable learners to take some control over their learning making use of tools that excite them and which they are using in their everyday lives; they seem to offer engagement in active rather than passive learning, in process as well as product; learners can also potentially engage in discourses that take them beyond the classroom.

On the MA program in Educational Technology and TESOL at the School of Education, University of Manchester, it is important to explore Web 2.0 technologies in language education and help teachers understand generic functions of Web 2.0 in order to facilitate their evaluation of its potential uses. An evaluation of this potential should not, and cannot, be divorced from considerations pertaining to the “ecology” of the teaching environment in which teachers work, or have worked in the past and how that pertains to the wider changes in society. The use of the term “ecology” here signifies all of the rich, interacting elements that create the dynamic of teachers’ teaching contexts including top down societal, curricula and institutional influences and the bottom up influences which may stem from teachers’ knowledge of and enthusiasm for Web 2.0.

In this regard the MA tutors continue to observe that the way teachers “make sense” of Web 2.0 genres, i.e. understand how their various intrinsic operations — a process which is intimately bound up with teachers’ “evaluation” of the potential affordances of such software for their contexts — is mediated by the teachers’ perceptions of the context in which they work, or have worked in the past.

Considerations of context are bound up on the MA program with those of pedagogy, and the “fit” of Web 2.0 genres to pedagogical approaches. We have always worked as teacher-educators on the principle that the use of technology in language education should be firmly underpinned by considerations of pedagogy and appropriate methodology (Holliday, 1994). The functions of the technologies explored on the MA program are therefore considered according to how they can facilitate and possibly enhance pedagogical approaches that respond to the specificities of different contexts and the needs of learners in that context. In other words we adopt a context and pedagogic driven rather than a technology driven model. This central focus on pedagogy as it is relates to context and the role that technology can play in contexts has led us to evaluate the affordances of Web 2.0 as they might respond to contextual factors, what we have started to term a "niche" approach to evaluating Web 2.0

The main discussion in this chapter will centre on one of eight component modules that the MA program participants take, entitled "Courseware Development for Distributed and Blended Learning" (CDDBL) and the context-based approach to CDDBL introduced above. This module explores a range of Web 2.0 tools and how they may be exploited both for blending and distributing courseware. The chapter describes the module's explorations of Web 2.0 and tutors' evolving thinking about the way to best approach Web 2.0. The changing nature of participants on the module since its inception in 2005 is discussed and two case studies will illustrate ways in which participants have employed or how they envisage employing Web 2.0 in their own contexts.

In what follows the chapter gives the reader an overall view of the MA course and how Web 2.0 is included, describing our current approach to exploring Web 2.0 in CDDBL and discussing aspects of the literature informing the module. It then focuses on two case studies illustrating how former participants on the course evaluated Web 2.0 in relation to their own contexts. The chapter concludes by discussing possible future developments in our context approach to Web 2.0 on CDDBL.

THE MA PROGRAMME

Brief Overview of the Program

The participants on the Master's program in Educational Technology and TESOL come from different parts of the world including South America, Asia, the Middle East and Western and Eastern Europe and may be either non-native or native speakers of English. They all have at least three years teaching experience. This level of experience is a prerequisite for entry onto the degree as its whole focus is not on our (the tutors) forming and shaping of the participants' thinking about teaching, but on facilitating the reflective process that will allow the participants, drawing from their teaching experience, to shape their own thinking about their teaching. The participants' experiences of using technology vary and range from no use to a significant engagement. This obviously affects the extent to which they can reflect on their own practice using technology.

The MA was set up in the 1980s (in those days it was an MEd) to meet the needs of teachers who were becoming interested in using video and computers as part of their language teaching processes. The course has changed considerably over time, but still keeps as its main foundational aim a focus on the pedagogical implications of the uses of technology (see Wildner, 1999). The specific modules that are relevant to technology and language learning include: Language Learning and Technology, which explores the general uses made of technology in language classrooms to support language skills development; Multimedia in Language Education, which combines an exploration of second language acquisition processes with the design and development of language tasks using Web 1.0 technologies; Teaching and Learning Online, in which we ask the participants to explore and reflect on experiences of online learning. The fourth technology-focused module, Courseware Development for Distributed and Blended Learning (CDDBL), is the one that is described in detail in this chapter. Other modules that students do reflect a more general TESOL diet.

We have both onsite and offsite (distance) participants. The offsite participants study part time as they are generally practicing teachers and via an online virtual learning environment (VLE – currently WebCT). Onsite participants are studying for the most part full-time and are therefore removed from their teaching context, particularly if that context is not in the locality of Manchester or is overseas.

COURSEWARE FOR DISTRIBUTED AND BLENDED LEARNING

The nature of the program we offer means that there is a continual refreshment of the modules and CDDBL is the latest re-working, the first run of this new module taking place in 2005. The aim of the module is to assist in the development of skills that will enable teachers to review and create effective blended and distributed learning materials for their context, with all of the attendant considerations that this involves. While "Multimedia in Language Education" looks at materials design at task level, CDDBL considers the integration of activities at the level of a course or scheme of work.

Early on in the development of CDDBL we took the decision to focus primarily on Web 2.0 tools and their affordances in distributed courseware. We had originally intended to focus solely on Virtual Learning Environments (VLEs) but realized that in doing so we would be missing the opportunity to explore emerging technologies from the perspective of courseware development, technologies that potentially change the way we view that development, both in terms of the greater ease with which courseware might be created by tutors but also in the degree of control that the participants themselves have in the materials design process.

As we designed the first iteration of the module it occurred to us that unlike Web 1.0 technologies, where the extent of interactivity that a learner can engage in is more likely to be determined by the designer/tutor, with Web 2.0 technologies the development of courseware need not be the preserve of the tutor designer, but also of learners. Such differences between Web 1.0 and Web 2.0 technologies are explicitly discussed on CDDBL as are the ways that the two technology types can be effectively combined to suit the specificities of teaching contexts.

The current iteration of CDDBL therefore explores the following Web 2.0 tools: blogs; wikis; social bookmarking; e-portfolio software; and podcasting. It further focuses on two VLE platforms, WebCT and Moodle, an open source VLE which, in response to ongoing feedback from designers/tutors using the software, continues to have new tools incorporated into it, the majority of which are Web 2.0 tools such as blogs.

The assessment procedure for CDDBL requires participants to create sample courseware materials which combine Web 2.0 technologies and which address issues related to language learning in their context. They are currently asked to articulate the thinking behind their courseware through a 30-minute presentation and short executive summary and to discuss the courseware in relation to relevant educational literature.

PERSPECTIVES ON WEB 2.0 IN THE LITERATURE AND CDDBL

There are areas of difference and confluence between perspectives on Web 2.0 in the educational literature and our own perspective on important considerations relating to Web 2.0. in courseware development. We have said that the way participants on CDDBL evaluate the potential of Web 2.0 technologies is intertwined with their perceptions of the contexts in which they teach; this is having an increasing influence on the ongoing development of the module and on our approach to the exploration of Web 2.0 genres.

It is fair to say that the relationship between considerations of context in the educational literature and the nature and potential of Web 2.0 has not, as yet, been extensively explored. Much of the current literature on Web 2.0 in education discusses it from a general perspective, e.g. with regard to the uptake of Web 2.0 in society and particularly among the digital or net generation, and mainly with regard to tertiary education (see Oblinger, 2005; Bryant, 2006). Little of the discussion on Web 2.0 is, as yet, localized. This is not the case with discussion on Web 1.0 technologies, where a number of studies relate to the specificities of different local contexts (see, for example, Zhong & Shen, 2002). In CDDBL we explore with participants the general themes in the literature on Web 2.0; we provide a summary of some of these below.

Those in the field of education who write on Web 2.0 technologies see it as holding significant possibilities for the field. A lot of Web 2.0 discussion is subsumed under the epithet of "social software" which is perhaps both indicative of the cryptic nature of the term Web 2.0 and of the significance of the term "social" in the educational field where it is widely argued, partly based on the ideas of socio-cultural theorists such as Vygotsky (1978), that learning takes place through mediated social interactions. This potential is discussed in relation to the creation of new learning communities which may offer the "personalised collaborative learning experiences such as those that are already emerging in the world outside the school gates" (Owen et al., 2006, p. 11). Such communities can expand discussion beyond the classroom and provide new ways for students to collaborate within their class and across the world (Bryant, 2006). Wenger (1998) is regularly cited when discussions of the building of communities beyond classrooms is proposed.

As with the discussion in the educational field generally, the term Web 2.0 has not, as yet, been used extensively in the literature on language teaching. The tools associated with the term tend to be subsumed under the umbrella terms of Computer Mediated Communication (CMC) and social networking. Nevertheless the potential of those tools, as articulated in relation to social networking and CMC, is increasingly recognized. They may offer scope for exposure to, and production of, authentic language use in real life intracultural and intercultural Internet contexts (Kern, Ware & Warschauer, 2004). They also offer the learner the chance to use language as it is used on the Internet and be exposed to "emerging genres of language use" (Thorne & Payne, 2005, p. 372). Such opportunities for authentic language output and the concomitant opportunities for "noticing" and "negotiation of meaning" sit comfortably with notions of how SLA takes place (See Chapelle, 1998). Wikis for example, with their text editing features, may provide the learner opportunities to "correct their linguistic output" and "engage in target language interaction whose structure can be modified for negotiation of meaning" (Chapelle, 1998, pp. 23-24). With these opportunities for greater levels of authentic, autonomous language engagement more emphasis will need to be placed by teachers on the development of metacognitive skills among learners, i.e. the skills that learners need to order and develop their own learning. In some ways many Web 2.0 genres have inbuilt features, e.g. the wiki edit facility which can facilitate metacognitive thinking. Web 2.0 therefore, may offer the most genuine medium yet for breaking down the barriers between the classroom and the real world as not only can the learner use English in an authentic medium but that medium also provides the tools which allows learners to focus in an authentic way on how language is used. However, in the same way that Web 2.0 is an extension of Web 1.0 and can be seen to have some of its characteristics, it is not a good idea to view the uses of Web 2.0 technologies as somehow divorced from what has been occurring for many years in the world of Computer Assisted

Language Learning. This field has certainly advocated extensive use of a variety of technologies to promote language learning and has made use of a wide variety of tools to do this. It has also drawn heavily on popular theory from a range of contexts to support its practices. What is potentially different is the way that uses are more easily managed by the learners themselves and materials can be more easily learner generated.

In all the above examples the value of the learning that can potentially occur through Web 2.0 is seen in relation to the extent to which it is allied to, driven by and a part of the social, cultural and economic trends that are shaping the world. There is a prevailing discourse of urgency evident in some of the literature relating to technology in education, perhaps most pithily encapsulated in the phrase, “You can’t not do it” (Collis & Moonen, 2001). This discourse sees the world changing at speed, where economies will be driven by a technologically savvy population, where academic institutions will need to gear themselves to offering flexible learning programs through various technologies and where the “digital/net generation” is not only at home with digital technologies but will be increasingly mystified as to why they are not an integral feature of their education (Oblinger, 2005). If the net generation’s thinking and expectations are shaped by their experiences as net citizens and participants then they will bring those expectations into the educational context where Web 2.0 which is geared around interaction, will really count.

The literature identifies important caveats relating to the uptake of technology, not least the need for pedagogy to drive the way technology is used rather than the contrary. Salaberry, (2001) in his overview of the uses of technology and their impact on language learning during the twentieth century, makes the point that, “new technologies—revolutionary as they may be from a strictly technological point of view—are normally regarded as revolutionary from a pedagogical standpoint as well” (p. 39). Pedagogical approaches rooted in socio-cultural theory which views humans as embedded in learning communities where social activity, collaboration and interaction are prime factors in the learning process, are seen as fortuitously consonant with what Web 2.0 appears to have on offer. However, there is perhaps a tendency in the literature to assume that there is a direct unmediated link between Web 2.0 and socio cultural pedagogical approaches and that the introduction of Web 2.0 automatically engenders greater learner participation and interaction. Web 2.0 tools may be predicated on the user as broadcaster rather than audience, as creator rather than recipient (Horizon Report, 2007) but when such tools are harnessed in educational contexts, the way that the teacher designs and scaffolds activities within these tools has a prime affect on the extent of and ways that students participate. Web 2.0 tools may offer the teacher a malleable medium for moulding learner development but it is the teacher’s understanding of how best to work and craft that medium which may well determine how it works in a language teaching context. The importance of the tutor as designer is stressed to CDDBL participants.

The way that teachers choose to harness Web 2.0 will depend in large part on their teaching context and we are particularly careful on CDDBL that in focusing on the way pedagogy can be enhanced by technology we do not neglect considerations of context. While in the literature on language education there has been some discussion on the need for a “context approach” i.e. “an approach that places context at the heart of the profession” (Bax, 2003b, p. 278) and on an “ecological perspective” which looks at the dynamic and negotiated relationship between the richness of a teaching context and methodology (Tudor, 2003), a “context approach” tends to be sidelined when it comes to thinking of the triadic relationship between pedagogy, technology and context. In arguing that there are snug and beneficial fits between a technology and a single pedagogical approach, e.g. social constructivism, there is a risk of propounding a one-fits-all pedagogy which is unresponsive to specificities of context. In fact what we would contend is that the inherent flexibility of Web 2.0 can allow for a blended pedagogical approach which can respond to local educational contexts. There is an expanding strand in the literature that argues that technology, as it is harnessed in careful instructional design, can be effectively used in this way (see, for example, Alonso et al., 2005).

THE PARTICIPANTS' CONTEXTS

The importance of a "context approach" on CDDBL is underscored with every new cohort that participates on the module. CDDBL participants come from a multiplicity of teaching contexts around the globe, from South and South East Asia to the Middle East to South and North America to Eastern and Western Europe.

A preponderance of participants on CDDBL comes from low and mid-tech contexts. We describe low, mid and high-tech contexts here both from the perspective of the institution and of the learner (see Figure 1). The extent to which we consider an institution as low, mid or high tech depends on a number of factors; primary among these is the level of computer resources available to learners and teachers in the institution and the level of computer know-how among staff in the institution. From the learner perspective we describe context as low, mid or high-tech ostensibly according to the level of access they have to computer technology inside or outside of the teaching institution and their familiarity with that technology. While in a low or mid-tech context some teachers do use Web 1.0 technologies such as PowerPoint, it is more uncommon for Web 2.0 technologies to be used although a clutch of participants on the current run of the module are using Web 2.0 largely as resource areas for their learners. As yet we have not had any participants that we consider to be from high-tech contexts. By high-tech contexts we mean contexts where the use of technology has become "normalized," in the institution, or as part of the learning process (Bax, 2003a) that is to say where the use of technology has become an integral, assumed and unnoticed aspect of the learning process and where learners consider it perfectly natural to engage in language learning, as facilitated by technology, in the institution as well as in their own time outside of the educational environment.

THE IMPACT OF CONTEXT ON EVALUATION OF WEB 2.0 TECHNOLOGIES

We have mentioned the complex and interacting factors that make up the teaching contexts of participants on CDDBL. These will be different for every participant on the module and therefore are best represented on an individual basis (See section entitled Two Case Scenarios). The low, mid and high-tech categories however, provide us with a general starting point for analyzing the way participants on the CDDBL evaluate Web 2.0 technologies. A top down and bottom up perspective on the use of technologies also provide a useful conceptual framework for analyzing this dynamic. By top down is meant societal, curricula and institutional factors that push for the further integration of technologies. Bottom up means influences that may derive from teachers who are enthusiastic technology users and may see its potential in language education.

Our observations to date have led us to conclude that such contextual factors and the way participants represent these as "context-in-mind," in other words their perceptions of context as they see it in their mind's eye, (Brown, in preparation), have some impact on the way they evaluate the potential of Web. 2.0. We have, over recent months variously described this representation; we have used a "smorgasbord" versus "empty table" metaphor where the smorgasbord represents a participant's perception of the major potential of Web 2.0 in relation to context and where the empty table is

Figure 1. High-tech and low-tech context

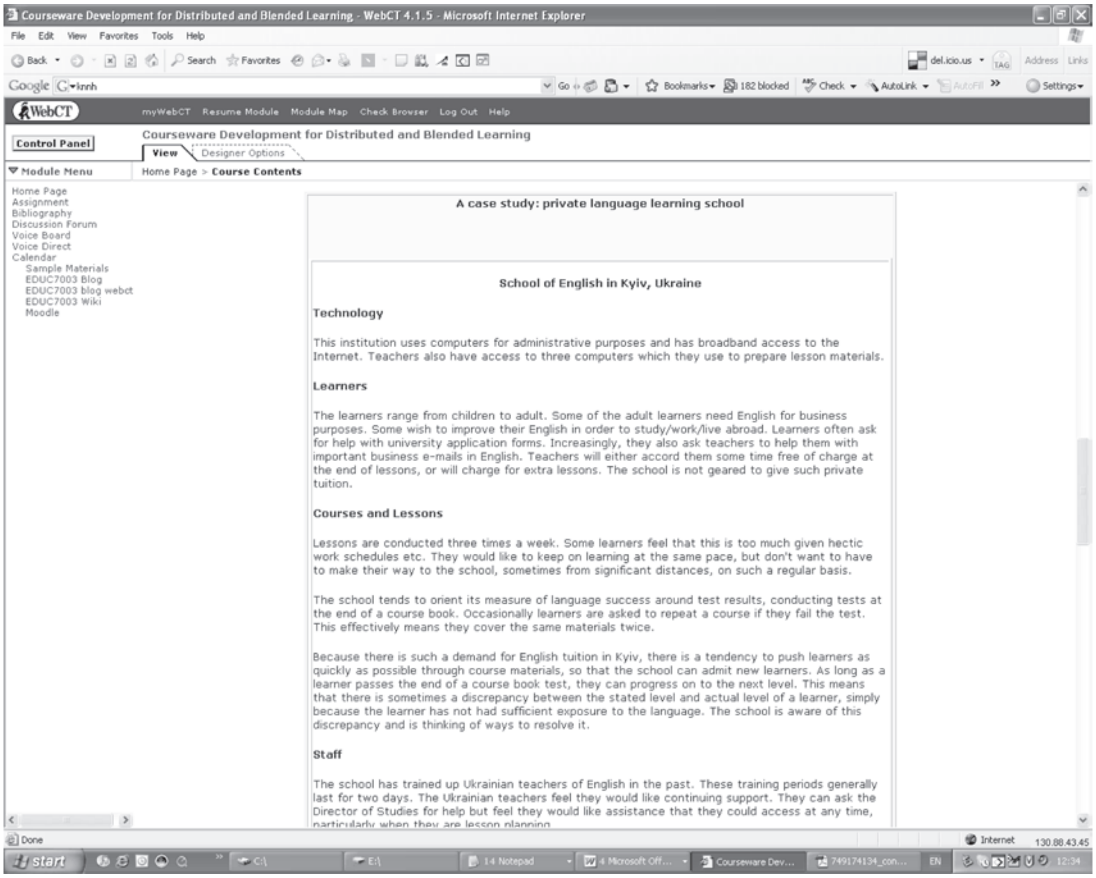
High-tech context	
Full access to and integration of technology with extensive know-how as to when and how to use technology. Technology is normalized.	Learners have extensive access to technology, are used to working with technology and presume that technology will be used in the educational context.
Institution / Teacher / Learner	
Little or no access to computers/digital technologies/broad band. Little or no know-how among staff.	Little or no access to computer/digital technologies.
Low-tech context	

indicative of a context that a participant sees as entirely unconducive to the use of such technology. Generally, the empty table metaphor applies to the perceptions of participants working in low-tech contexts. There have been perhaps three or four participants on CDDBL out of the fifty or so participants since the inception of the module who have perceived Web 2.0 as a smorgasbord. These participants work in mid-tech contexts and have generally excellent IT skills. They are all to greater or lesser extents bottom up introducers of Web 2.0 in institutions that are generally receptive to their ideas.

We also characterize the way participants represent their context in terms of "considerations," "challenges," and "constraints." We find that a participant talks in terms of considerations when they can see ways of using Web 2.0 technologies in their context, but where there are issues they feel they need to take into account, e.g. the level of the language learner in order to effectively do that. The word "challenge" we relate to when a participant sees obstacles to the use of Web 2.0, obstacles that they see as surmountable and where they can envisage themselves playing a role in overcoming them. We use the term constraint to signify times when a participant sees their context as hostile to the use of Web 2.0, and where they feel they have no power to alter that situation (once again such a perception of constraints generally, but not exclusively, relates to participants working in low-tech contexts).

In whichever way we choose to describe participant perceptions of Web 2.0 in relation to context, such perceptions do seem to be the primary factor in how participants relate to and evaluate Web 2.0. In CDDBL therefore, we are increasingly trying to steer our approach so that participants perceive strong connections between what they are doing on the module and what they will be able to do in context. While discussion of the literature relating to the use of such technology can give the participant a general sense of the possible value of Web 2.0, it does not seem to lead to those moments of recognition and connection when a participant "visualizes" themselves using the technology in a way that will beneficially address issues they have in their context or come to a keen understanding of why a particular Web 2.0 tool is not useful. A decision that a Web 2.0 tool cannot offer useful affordances for a specific context should be an informed decision coming from strong critical engagement with, and analysis of, the tool.

Figure 2. Private language school case scenario



A CLOSER LOOK AT THE "CONTEXT APPROACH" ON CDDBL

We are using different strategies to facilitate the moments of recognition and connection which allow participants to visualize a role for Web 2.0 in their teaching. One of these is a "case study approach." One facet of this approach is the use of a semi-authentic case scenario centering on a private language school. In this scenario (illustrated in three descriptions: see Figure 2 for the first of these) a number of issues are flagged. CDDBL participants are asked to think about the role Web 2.0 genres may play in addressing these. The issues relate, for example, to the number of face-to-face sessions the learners at the school are required to attend, which, given their busy lives and the location of the school can prove difficult, and to a pervasive testing system where learners who do not pass the test have to repeat exactly the same course book materials.

Along with the case scenario, we also draw on actual Web 2.0 courseware examples (see Figure 3 for the Web resource page containing these examples) which correspond to specific contexts. Where possible we try to obtain accounts from the courseware developers themselves about the way they have developed the courseware in response to considerations of context. Increasingly we are using courseware examples developed by previous participants on CDDBL and hope to create a bank of exemplars which address various aspects of context. Some previous participants on CDDBL have begun to use courseware conceived on the

Figure 3. Courseware examples

module in their local contexts and we hope that we will be able to tap into their experiences of this in order to see how their evaluation of the affordances of Web 2.0 works out in practice. We hope that this will lead to the formation of a "community of practice" (Wenger, 1998) centering on the use of Web 2.0 for language education.

Another approach we are currently exploring to facilitate "moments of recognition and connection" relates to our "niche" evaluation of Web 2.0. Earlier in this chapter we described Web 2.0 as a malleable medium that teachers can harness to suit the needs of their local contexts and it is this inherent flexibility which is at the root of the niche approach. The emphasis in this approach is not on Web 2.0 tools as killer applications that will change the face of language teaching. Instead, we look at the intrinsic functions of Web 2.0 genres to get a better sense of how they may address specific and perhaps seemingly minor issues within various contexts e.g. a lack of time to develop process writing skills in face-to-face language lessons. Web 2.0 has a broad and expanding set of functions — that expansion of functions stemming from the loop development of Web 2.0 genres in response to the way users employ or wish to employ those genres. These functions resultantly offer a varied range of affordances for language education. We feel that the range of functions of Web 2.0 is well suited to a blended pedagogical approach.

An Exploration of Blogs

It is perhaps best to illustrate this emerging niche approach to Web 2.0 through a discussion of the functions and affordances of one Web 2.0 tool. We have chosen to focus on blogs to do this partly because the educational literature on blogs is currently more substantial than the literature on other Web 2.0 tools. This literature is beginning to analyze the increasingly varied nature of blog types and blog interactions and flesh out pedagogical approaches that relate to these. In doing so it usefully corresponds to our own emerging thinking around Web 2.0 and we can therefore discuss our approach in close relation to recent literature.

A growing number of articles on blogs in education discuss their role in fostering communicative and collaborative interactions (Belderrain, 2006; Cereijol & Myers, 2006; Efimova & Moor, 2005; Owen et al., 2006; Williams & Jacobs, 2004). Owen et al. (2006) define the interactional aspect of Weblogs as those properties that allow "readers to comment on postings, to post links to other blogs and through using pingback and trackback functions (which essentially constitute referencing systems between comments on different blog sites) to keep track of other blogs referencing their posts" (Owen et al., 2006, p. 41). Efimova and Moore discuss the "distributed" generally "spontaneous" conversational interactions which blogs can engender (2005, p. 1), conversations that are tightly associated with the functions of blogs, namely the "comment" feature, "trackback" function and RSS aggregator. Efimova and Moor's research into conversational blogging is particularly useful in its explorations of how specific functions of blogs may relate to the types of interactions that take place through it.

As blogging becomes increasingly popular new tailored blogging environments have been created that respond to and cater for changes in interactional types. Twitter (http://twitter.com/) and Jaiku (http://www.jaiku.com/) are both mobile blogging environments which support brief and frequent "What I'm doing now" type interactions. Such mobile software applications with their ability to provide for embedded/contextualized interactions may potentially offer "virtual and real-world support for social interactions and collaboration in a real-world context" (de Jong et al., 2008, p. 121). The thinking behind such interactive software applications may be juxtaposed with the thinking behind the "slow blogging" movement. The slow blogging movement has its own manifesto rejecting "immediate" blogging and the "disintegration into the one-liners" (Slow blogging manifesto, online) "what I am doing now" type blogging. It expounds, in contrast, an unhurried, reflective "speaking like it matters" approach to blogging which has its roots in the conception of blogging as a diary space.

We can see from such discussion in the literature that blogging may be used in educational settings for a variety of purposes serving to promote interactions and conversations of various types and reflective thinking that is not predicated on interaction. While blogging may be consonant with pedagogical approaches rooted in socio-cultural theory and therefore predicated on social interactions, it might equally support approaches that are not necessarily intrinsically connected with such interaction, approaches, by way of example, rooted in cognitive theory. On CDDBL we discuss how the various uses that blog environments can be put to and the types of user behavior they can engender can relate to the specificities of CDDBL participants' teaching contexts. To this end we are increasingly using terms which reference the specific character and variety of blog spaces such as micro blogs, slow blogs, soap blogs etc.

We also focus on the setting panel in blogs, an aspect of blogs that has been little explored in the educational literature. The settings area of blogs (the areas that provides customization functions) allows the blog owner to disable blog comments and trackback functions which can effectively seal a blog off from interaction. Comments can be approved or rejected by a blog owner before they appear on the blog and the owner can also determine who views their blog and who has a role as an author.

Knowing about these blog properties is important for teachers as they allow for a nuanced methodology in relation to context. Group blogs may be set up by teachers to allow the full range of interactions that blogs can afford, including learner permission to edit the blog. Learners can set up their own blogs and have full control of the permissions on that blog. However a teacher can take a more prominent controlling role of a group blog space, in order to facilitate scaffolded interaction. Such setting functions may prove useful in some contexts.

The participants on CDDBL who work in South Asian and South East Asian locations talk about the "teacher-centered" contexts they work in where the teacher is expected to be the "sage on the stage," rather than the "ghost in the wings," a metaphor for a teacher who plays a hands-off facilitory role (see Mazzolini & Maddison, 2003). Such participants, while they see major potential for a more learner-centered approach — which they view particularly as a beneficial means of developing fluency skills in English — council a carefully scaffolded approach which slowly introduces learners to greater autonomous peer-peer interaction (see the case study on Andrew Prosser). The setting affordances of blog environments can support that transition.

TWO CASE STUDIES

This chapter has discussed the reasons why a context-based approach has been adopted, particularly as it regards CDDBL participant perceptions of their teaching context. It has explored the way the context-based approach aims to help teachers critically evaluate Web 2.0 for their contexts by focusing on the flexibility of Web 2.0 and the range of pedagogical approaches it may be associated with. In what follows, ways in which two CDDBL participants have perceived the potential of Web 2.0 for their context are presented and their approach to harnessing the affordances of Web 2.0 tools to address aspects of those contexts explored. The two participants work in two diverse contexts. One of the contexts may be characterized as verging on high-tech and one as mid-tech. Both participants have called on the affordance of Web 2.0 tools in interesting ways to address issues within their context.

Vida Zorko: University of Ljubljana (Offsite Participant on the First Run of CDDBL, 2005)

Vida teaches English for Specific Purposes (ESP) and develops courseware for groups of sociology students studying at the Faculty of Social Sciences at the University of Ljubljana who receive ESP tuition as a part of their degree. Vida felt that a move from a "traditional" lecture-based approach to Problem Based Learning (PBL) (Savin-Baden, 2000) would better serve the students in their learning. She was instrumental in introducing that approach, an approach for the most part approved of in the Sociology Department. This inevitably required a change in ways of working, both on the part of teachers and students, and impacted on the ESP provision. Vida felt that the introduction of a wiki (in this case a pbwiki), which she considered well suited to the social constructivist learning that underpins PBL, could play an instrumental role in facilitating this change. She combined the use of a wiki, in which students working in small groups solved real-life sociology related problems, with a blog space, which she used to co-ordinate aspects of the blended online and face-to-face learning approach and to offer advice and help when students encountered problems. The students could also access Web 1.0 html pages, which were used for the delivery of language learning resources and activities.

Vida felt that the role of the wiki would prove valuable in:

- Promoting peer-to-peer, teacher-teacher and student-teacher interactions necessary (as Vida sees it) for the successful institution of a problem based teaching approach.

- Increasing motivation by publicly displaying group products.
- Facilitating the sharing of knowledge among students and teachers.
- Empowering students with the authority to construct their own knowledge.
- Enabling teachers to better assess students' progress by monitoring the history of the process.

Vida felt that these potential collaborative affordances link to the following features of the PBwiki:

- An interface which is easy to modify to make it more transparent for users.
- 1-click access to all areas, thus promoting greater sharing of knowledge, making student and tutor contributions easily accessible and allowing tutors to better monitor student progress and to collate reoccurring language problems and deal with them in a face-to-face environment.
- The "whose online" and "edit" function that allow tutors to see who is working in the wiki at a given time, and to respond almost immediately to student contributions. From this perspective Vida characterizes the wiki as "almost a synchronous space."
- A comment area that allows for easy dialogue between student and student, and student and tutor.

However, Vida was aware that such functions, in and of themselves, would not bring about the benefits she felt the use of a wiki would introduce. The student wiki pages all adhere to a certain format (though students can adjust that format themselves) that scaffolds the way the students work with each other, as in the example in Figure 4.

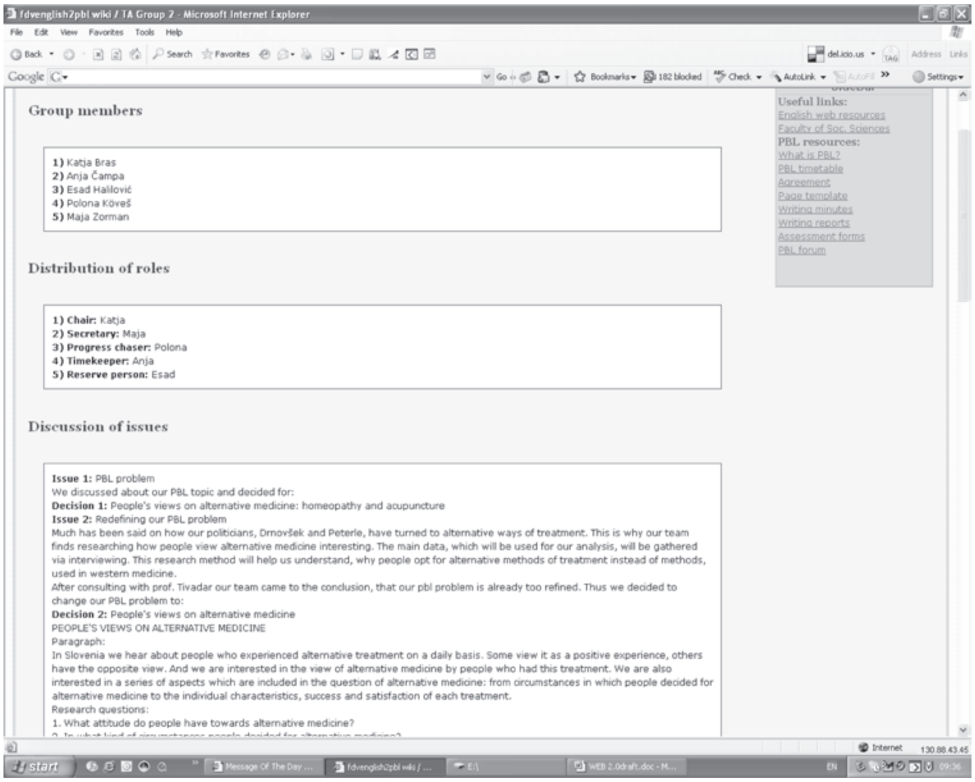
In Vida's case the "smorgasbord" metaphor referred to earlier in this chapter is entirely appropriate as she saw, as a participant on CDDBL, an abundance of opportunities offered by Web 2.0 for her context. She looked on wikis as a tool that with careful scaffolding could facilitate the PBL approach that she had instituted and support a sea change in ways of learning in the Sociology Department. The PBL approach adopted by Vida and the Sociology Department is generally perceived, in the literature, as consonant with the nature of Web 2.0. The Sociology Department was, moreover, amenable to Vida's ideas. In these respects Vida's context arguably offered fertile ground for the introduction of Web 2.0 and may have made it easier for Vida to evaluate Web 2.0 and envisage for it a concrete role. Nevertheless, the introduction of Web 2.0 stemmed largely from Vida's bottom up initiatives and efforts to persuade tutors of the value of the wiki. She saw this process as an enjoyable "challenge," referring back to the three "c" considerations, challenges and constraints framework, rather than as a constraint that would impede the introduction of Web 2.0.

It is perhaps possible that the bottom up influence from Vida, and increasingly her fellow tutors, along with the top down department approval will conspire to normalize Web 2.0 in the faculty, making it the first context we have encountered where this is the case. Early indications through research Vida conducted for her MA dissertation show that the wiki has proved valuable in instigating greater collaborative learning and is fast becoming a "normalized" tool (Bax, 2003a).

Andrew Prosser, Private Language School, Seoul (Offsite Participant on Second Run of CDDBL, 2006)

We might contrast Vida's context with that of Andrew Prosser's. Andrew's is a mid-tech context and is in some ways amenable to Web 2.0 use in that his learners all have access to computers and have some familiarity with Web 2.0 as the majority of them enjoy blogging, a popular Web 2.0 tool in Korea.

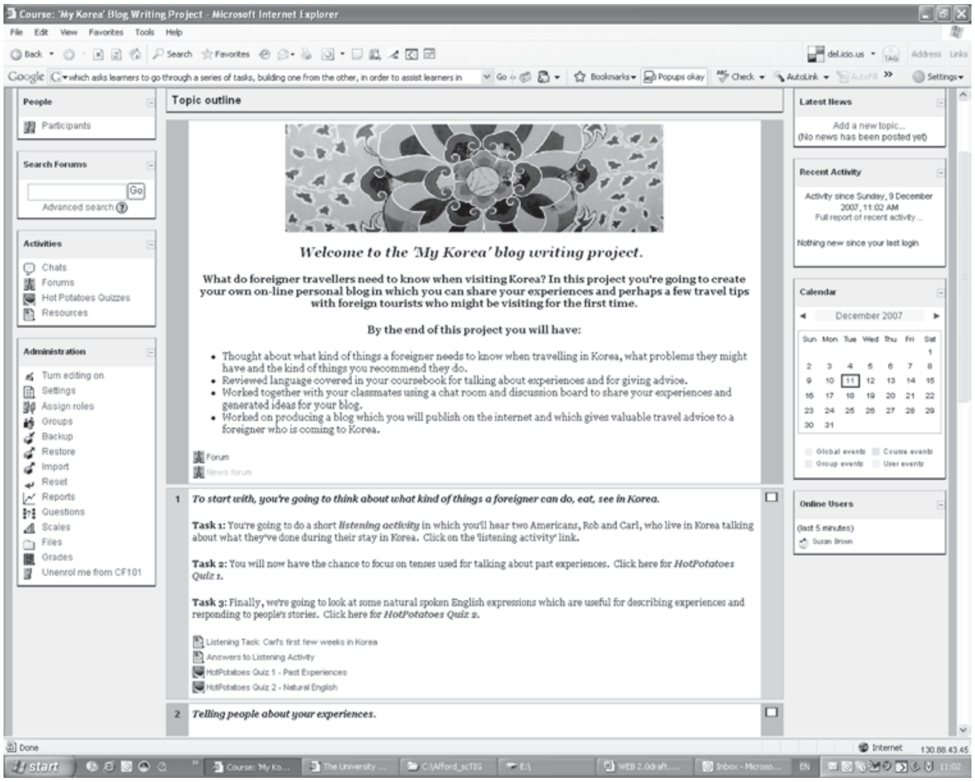
Figure 4. Example page of Vida's wiki environment



It would, however, be an exaggeration to say that Andrew saw Web 2.0 in terms of a smorgasbord of opportunity for his context, particularly where Web 2.0 is associated with highly learner-centered, autonomous learning. Andrew described his context as essentially teacher-centered, where the teacher is viewed as the "sage on the stage" (Mazzolini & Maddison, 2003), "transmitting" knowledge to be memorized by learners. He saw the value of a more learner-centered approach in encouraging greater learner autonomy and learner interactions which would, in turn, have dividends in terms of language development. However, he did not advocate a total shift to a learner-centered approach and argued that nudging learners towards greater autonomy would require a careful structuring of courseware. He had a cautiously optimistic approach to the value of Web 2.0 in such courseware but once again felt that its value would be highly contingent on careful scaffolding.

Andrew harnessed the popularity of blogging in his teaching context by creating courseware with blogs and in Moodle that would lead ultimately to the learners creating their own "tourist guide to Seoul" blogs. He drew, in the initial stages of his courseware, on those affordances of Moodle and blogs that he considered in keeping with an "associative" instructional design approach to courseware design, an instructional design approach which is mainly tutor determined and uses a linear navigational design structure which asks learners to go through a series of tasks in order to assist learners in mastering a specific language structure or skill. To this end he created controlled Web 1.0 practice exercises in Moodle (see Figure 5), and a blog that modeled the way learners may approach their tourist blog. Through this scaffolding process he gradually shifted from a teacher-centered approach, to the more learner-centered

Figure 5. Andrew's Moodle environment



blog task that tapped into the learner-centered affordances of blogs.

Andrew's Web 1.0 and Web 2.0 meld and his perceptions of the potential of Web 2.0 for blending pedagogies in many ways constitute a "niche" approach to his own context.

CONCLUDING REMARKS: THE FUTURE OF CDDBL

In choosing to focus on Vida and Andrew we have illustrated two contexts in which both teachers have envisaged a key role for Web 2.0. and successfully incorporated it into courseware. There is not scope in this chapter to explore case scenarios where CDDBL participants have perceived the introduction of Web 2.0 in their contexts entirely in terms of "empty tables" or "constraints" militating against their incorporation. However, as we have discussed earlier a number of participants on CDDBL perceive their contexts in these terms and we do not anticipate that this situation will change any time soon. This said, if Web 2.0 continues to be integrated into the fabric of societies at its current speed then it is likely in the longer term that institutions and the teachers within them will increasingly need to negotiate their use. The issue then will be less one of "empty tables" and "constraints" and more one of how Web 2.0 can best be used. This may well engender the type of bottom up thinking demonstrated by Vida Zorko and the "niche" thinking of Andrew. Teachers may increasingly also need to negotiate top-down decisions about the use of technology and the extent to which the use of technology should be Web 1.0 based and Web 2.0 based. We hope that CDDBL will help participants see clear ways to play a role in, and negotiate these influences.

In the latest 2007 offsite run of CDDBL a small proportion of the participants were already enthusiastically using Web 2.0 in their contexts before the module commenced and are indeed bottom up initiators of its use. They have a strong grasp of the functions of the technologies, even if they have not greatly explored their pedagogical possibilities. By the end of the module all of the participants are actively contemplating using Web 2.0 technologies in their teaching and learning situations. Assignment presentations that we have viewed include: the use of online video to encourage better presentation skills for trainee teachers in Japan; the introduction of blogs to encourage more accurate writing skills in Mexico; the use of Moodle as a delivery platform to supplement in class activity; the use of Ning as a tool to increase participation in e-learning; the use of RSS feeds to support the development of learner autonomy in Japan; the use of blogs as an e-portfolio in primary schools in Greece; the use of Moodle, wikis and instant messaging to introduce a greater language element into cultural visits in the UK.

There will continue to be participants who view Web 2.0 with skepticism largely because they see their context as militating against its use. However, interest in the module grows and assignment presentations show the inventiveness of the module participants, their increased ability to analyze their contexts, to bring theory to bear and to integrate a variety of Web 2.0 technologies into their teaching. We feel that the context-based, niche approach we are adopting, which we will continue to develop, will help to increase this interest and give our module participants opportunities to use Web 2.0 in ways they feel will enhance their learning context however minor or substantial these modifications may be.

There will continue to be debate and argument about whether Web 2.0 is somehow different and transformative in its very nature. We have argued here that, given its flexibility, which we see as conducive to a bended pedagogical approach, and the possibilities that it offers for breaking down the barriers between the real world and the classroom, it does have a different and possibly transformative potential. We have also argued that the potential for Web 2.0 goes hand-in-hand with the way it is harnessed by tutors/designers to suit local contexts.

It is clear that information/digital technologies are an increasing feature of net migrants' lives and that for the generations coming through the ability to stay socialized via technologies will be a significant part of their identities. Of course, this landscape will continue to change and the elements on the table will constantly refresh, however, we believe that our particular approach will enable both ourselves and the module participants to deal with these changes in an informed and pedagogically appropriate way.

ACKNOWLEDGMENTS

The authors would like to thank Vida Zorko and Andrew Prosser for their kind permission to write about their perceptions of Web 2.0 and their contexts and for their invaluable assistance in ensuring these were faithfully represented.

REFERENCES

Alonso, F., Lopez, G., Manrique, D., & Vines, J. (2005). An instructional model for Web-based E-Learning education with a blended learning process approach. *British Journal of Educational Technology*, *36*(2), 217–235. doi:10.1111/j.1467-8535.2005.00454.x

Bax, S. (2003a). CALL — past, present and future. *System*, *36*(1), 13–28. doi:10.1016/S0346-251X(02)00071-4

Bax, S. (2003b). The end of CLT: A context approach to language teaching. *English Language Teaching Journal*, *57*(3), 278–287.

Beldarrain, Y. (2006). Distance education trends: integrating new technologies to foster student interaction and collaboration. *Distance Education*, *27*(2), 139-153. Retrieved November 24, 2007, from http://www.informaworld.com/smpp/content~content=a749174134

Brown, S. (in preparation). *Context-in-mind: A cultural historical perspective on language teachers' perceptions of social software.*

Bryant, T. (2006). Social software in academia. *EDUCAUSE Quarterly*, *2*, 61-64. Retrieved November 25, 2007, from http://www.educause.edu/ir/library/pdf/eqm0627.pdf

Cereijo, M., & Myers, C. (2006). Weblogs: New communication technology uses in resource limited environments. In Gonzales et al., (Eds), *Current Developments in Technology Assisted Education, 1 (*pp.720-726). Retrieved November 24, 2007, from http://www.formatex.org/micte2006/Downloadable-files/oral/Weblogs.pdf

Chapelle, C. (1998). Multimedia call: Lessons to be learned from research on instructed SLA. *Language Learning & Technology*, *2*(1), 22–34.

Collis, B., & Moonen, J. (2001). *Flexible learning in a digital world.* London: Kogan Page.

De Jong, T., Sprecht, M., & Koper, R. (2008). A Reference model for mobile social software for learning. *Int. J. Cont. Engineering, Education and Lifelong Learning, 18*(1), 118-138. Retrieved November 25, 2007, from http://dspace.learningnetworks.org/bitstream/1820/996/6/08+De+JONG_16.pdf

Efimova, L., & Moor, A. (2005). An exploratory study of conversational blogging practices. *Proceedings of the 38th Hawaii International Conference on Systems Sciences.* Retrieved November 24, 2007, from http://ieeexplore.ieee.org/iel5/9518/30166/01385452.pdf?tp=&isnumber=&arnumber=1385452

Holliday, A. (1994). *Appropriate methodology and social context*. Cambridge: Cambridge University Press.

Kern, R., Ware, P., & Warschauer, M. (2004). Crossing frontiers: New directions in online pedagogy and research. *Annual Review of Applied Linguistics*, *24*(1), 243–260.

Kumaravadivelu, B. (2001). Towards a postmethod pedagogy. *TESOL Quarterly*, *35*(4), 537–559. doi:10.2307/3588427

Mazzolini, M., & Maddison, S. (2003). Sage, guide or ghost? The effect of instructor intervention on student participation in online discussion forums. *Computers & Education*, *40*, 237–253. doi:10.1016/S0360-1315(02)00129-X

Oblinger, D. (2005). *Educating the Net generation*. EDUCAUSE e-book, Boulder, Colorado. Retrieved November 25, 2007, from http://www.educause.edu/LibraryDetailPage/666?ID=PUB7101

Owen, M., Grant, L., Sayers, S., & Facer, K. (2006). *Social software and learning: An Opening Education report from Futurelab*. Retrieved November 30, 2007, from http://www.futurelab.org.uk/research/opening_education/social_software_01.htm

Report, H. (2007). *New Media Consortium*. Retrieved January 29, 2007, from http://www.nmc.org/pdf/2007_Horizon_Report.pdf

Salaberry, R. (2001). The use of technology for second language learning and teaching: A retrospective. *Modern Language Journal*, *85*(1), 39–56. doi:10.1111/0026-7902.00096

Savin-Baden, M. (2000). *Problem-based learning in higher education: untold stories*. SRHE & Open University Press, Buckingham.

Thorne, S. L., & Payne, J. S. (2005). Evolutionary trajectories, Internet-mediated expression, and language education. *CALICO Journal*, *22*(3), 371–397. Retrieved November 30, 2007, from http://language.la.psu.edu/~thorne/thorne_payne_calico2005.pdf

Tudor, I. (2003). Learning to live with complexity: towards an ecological perspective on language teaching. *System*, *31*(1), 1–12. doi:10.1016/S0346-251X(02)00070-2

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. New York: Cambridge University Press.

Wildner, S. (1999). Technology integration into preservice foreign language teacher education programs. *CALICO Journal*, *17*(2), 223–250.

Williams, J. B., & Jacobs, J. (2004). Exploring the use of blogs as learning spaces in the higher education sector. *Australasian Journal of Educational Technology*, *20*(2), 232–247.

Zhong, Y., & Shen, H. (2002). Where is the technology-induced pedagogy? Snap-shots from two multimedia EFL classrooms. *British Journal of Educational Technology*, *33*(1), 39–52. doi:10.1111/1467-8535.00237

KEY TERMS

Context-Based Approach: An approach that encourages teachers to have the confidence to creatively reflect on their teaching practice as it responds to the particularities of their own teaching contexts. Kumaradivelu refers to this as a "teacher generated theory of practice" (2001, p. 541). This means that the potential of technologies cannot be evaluated in abstract terms but as it is interlinked with contextually appropriate practice.

Ecology (Teaching Environment): The teaching ecology refers to all the rich, interacting elements that create the dynamic of a teacher's context. These may include top down societal, curricula and institutional elements and bottom up elements such as learner requests to use more technology in the classroom.

High-Tech Contexts: One where the use of technology is integrated into everyday life, so you would expect there to be easy access to the internet, probably these days through wi-fi; for the bulk of the population to carry mobile phones and for technology to feature strongly in the education system.

Low-Tech Contexts: Whilst the middle classes may have access to mobile phones and access to the internet at home, schools may only have traditional computer rooms which may not well be networked. Access to the internet for the general population is via internet cafes in urban areas rather than through wi-fi.

Pbwiki: One of a burgeoning number of wiki environments. The following page provides a useful comparative analysis of different wiki environments: http://www.wikimatrix.org.

Trackback/Pingback: Links that allow blog users to reference content on each others' blogs. For example, say every learner in a class has their own blog and one learner embeds a video file in their blog about a trip they have been on, if other learners comment on the video in their own blogs and use the trackback function, this will automatically show in the blog of the learner who embedded the video. Note that blogger.com does not currently offer the trackback and pingback function

This work was previously published in Handbook of Research on Web 2.0 and Second Language Learning, edited by M. Thomas, pp. 119-136, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 4.16

An Adaptive and Context-Aware Scenario Model Based on a Web Service Architecture for Pervasive Learning Systems

Cuong Pham-Nguyen
TELECOM, France

Serge Garlatti
TELECOM, France

B.-Y.-Simon Lau
Multimedia University, Malaysia

Benjamin Barbry
University of Sciences and Technologies of Lille, France

Thomas Vantroys
University of Sciences and Technologies of Lille, France

ABSTRACT

Pervasive learning will become increasingly important in technology-enhanced learning (TEL). In this context, development efforts focus on features such as context-awareness, adaptation, services retrieval and orchestration mechanisms. This paper proposes a process to assist the development of such systems, from conception through to execution. This paper focuses mainly on pervasive TEL systems in a learning situation at the workplace. We introduce a context-aware scenario model of corporate learning and working scenarios in e-retail environments such as shops and hypermarkets. This model enables us to integrate contextual information into scenarios and to select how to perform activities according to the current situation. Our pervasive learning system is based on a service oriented architecture that consists of an infrastructure for service

management and execution that is flexible enough to reuse learning components and to deal with context changes that are not known in advance and discovered on the fly.. [Article copies are available for purchase from InfoSci-on-Demand.com]

INTRODUCTION

Nowadays, technology-enhanced learning (TEL) systems must have the capability to reuse learning resources and web services from large repositories, to take into account the context and to allow dynamic adaptation to different learners based on substantial advances in pedagogical theories and knowledge models (Balacheff, 2006). This is particularly true of mobile learning, where context is variable. The reuse of learning resources and web services requires interoperability at a semantic level. In other words, it is necessary to have a semantic web approach to design TEL systems. Moreover, knowledge models and pedagogical theories can be fully represented by means of a semantic web approach. In the mobile learning area, a number of terms are commonly used; mobile, pervasive and ubiquitous learning systems (Brodersen, Christensen, Gronboek, Dindler, & Sundararajah, 2005; Hundebol & Helms, 2006; Sharples, 2005; Thomas, 2007). In computer science, mobile computing is mainly about increasing our capability to physically move computing tools and services with us. The computer becomes an ever-present device that expands our capabilities by reducing the device size and/or by providing access to computing capacity over the network (Lyytinen & Yoo, 2002). In mobile computing, an important limitation is that the computing model does not change while we move. This is because the device cannot obtain information about the context in which the computing takes place and adjust it accordingly. In pervasive computing, the computer has the capability to inquire, detect and explore its environment to obtain information and to dynamically build environment models. This process is reciprocal: the environment also does it and becomes "intelligent". In ubiquitous computing, the main goal is to integrate large-scale mobility with pervasive computing functionalities.

In this article, we consider that mobile, pervasive and ubiquitous learning systems have the properties of mobile, pervasive and ubiquitous computing systems respectively. We focus our attention on pervasive learning systems. Mobile learning is not just about learning at anytime, at any place and in any form using lightweight devices, but learning in context and across contexts. It is best viewed as providing mediating tools in the learning process (Sharples, 2006). Many definitions of pervasive learning are given in the literature (Bomsdorf, 2005; Hundebol & Helms, 2006; Jones & Jo, 2004; Thomas, 2007). One useful definition is that a "pervasive learning environment is a context (or state) for mediating learning in a physical environment enriched with additional site-specific and situation dependent elements – be it plain data, graphics, information -, knowledge -, and learning objects, or, ultimately, audio-visually enhanced virtual layers" (Hundebol & Helms, 2006). One could consider pervasive learning as an extension to mobile learning where the roles of the intelligent environment and of the context are emphasized (Laine & Joy, 2008). In pervasive learning, computers can obtain information about the context of learning from the learning environment where small devices, sensors, pads, badges, large LCD screens, people, and so on, are embedded and communicate mutually. The physical environment is directly related to learning goals and activities. The learning system is dynamically adapted to the learning context. Consequently, a pervasive learning system needs to have an appropriate software architecture to support these features.

In the workplace context, learning can occur in purposeful situations in which there is an explicit goal to learn as well as in incidental situations in which there is no explicit learning goal or interest. Working involves an activity or a related set of activities that require effort and are aimed at achieving one or more objectives. Learning emphasizes what a learner knows or is able to do, while, in contrast, working is related to performance improvement (Michael-Spector & Wang, 2002). In other words, when performing a work task, it often happens that learning also occurs. The performance and quality of work may also be enhanced following learning experiences. Working activities are mainly about solving problems, and in knowledge-intensive organizations this implies continuous learning. Carrying out the particular working task is the priority; learning is just a means (Farmer, Lindstaedt, Droschl, & Luttenberger, 2004). The distinction between learning and working activity is blurring, working being a way of learning, and vice versa. Simon (2007) asserts that traditional methodologies such as formal classroom teaching and even Internet based, content oriented courses and programs have their place at the worksite. Nevertheless, these approaches are generally inflexible to the demands of contextualised, learner centred, performance related challenges (Simon, 2007). Thus, learning processes need to be embedded in organizations, so that learning becomes pervasive and a natural part of work. A particular architecture is required to facilitate the redefinition of learning to mean a work activity and to provide an infrastructure for seamless work-learning integration (Simon, 2007). In such framework, situated learning can be used, where the location, time, environment and tasks, etc. are taken into account. It provides the right learning support at the right time according to the situation parameters and to the goals in the working context. Situated learning increases the quality of learning and is attractive for learning at the workplace and for work-learning integration (Oppermann & Specht, 2006).

In the p-LearNet project (p-LearNet, 2006), a pervasive learning system aims to integrate context-aware corporate learning and working activities within the e-retail framework (retail activities through shops and hypermarkets). In such a framework, we are interested in the following learning issues: the combination of formal learning (formal classroom at the workplace, etc.) and work-learning integration, integration of mobile devices in broader lifelong learning and working scenarios, learning in context, seamless learning across different contexts and context-as-construct (Balacheff, 2006; Sharples, 2006; Vavoula & Sharples, 2008). In such a framework, we focus on a scenario-based approach for TEL system design. Scenarios are used to describe the learning, working and tutoring activities to acquire some domain knowledge and know-how, solve a particular problem or support working activities. Scenario analysis reveals that learning and working situations can be modeled by an explicit hierarchical task model because working and learning activities are well structured and stable. In pervasive learning systems, activities, represented by tasks, can be achieved in different ways according to the current situation. Methods associated with tasks enable us to provide different ways to carrying out those tasks. Activities need to have access to supporting resources or web services. Thus, a context-aware and adaptive mechanism is necessary to select relevant methods associated with a task and their corresponding resources and web services. For a particular couple (Task, Method), resources and web services may also be selected according to the current situation.

In pervasive computing, the computing device has to seamlessly and flexibly obtain information about its environment in which the computing takes place and to adjust itself accordingly. From a software architecture viewpoint, a pervasive learning system has to be flexible enough to reuse learning components (learning resources or learning web services) which are not known in advance and discovered on the fly. A service oriented architecture (SOA) approach facilitates the deployment of an adaptive learning environment based on the aggregation and orchestration of the services needed by an organization. This approach can be effective for pervasive learning systems if one provides for continuous adaptation based on the available services and other contextual information.

The main contribution of this article is an adaptive and context-aware model of scenarios for a pervasive learning system supporting working and learning activities. The pervasive learning system architecture is based on a service oriented architecture to meet pervasive computing requirements. Web services are retrieved and orchestrated, and can be used for different working and learning activities. Thus, the scenario model can invoke web services to undertake activities. The scenario model and the web service retrieval and orchestration are based on a semantic web approach which enables us to represent the explicit common knowledge of the communities of practice involved in the p-LearNet project. The scenario model is based on a hierarchical task model having the task/method paradigm. An activity, represented by a task, may have several associated methods. A method represents a way of performing a task in a particular situation. The context-aware and adaptive mechanism can be viewed as the selection of the relevant content (methods or web services) for a given task according to the current working and/or learning situation. This mechanism is based on matching content description to the current situation for filtering, annotation and ranking. Content and situation need to have corresponding features for adaptation purposes. Methods are described by contextual features while web services are described by metadata. Situations are described according to a context model. For managing web services, we also define the service requirement specification for web service retrieval. Moreover, a pervasive learning system architecture is proposed to facilitate its design and execution in a workplace environment, based on Open Services Gateway initiatives (OSGi) and Universal Plug and Play (UPnP). SOA enables us to design an architecture that is able to inquire, detect and explore its environment to obtain information and to dynamically build environment models. As web services are closely related to learning needs by means of the scenario model, we can provide the right learning support according to the current situation and deal with pervasive computing issues.

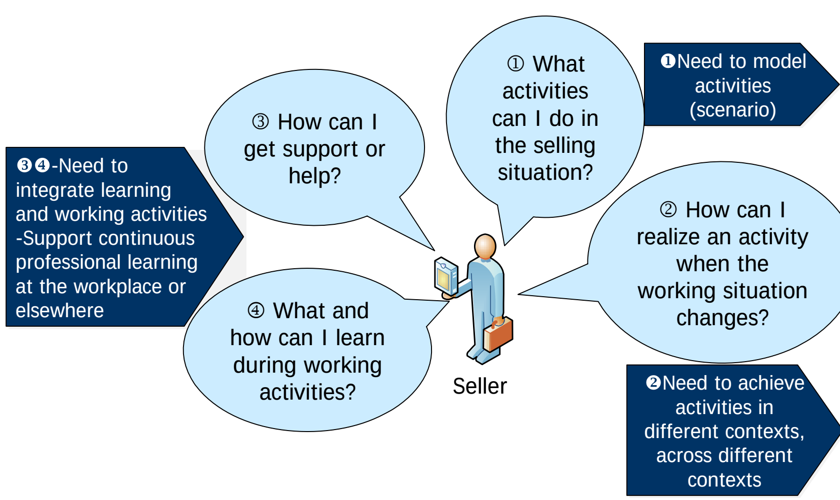
In the next section we begin by presenting the p-LearNet project and e-retail system goals. Secondly, we present our pervasive learning system architecture based on SOA that serves as a platform of service management in e-retail systems. Thirdly, our context management model (organization, features) is detailed and linked to a situation. Fourthly, we present our context-aware scenario model and its relationships with the context management model. Fifthly, the web service requirements and specifications are presented and also the semantic metadata schema describing the web services. Sixthly, the context management process, i.e. the detection and creation of new situations are explained. Seventhly, the adaptation process which links the context model, the scenario model, the metadata schema and the context management is presented. Finally, a conclusion highlights the main results of this study and some perspectives on the results.

P-LEARNET PROJECT AND E-RETAIL SYSTEMS

The 3-year p-LearNet project is an exploratory work on adaptive services and usages for human learning in the context of pervasive communication. One aim is concerned with providing learning to professionals during their work activities. This project therefore addresses concepts and methodologies to facilitate this type of learning. In such a framework, the main issues of the p-LearNet project are: work-integrated learning and customer learning support, continuous professional learning at the workplace, professional learning whatever the place, the time, the organisational and technological contexts of the individual or collective learning and working processes, context-as-construct and seamless learning. The design and engineering of pervasive learning systems must be considered as an interdisciplinary problem requiring the integration of different scientific approaches from computer science, education, commerce, social sciences, etc. Learning focuses mainly on how to support individual and group learning processes through pedagogical guidance and how to enhance the learner's knowledge and know-how. In this project, one of our corporate partners is an international retail companies having chains of shops and hypermarkets wishing to explore sales staff learning at the workplace. Corporate partners identify the problems and requirements about quality and efficiency of information and services to increase market share and the corresponding learning goals. Several innovative scenarios have been set up according to two main learning and/or working situations for both a seller and a customer as learners: i) Seller or customer, outside the shop counters: seller in the back office or storage areas, customer at home or elsewhere; ii) Seller in his department, alone or with a customer having resources from the Smart Spaces (large LCD screen, printers, RFID, etc.) surrounding them (Derycke, Chevrin, & Vantroys, 2007).

In the e-retail framework, the context is continuously evolving during the selling process: a seller can communicate with customers while revising his/her knowledge, checking the inventory or contacting the supplier about products, etc. Sometimes learning and working are interwoven in a pervasive environment. A substantial part of learning does not happen during training but during working activities. Learning and professional activity support must therefore be integrated. A learning system must overcome three main obstacles: time pressure, inadequate learning support in the working context and cognitive and structural disconnection between work, knowledge, and learning (Farmer, Lindstaedt, Droschl, & Luttenberger, 2004). For example, a seller equipped with a portable device, for example a PDA or a UMPC, close to shelves (without a customer) can revise thier knowledge about products and selling techniques or can continue thier previous learning activities to improve their knowledge, or they can verify product labeling or access to product information. During the selling process or the selection / decision phase, the seller can use his/her mobile device as a coach to help them, etc.

Figure 1. Learning requirements and needs at the workplace



In such a framework, the following needs are required (cf. Figure 1).

One of the key problems is the requirement for continuously following the user into a wide variety of contexts, and our efforts address this issue inside the intermediation infrastructure, between the service universe and the communication universe, by adding "intelligence" to provide seamless services regardless of channel. The next section will present the pervasive learning system architecture we developed to deal with our requirements. We then explain in more detail the context and the context-aware scenario model.

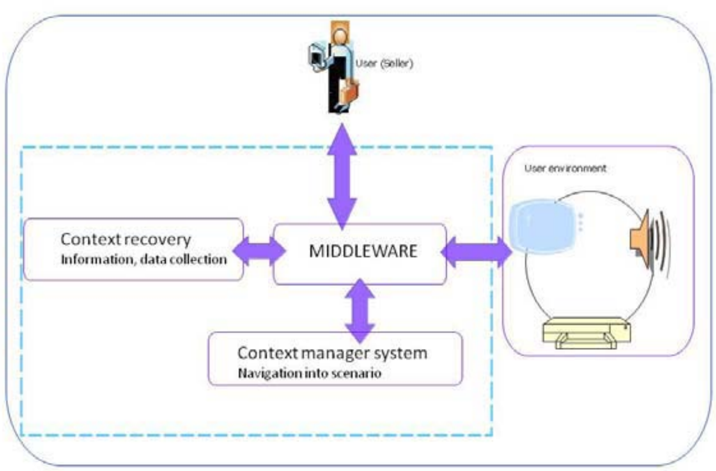
PERVASIVE LEARNING SYSTEM ARCHITECTURE

We developed a generic and flexible software infrastructure (cf. Figure 2) to support learning and working activities at workplace. As described in the introduction, learning in context must support certain functionalities, as listed below.

Firstly, we have to recover information or data from the context, such as activities (current and available), location, and devices available in the environment. We identified two main sources for context recovery. One is the user's environment, which is composed here of a set of devices offering services. These devices can communicate with each other and they can be reachable by anyone, within the constraints of access rules (e.g. sellers can access more services than customers, etc.). Thus, the selling space becomes a part of the process and can collect information (Derycke, Vantroys, Barbry, & Laporte, 2008). Our pervasive learning system can access services provided by devices, external services and semantic composition services. External services may be provided by the company information system. Semantic composition services are mainly composed of four modules (scenario, metadata, context and domain) corresponding to the models mentioned in the next sections. Each module is itself a set of services and managed by the Context Manager System component (cf. Figure 2). These services can be used by the devices and by anyone in the sales department (sellers, customers, etc.). Activities are another source for context.

Secondly, our architecture allows us to aggregate and to coordinate all available services, both from software (learning system, database access, information system, etc.) and from the environment (televisions, webcams, printers, etc.). It is what we call extensibility and modularity. Extensibility means that we can integrate

Figure 2. An overview of pervasive learning system architecture



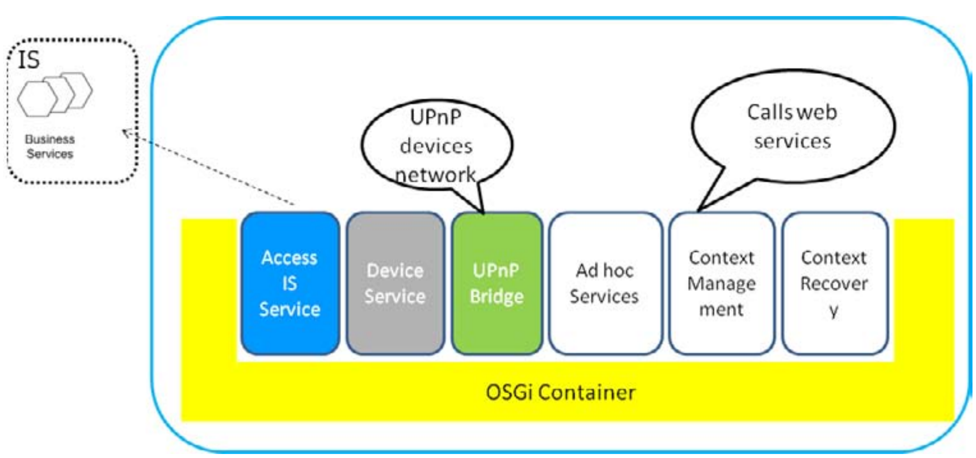
and use any new services (e-learning modules) or devices (printers, televisions, etc.) We have to support live detection of service abilities. For example when a device is turned on or off, the system knows if its services are available or not. Then, we need semantic discovery of all available services. Modularity means addition or removal of some interactive elements, for adaptation to the needs of a particular retail company. The services provided have to be modular too: the system can provide services with different implementations to match the location, time, users, devices, etc. And, finally, our system has to be compliant with existing services. Our system can use services which have already been developed by a particular company. For practical purposes, we do not want to re-develop services already used for e-learning activities.

Thirdly, in order to present all services to the user in a relevant way, we need adapt available services for user applications (see the section "Adaptation" below). Our system uses multi-channel and multimodal intermediations between a mobile personal user device and a collection of e-services (Chevrin, Sockeel, & Derycke, 2006).

To satisfy these requirements we have chosen the following technologies. The modular architecture is possible thanks to OSGi (OSGi, 2007). OSGi is a software layer over Java. It reuses the dynamic class loading capabilities of the Java language. It enables a development model where applications are (dynamically) composed of many different (reusable) components (bundles). It acts as a shell over the Java Virtual Machine: it manages the component life cycle, so we can dynamically add or remove components. Each component hides their implementation from other components while communicating through *services*, which are objects that are specifically shared between components. A similar implementation example is the Eclipse IDE, which is also built on an OSGi framework. Each Eclipse plug-in is an OSGi bundle, which is why we can easily add or remove plug-ins.

To communicate easily with external services, such as our context management system, we used web services access (using a SOAP-based implementation). To discover devices present in the selling space, we use the UPnP (UPnP, 2007) protocol. UPnP is a communication protocol allowing the creation of spontaneous networks of devices (TV sets, HVAC, light control etc.) and control points (PDA, Smartphone, touch panels, etc.). UPnP enables live detection of devices and

Figure 3. Middleware architecture



their use via the SOAP protocol, so the services provided by devices are also seen as web services.

As a consequence, we have to allow communication between different services from different protocols, each with their own formalism. Therefore we have implemented a dedicated communication middleware based on OSGi technology, as shown in Figure 3. Our system is built by assembling services in an OSGi container. Each service can communicate with external components using web services. For example, the context recovery service calls our context manager system as a web service. Moreover, thanks to different OSGi bridges, physical devices can be seen as UPnP devices. The gateway allows us to plug in any kind of device as long as a suitable OSGi driver is available. Each device company can provide such drivers, OSGi and UPnP being open protocols.

CONTEXT MODELING

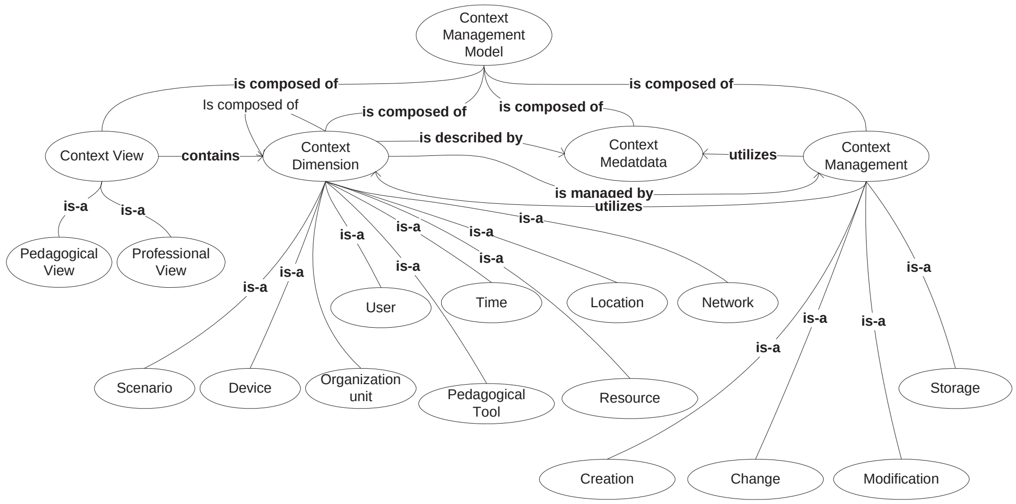
In mobile learning, context may be viewed as "context-as-construct", i.e. "context should be reconceived as a construct that is continually created by the interaction of learners, teachers, physical settings, and social environments" (Sharples, 2006) and "learning not only occurs in the context, it also creates context through continual interaction" (Balacheff, 2006). Dourish (2004) outlines certain modeling characteristics of the context that need to be tailored for structuring contextual information: i) Contextuality is a relational property that holds between objects or activities; ii) The context is not defined in advance. On the contrary, the scope of contextual features is defined dynamically; iii) Context is an occasioned property, relevant to particular settings, particular instances of activities; iv) Context arises from the activities. Context cannot be separated from activities. It is actively produced, maintained and enacted during activities. We claim that activities embedded in a particular physical world (or environment) are key issues to give us intention and meaning according to different situations and finally to determine the relevant features describing the different situations. Many definitions of context are given in the literature, including: *"learning context is used to describe the current situation of a person related to a learning activity; in addition to attributes relying on the physical world model* (Derntl & Hummel, 2005); *"information and content in use to support a specific activity (being individual or collaborative) in a particular physical environment"* (Kurti, Milrad, Alserin, & Gustafsson, 2006).

In our framework, the context model has to deal with learning in context, seamless learning across different contexts and context-as-construct. In other words, we have to manage the context to provide the right learning support. According to these features, a context management model must have the following properties: dynamic and "context-as-construct". Therefore, context settings could be "unpredictable" because it is dynamically created and modified through continual interactions with users and the environment. Context and activities are central to determine learning goals and to choose how achieve them, more particularly in a working environment.

According to these requirements and issues, our context management model is composed of a context model, a context metadata schema, several context views and a set of situations. It is organized as follows (cf. Figure 4).

There is a context model, defining a set of relevant context dimensions and their features. The context dimensions are divided into two categories: abstract dimensions and atomic dimensions. An abstract dimension can be recursively broken down into sub-dimensions which are either abstract or atomic. An atomic dimension consists only of a set of features. Potentially, our context dimensions are the scenario (a hierarchal task model having a task/method paradigm), the user (an employee, a sales person, an expert,

Figure 4. Context Management Model



etc.) with sub-dimensions: the role, previous knowledge, know-how, preferences, loyalty card, purchase intentions, intention of use), the device, the location (office, shelf, stockroom, cash desk, etc.), the time, the learning tools, the network, the physical environment and the resources (e.g. learning object, services, media resource, system resource, etc.). The context metadata schema is described by a set of properties (some of these are shown in Table 1). Dimensions and features are described by context metadata to deal with "unpredictable" context settings. The main idea is to specify in advance for each type of context property how to manage it for creation, change, and adaptation. In Table 1, the four displayed features are used to manage adaptation. Thus, we apply a specific adaptation policy according to the values of these features.

There is also a set of views. A view consists of a subset of context features which are relevant to a given content category (methods, resources, or web services) and a given activity type (learning or working) for adaptation. An adaptation process does not manage the same features for different content categories. In the adaptation process, the current situation, filtered by a view and the corresponding content description are compared. Potentially, a view can also be created for a sub-category of content category or activity type (for example, a sub-category of web services or learning activities). The predicate contextViewFeatures described below in F-logic specifies all context features for a view identified by a content category (*AC*) and an activity type (*AT*). Context views are used to define different viewpoints for adaptation. Consequently, different adaptation categories are specified accordingly (see Box 1).

A set of situations is organized in historical dependencies. A situation is a partial instantiation of the context model consisting of the obtained

Table 1. A subset of context metadata properties

Some context metadata features	Description
State	{Mandatory, Transitory}.
ContentType	{methods, resources, web services}
ActivityType	{learning, professional, etc.}

features describing the current learning or working situation and its physical environment. It defines a complete context state associating all interactions between the user and the learning system at the workplace in a given time interval. A user activity can be influenced by his previous work and learning activities. As soon as a new situation occurs, the next stage can be chosen according to the historical dependencies, for instance, to ensure seamless learning. So far, we do not have enough details from our experiments to define relevant strategies to manage seamless learning across contexts. Future experiments will gives us more detail about this issue and also about the real set of features describing the context management model. At present, it is mainly a kind of generic context management model having an accurate structure and associated principles to manage learning in context and seamless learning across contexts in our pervasive learning system, in other words adaptation. Consequently, our context-aware scenario model is defined according to this context management model.

CONTEXT-AWARE SCENARIO MODEL

In our framework, the goal of scenarios is to describe the learning, working and tutoring activities to acquire some knowledge domain and know-how to solve a particular problem or to support working activities. The main role of a scenario model is to integrate mobile devices into broader learning and working scenarios, to combine formal (formal classroom at the workplace) and work-learning integration and to enable us to manage seamless learning across contexts. As a scenario describes user activities, an author/designer can manage a global activity consistently. In this article, we focus on the work-learning integration context given in Figure 6 (described in detail later).

In pervasive learning systems, activities cannot be achieved in the same way in different situations. It is necessary to have a context-aware and adaptive mechanism to decide how to perform an activity according to a given situation. Derntl and Hummel address these needs by introducing a UML-based modeling extension for including relationships between context and learning activities in the learning design models (Derntl & Hummel, 2005). Several research studies in artificial intelligence focus on the hierarchical task model using the task/method paradigm (Trichet & Tchounikine, 1999; Wielinga, Velde, Schreiber, & Akkermans, 1992; Willamowski, Chevenet, & François, 1994). In a learning environment, hierarchical task models were also used for designing, for instance, authoring tools (Ikeda, Seta, & Mizoguchi, 1997) and learning systems (Betbeder & Tchounikine, 2003; Choquet, Danna, Tchounikine, & Trichet, 1998; Ullrich, 2005). The mechanism of hierarchical and recursive decomposition of a problem into sub-problems is one of the basic characteristics of the hierarchical task model (Trichet & Tchounikine, 1999; Wielinga et al., 1992; Willamowski et al., 1994). In the MODALES project, an adaptive and context-aware model of scenarios has been successfully proposed and implemented. It is based on an interdisciplinary approach (didactics, physics and its epistemology, computer science and education) (Tetchueng, Garlatti, & Laubé, 2008). This hierarchical task model has been reused and modified according to our new requirements.

Box 1.

```
   FORALL AT, CC, ViewFeature  contextViewFeatures(AT, CC, ViewFeature)
     <- EXISTS cm cm:ContextMetadata AND cm[ActivityType->>AT; ContentType-
>>CC; ForContextFeature->>ViewFeature].
```

The Task/Method Concept

Within the framework of the Task/Method paradigm, tasks represent activities and sub-activities managed by a knowledge-based system (cf. Figure 5). A method describes how a particular task can be achieved. There are two types of tasks: abstract tasks and atomic tasks. An atomic task is not composed of sub-tasks. It can be achieved by a simple procedure defined inside a method. An abstract task represents a high level activity composed of sub-activities. A method defines how an abstract task is composed of sub-tasks which can be either abstract or atomic. A method associated with an abstract task defines a control structure that allows the recursive decomposition of tasks into sub-tasks and the sub-task order at runtime by means of operators. At present, three different operators are used: sequence, alternative and parallel. A method associated with an atomic task can have: i) a resource specification for resource retrieval; ii) a service specification for web service retrieval; iii) a procedure/function specification for a simple procedure or function. For a given task, several methods can be used to achieve it. Methods are described by contextual features for selection and adaptation.

Learning Scenario for Sellers at the Workplace

In our framework, all activities (working and learning) are integrated into a single scenario that is modeled by a hierarchical structure of tasks/methods (cf. Figure 6). The selection of the relevant method to achieve a given task according to the current situation, restricted to the relevant context view and the context descriptor, are compared.

Figure 6 shows a part of the e-retail scenario, which represents the decomposition of the task "S.3 – Sale assistance in situation" by a method "M13" in a sequence order of sub-tasks ("S.3_T.1" and "S.3_T.2"). The task "S.3_T.1" can be achieved in two different situations: one without a customer and one with a customer. Thus, the decomposition of the task "S.3_T.1" can be made by two different methods ("M131", "M132"). Each one is associated to a set of contextual features (context descriptor), which specifies the relevant situations for which the method could be selected. Because the task "S.3_T.1" is an abstract task, its methods decompose it into sub-tasks by means of operators determining the sub-task execution order. The method "M131" aims to provide to the seller

Figure 5. Context-aware task model description

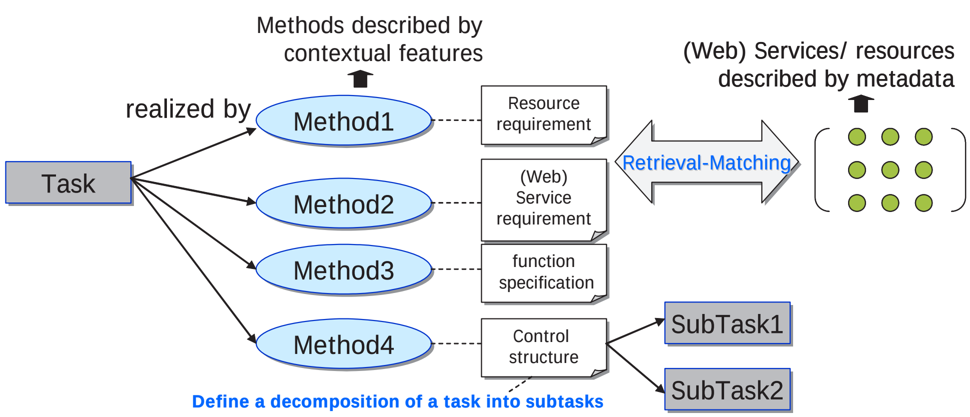
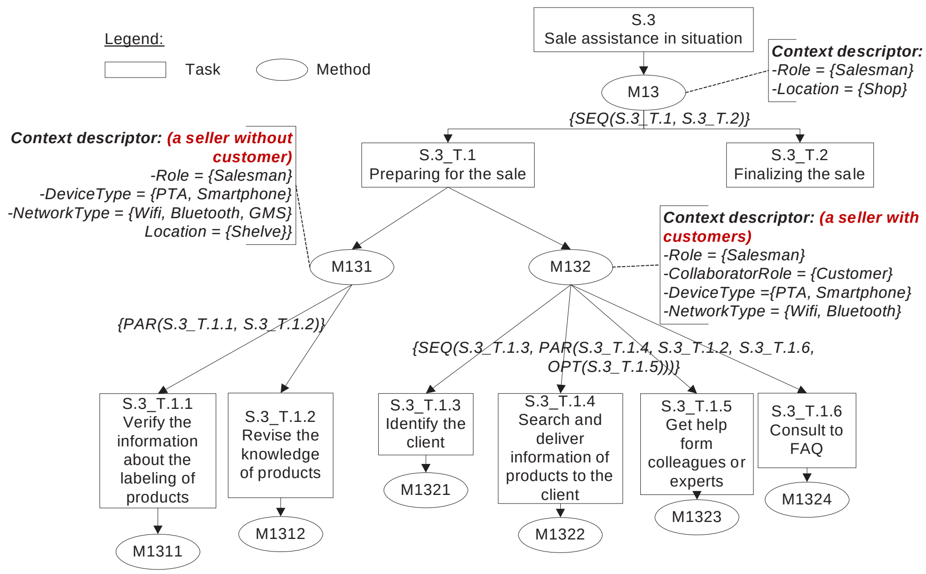


Figure 6. A small part of the e-retail scenario



with both activities available ("S.3_T.1.1 - verify information about the labeling of products" and "S.3_T.1.2 - revise the product knowledge") when s/he approaches shelves without customers (Location = {Shelves}). The method "M132" achieves the task "S.3_T.1" for the seller with a customer (CollaboratorRole = {Customer}). It is divided into four sub-tasks that are carried out as follows: (SEQ (S.3_T.1.3, PAR (S.3_T.1.4, S.3_T.1.5, S.3_T.1.6))) means that we start with the sub-task "S.3_T.1.3" then follow with one of three sub-tasks ("S.3_T.1.4", "S.3_T.1.5" and "S.3_T.1.6") that will be executed in parallel.

We shall use this part of scenario as an example. A task can be achieved by using web services, so in the next section we will show how services are described and annotated for retrieval, orchestration and adaptation in the scope of the task requirements.

SERVICE SPECIFICATIONS AND REQUIREMENTS

In pervasive computing, service oriented architectures are increasingly used to design learning systems. A review of current literature reveals numerous ongoing efforts aimed at exploiting semantic web technologies for web service retrieval and adaptation (Janssen, Lins, Schlegel, Kühner, & Wanner, 2004; Keidl & Kemper, 2004; Pathak, Koul, Caragea, & Honavar, 2005; Sheshagiri, Sadeh, & Gandon, 2004). Semantic modeling of web services is a pre-requisite for successful service retrieval. Service descriptions should specify what the service offers, how the service works and how to access the service (Yang, Lan & Chung, 2005). Work in this area utilises WSDL+DAML-OIL (Sivashanmugam, Verma, Sheth, & Miller, 2003), WSDL-S (Akkiraju et al., 2005) and OWL-S (Martin et al., 2004) among other technologies. Pathak et al. (2005) incorporate OWL-S descriptions to describe service requests and providers that can be processed by a matchmaking engine that is aware of the relevant domain ontologies. The framework supports selecting services based on the user's functional and non-functional requirements, which are then ranked based on user-specified criteria. Dolog et al. (2004) propose a service-based architecture for establishing personalized e-Learning provided by various web-services, including personalization services and support services. Their work encompasses annotation schema and ontologies for learning resources which may include web services, though these are not mentioned explicitly (Dolog, Henze, Nejdl, & Sintek, 2004).

In our learning and working scenarios, an activity may be realized by either a simple internal service or a set of external services. The problem to solve is to search for a service or a number of services that may be composed in a certain order to fulfill the requirements of the current task / method pair. Towards this goal, services have to be semantically annotated. In systems without context adaptation, web services are statically bound to a method at design time. In this regard, the problem is to bind, invoke and execute these web services in a "known" order. Take for instance Task "S.3_T.1.4" in Figure 6. Web service A can be bound to this task (through its method "M132") at design time for searching for product information and resources (images, sound messages, documents, etc.). At runtime, logic is built-in to web service A to search for relevant resources annotated with description metadata from repositories or from the database and deliver them to the client making the request.

The proposed solution is semi-dynamic service searching and matching. The relevant services in this regard can be a single matched service or a set of matched services. The latter is aimed at searching for all the relevant services provided by service providers according to the current situation and the service requirement identified in the selected method. This is done based on an adaptation process applied to services (Please refer to the Context Management, Service Requirement Specification and Adaptation Process sections). Taking Task "S.3_T.1.4" above for example, the realization solution consists of four phases. Each phase is considered as a single service or a set of matched services:

- Phase 1: a search of web services (in a service repository) which provide information about a selected product or a product type required by the Service Requirement Specification of the method "M132" is first carried out according to the current situation.
- Phase 2: All relevant web services of all the suppliers are organized, invoked and executed for searching for product information and resources.
- Phase 3: The found resources are assembled and adapted according to the current situation. They are filtered and sorted according to their degree of relevance.
- Phase 4: The relevant resources are delivered to the target devices. All services are managed and executed by our pervasive learning system architecture previously described.

SERVICE REQUIREMENT WSPECIFICATION

The characteristics and functionalities of a service required for a relevant task or activity have to be specified semantically to facilitate accurate and efficient discovery and matching of the relevant services. The primary goal of a service requirement is the description of how a service is to be "desired". It is a request issued by the system wishing to interact with a service provider in order that a task should be performed on behalf of the learner in the current situation. By our definition, a service satisfies a service requirement by providing a set of desired output parameters for a desired goal with a set of pre-existing input parameters and situational context features. Our proposed service requirement is summarized as follows:

- **Functional requirements:** describe the capabilities of web services desired by a user. It is characterized by input parameters, expected output parameters, pre-conditions and expected post-conditions;
- **Non functional requirements:** include the identity of a service (e.g. name, owner,

type etc.) as well as performance related characteristics, such as Quality of Service (QoS), security etc;

- **Content requirements:** specify a list of domain concepts or a query identifying the content requirements of services (e.g. Mark, Price, ProductModel etc.).

Table 2 shows an example of a service requirement specification "SRS_M1322" for the method "M1322" which realizes the task "S.3_T.1.4 – Search and deliver information on products to clients". This specification consists of three main parts. In the functional requirement part, the input is composed of three parameters (ProductType ?pt, ProductInfoType ?pi and DeliveredResourceType ?rt) describing desired capabilities of the service for retrieving the desired resource (e.g. document, media, voice file, pages, database, etc.). The pre-condition is a predicate that verifies, for example a mandatory presence of a *ProductType*. The output describes a list of relevant resources to be delivered to the client. The non-functional requirements portion represents QoS related context features that can be used for ranking purposes. Finally, the content requirement presents a list of domain concepts that covers a sub-domain (ontology of products) for query refinement. These concepts are used to "compare" with the DomainContent category of a service (see Table 3) through the domain ontology.

Service Description

Service requests and provision are modeled with a common service description metadata model to formally specify the functional and non-functional requirements of services. The fundamental consideration in describing a web service to support accurate and efficient service discovery and matching is to fulfill a three-part ontology (Milanovic & Malek, 2004): *function*, *behavior* and *interface*. *Interface* dictates how the service can be invoked and what resources are to be assembled to provide the desired functionalities of the service. It is syntactic in nature. We propose to follow the WSDL binding standard for message format and protocol details for interoperability. Hence, the problem of searching and matching the right service provision to a client service requirement in a web service architecture is basically reduced to a problem of matching the service

Table 2. An example illustrating the service requirement specification for the method "M132"

Service Request	SRS_M1322	
Functional Requirement	(	
	input	(ProductType ?pt) (ProductInfoType ?pi) (DeliveredResourceType ?rt)
	output	(ProductResource ?pr)
	pre-condition	(isNotNull ?pt)
	post-condition	(isNotNull ?rt)
	)	

function and *behavior* descriptions to the service requirement specification. In this article, we propose a service description ontology to describe web services with a service description feature set. Table 3 enlists features which are generally applicable to most web services.

These features can be extended to include features which describe more specific functionalities of web services such as pedagogy. The features on a feature list can either be mandatory or optional. They serve to index a service for searching, filtering and ranking purposes. As soon as a web service has been chosen according to the service specification and requirements, it is achieved and managed by our pervasive learning system architecture.

CONTEXT MANAGEMENT PROCESS

The context management process integrates all parts and models of the system in a permanent way to deal with complex and dynamic context changes and demands of situated learning at runtime. To this end, it detects the situation changes and then generates a new situation or updates the current situation for maintaining activity relevance, seamlessly. This process is divided into five main stages (see Figure 7):

- Stage 1: Context change detection and aggregation: this stage determines context changes and checks whether these changes lead to the creation of a new situation or an update of the current one. The changes can come from collaborators (colleagues, customers, tutors or learners), user interactions, location, network, device, time, scenario, etc.;
- Stage 2: Gathering of the initial context information to create the new situation, this is done by a partial or complete copy of the last situation and/or by querying the context ontology;
- Stage 3: Selection of the current task: the last task status can be restored in the new situation to ensure seamless learning and working based on the historical dependencies of

Table 3. Service global feature set

Sub categories	Feature set
General	{name, description, language, owner, type (name, taxonomy, value), entityType}
Meta-metadata	{metadataCreator, metadataValidator, creationDate, validationDate, language, format}
Life-cycle	{creator, dateCreated, version, status, contributor, publisher, dateUpdated, extentOfValidity}
Right	{IP, accessRight, signature, provenance, dateCreated, dateUpdated}
Technical	{URI, resource, resourceURI, resourceFormat, replacedBy, realisation, modeOfInteraction}
ServiceRequirement	{listInput(name, type, value, ontologyURI), listOutput(name, type, value, ontologyURI), expectedEffect}
DomainContent	{listDomainConcepts}
Context	{roleModels, location(coordinate, spatialLocation, locationRelativity), physical(deliveryChannel, deliverySystem, deviceModel, tool), informaticResource(hardware, software), temporal(temporalCoverage, frequencyRequirement)}

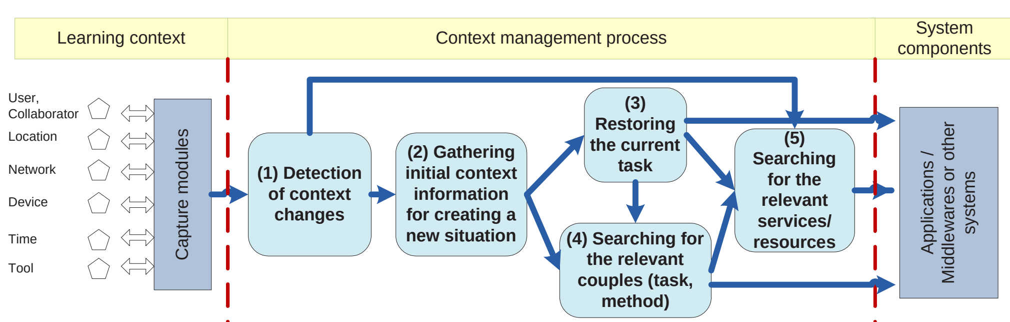
situations. When the last task is finished, the next one is selected according to the current scenario and the historical dependencies of situations. When the last task is not finished, it is necessary to continue it;

- Stage 4: Searching for the relevant couples (Task, Method): According to the current situation and the task, the adaptation process has to select the relevant methods. The main role of the adaptation process is to maintain consistency between the learning system, the physical environment and the current learning or working situation and also to ensure seamless learning and working;
- Stage 5: Discovery and execution of relevant web services to serve the current task: Realization of a task requires the discovery, orchestration, invocation and execution of relevant web services. Service requirement specification which specifies the required service functionalities and characteristics is defined in the selected method of the selected Task/Method pair in Stage 4. The adaptation process is hence aimed at searching for the relevant web services to realize the current task according to this specification and the current situation. The specification and management of web services is described in the next section.

ADAPTATION

The fundamental issue in a pervasive learning environment is how to provide learners with the right activities, the right learning content at the right time and in the right way. Thus, adaptation is mandatory to all types of learning activities in pervasive learning environments (Bomsdorf, 2005). At present, we focus on adaptation mechanisms dedicated to two content categories: method and web service adaptation. At the scenario level, adaptation is aimed at accomplishing an activity according to the current situation, or in other words, how to select the relevant methods for a given task. The learning system has to dynamically select the relevant way to perform the different tasks of a scenario. At the service level, adaptation has to refine the service retrieval process for perform an atomic task. Thus, web services are only available in atomic tasks. The learning system needs to select the relevant web services according to the current couple (task, method) and the current situation. As context is dynamic, it is not possible to know in advance how the next situation will be structured. In other words, it is not possible to anticipate the set of features composing the different situations. For managing adaptation, it is necessary to trigger rules using the contextual features of methods and the service

Figure 7. The context management process



description of web services which are defined a priori and the set of situation features which can be unexpected – not known in advance. Consequently, it is not possible to define rules for each possible configuration of features in a situation.

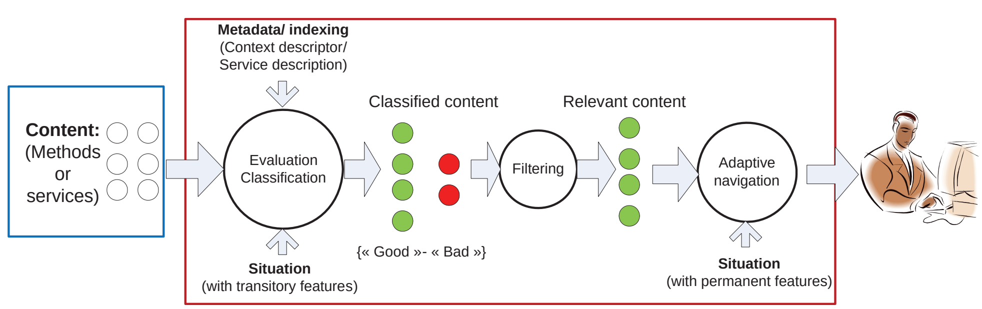
Like the Mobilearn European project (Lonsdale & Beale, 2004), we associate specific metadata with situation properties and/or dimensions for managing adaptation. Situation features are divided into two categories: permanent and transitory. Features describing scenarios and users which are available in all situations are permanent features in a situation. Thus, it is possible to manage them as usual because they are known a priori. Other features are transitory. For them, it is necessary to analyse how each property contributes to content adaptation – methods and services selection - differently according to its role in the adaptation process. Some of them are used to filter the content while others are used to rank or annotate it where filtering, ranking and annotation are the adaptation techniques. For instance, learning and working methods can be filtered according to learning places or used devices while it is annotated according to the user's knowledge or preferences. As soon as features are used to filter content (methods or services), it is easy to manage them. In other words, content will or will not be filtered out by a present transitory feature. At present, all transitory features of content are used to filter. Nevertheless, we shall have to investigate this issue in depth according to more detailed scenarios in future.

Adaptation Process

The adaptation process is specified for a content type (methods and web services) and an adaptation category. The input content of the adaptation process can be achieved in different ways depending on the content type: input methods are specified directly by the current task while input services are selected by a search process based on a query which "compares" the service requirement of the selected couple (task, method) with the service description. The three stages of the adaptation process are presented as follows (cf. Figure 8):

- Evaluation/Classification: input content is classified according to the current situation in several equivalence classes: two classes *{"Good", "Bad"}* for each transitory feature and up to five equivalence classes *{"VeryBad", "Bad", "ToConsider", "Good", "VeryGood"}* for all permanent situation features, together. The content belongs to an equivalent class if it satisfies its membership rules.
- Filtering: all content belonging to "Bad" classes are filtered out. In other words, these content are discarded. For example, with the network dimension, the class *"Good"* is considered as relevant while the class *"Bad"* is not. This means that the system hence will eliminate all content that belongs to the class *"Bad"*.

Figure 8. Adaptation principle



- Adaptive navigation: permanent situation features are used to evaluate and classify the remaining content. An adaptive technique can be chosen by the system or by the user according to an author decision. All content belonging to the same equivalence class are treated in the same way. Annotation and sorting are processed according to the total order of equivalence classes. For hiding, only contents belonging to the class "Good" and "VeryGood" are maintained.

Adaptation of Methods

Method adaptation is specified as follows: 1) All context dimensions for this adaptation category possess the same set of equivalent classes {"Good", "Bad"}; 2) All methods for which the current context and the user features match up to the corresponding method's contextual features (or "belong to" for multiple-valued features) belong to the class "Good" and others belong to the class "Bad"; 3) If the class "Good" is empty, it is considered as a problematic situation and required a designer action to remediate or to provide a new method and context adapted to the user and the task. Otherwise, all methods, belonging to the class "Good", can be provided to the user. The user can choose one of the relevant methods to carry out the current task.

Table 4 shows an example of two membership rules described in a reduced form using F-Logic syntax. The first rule *EvaluateBadMethod* verifies if a method belongs to the class "Bad" according to a situation. It is done based on the mismatching between the method's context descriptor and contextual features of a situation. The mismatching of each contextual feature is verified by each sub rule. The second rule *EvaluateBadLocationForMethod* is a sub rule that verifies the mismatching of the current situation's location feature and the corresponding the supported location list of a method. The predicate *contextViewFeatures* verifies if the location feature is one context view feature for methods. For the filtering, all methods which do not belong to class "Bad" are considered as relevant methods (see the rule *MethodFiltering* in Table 4).

Adaptation of Services

Firstly, a service retrieval process is carried out to search relevant services in the repository based on the Service description to fulfill the Service

Table 4. Illustration of a membership rule of the equivalent class "Bad" for methods

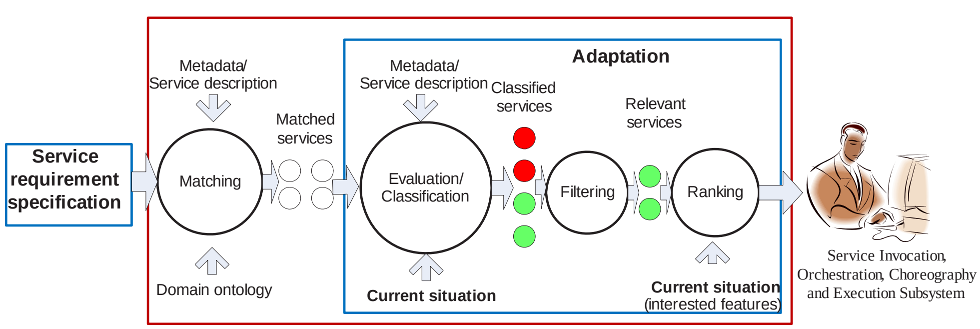
```
RULE EvaluateBadMethod:
FORALL aMethod, aSituation belongingToBad (aMethod, aSituation)
<- EXISTS aCxtDescriptor  (aMethod[has_Context_Descriptor->> aCxtDescriptor] AND (
  IsMismatchedLocation(aSituation->>hasLocationPlace, aCxtDescriptor->>Location) OR
  IsMismatchedNetwork(aSituation->>hasCurrentNetwork, aCxtDescriptor->>hasNextworkTypes) OR
  IsMismatchedDevice(aSituation->>hasUsedDevice, aCxtDescriptor->>hasDeviceType) OR
  IsMismatchedCollaborator(aSituation->>hasCollaborator, aCxtDescriptor->>hasCollaboratorRole) OR
  IsMismatchedRole(aSituation->>hasCurrentUser->>hasRole, aCxtDescriptor->>hasRole) OR
  IsMismatchedKnowHow(aSituation->>hasCurrentUser, aCxtDescriptor->>hasPrerequisiteKnowHow))).

RULE EvaluateBadLocationForMethod:
FORALL aLocation1, aListLocation2  IsMismatchedLocation (aLocation1, aListLocation2)
 <- NOT inList(aListLocation2, aLocation1)  OR NOT contextViewFeatures("Method", "Learning", getCorrespondingCon
textFeature(aLocation1)).

RULE MethodFiltering:
FORALL aMethod, aSituation isRelevant(aMethods, aSituation) <- aMethod:Methods AND NOT
belongingToBad(aMethod, aSituation).
```

request for a task. This is done by querying the Service descriptor metadata repository (matching particularly the ServiceRequirement and DomainContent of Services with the functional, non-functional and content requirements of a request). Secondly, the relevant Services serve as the input for the adaptation phase which is aimed at refining the relevant services according to the current situation and user's interest. Adaptation of services is specified as follows: 1) all transitory context features in the Context category are used to classify input services into two equivalent classes {"Good", "Bad"}; 2) all services belong to class "Bad" will be eliminated; 3) some context features from the Context, Quality, Financial, etc. categories are taken into account for the service ranking process. This process is based on user interest features. For example, users can choose *qualityRating* of services as with high priority while others may be more interested in *performance*. Moreover, ranking requirements also depend on the type of services. With services for voice communications, the performance is very important while with services for financial transactions, security is a high priority. Therefore, the user can build an "interested" feature priority list by annotating each feature with an "interest" level (from very low to very high). This information will be stored in the user's profile for later reuse and be part of the current situation information for this user. Based on the list of "interest" features, the system can rank all matched services.

CONCLUSION

We have proposed an adaptive and context-aware model of a scenario based on a hierarchical task model having the task/method paradigm - with methods defining how to achieve a task - for a pervasive learning system supporting working and learning situations. This model enables us to include contextual information in learning scenarios at the design stage and to choose how to achieve activities according to the current situation at runtime. In other words, the relevant methods are selected dynamically according to the current situation for performing activities. We have also integrated web services described at the semantic level for indexing services. From a scenario perspective, the system can dynamically select and adapt learning components (resources and services) that are not known in advance for achieving an atomic task. Our context model and adaptation process deal with dynamic "context-as-construct" by means of transitory and permanent situation features managed in different ways. A service oriented architecture approach is suitable for pervasive learning systems to deal with such dynamic learning content and environments. At present, because our scenarios are limited, issues of seamless learning and working across contexts are managed in a limited way. The integration of scenarios and learning resources is not actually taken into account. As an area for future work, we will enrich our scenarios and from that, we can study in greater depth the transitory and per-

Figure 9. Service retrieval and adaptation



manent situation features and adaptation policies needed to manage different adaptation categories and seamless learning. Our model will lead to experiments in the scope of the p-LearNet project with our industrial partners to be evaluated in real situations.

ACKNOWLEDGMENT

This work is supported by the p-LearNet project funded by the ANR "Agence Nationale de la Recherche" in France. It receives the support of the PICOM "Pôle de Compétitivité des Industries du Commerce".

REFERENCES

Akkiraju, R., Farrel, J., Miller, J., Nagarajan, M., Marc-Thomas, Schmidt, et al. (2005). Web Service Semantics - WSDL-S. Retrieved 15 November, 2008, from http://www.w3.org/Submission/2005/SUBM-WSDL-S-20051107/

Balacheff, N. (2006). 10 issues to think about the future of research on TEL. *Les Cahiers Leibniz, Kaleidoscope Research Report* (147).

Betbeder, M.-L., & Tchounikine, P. (2003). *Structuring collective activities with tasks and plans*. Paper presented at the IEEE International Conference on Advanced Learning Technologies (ICALT'2003).

Bomsdorf, B. (2005). *Adaptation of Learning Spaces: Supporting Ubiquitous Learning in Higher Distance Education*. Paper presented at the Mobile Computing and Ambient Intelligence: The Challenge of Multimedia, Dagstuhl Seminar Proceedings.

Brodersen, C., Christensen, B. G., Gronboek, K., Dindler, C., & Sundararajah, B. (2005). *eBag: a ubiquitous Web infrastructure for nomadic learning*. Paper presented at the Proceedings of the 14th international conference on World Wide Web.

Chevrin, V., Sockeel, S., & Derycke, A. (2006, June 26-29). *An Intermediation Middleware for supporting Ubiquitous Interaction in Mobile Commerce*. Paper presented at the ICPS'06, IEEE International Conference on Pervasive Services, Lyon.

Choquet, C., Danna, F., Tchounikine, P., & F. Trichet. (1998). *Modelling Knowledge-Based Components of a Learning Environment within the Task/Method Paradigm*. Paper presented at the Intelligent tutoring Systems, ITS'98, San Antonio (USA).

Derntl, M., & Hummel, K. A. (2005). *Modeling Context-Aware E-learning Scenarios*. Paper presented at the Pervasive Computing and Communications Workshops. *3rd IEEE PerCom 2005 Workshops*.

Derycke, A., Chevrin, V., & Vantroys, T. (2007, 13th May). *P-Learning and e-retail: a case and a flexible Sotware Architecture.* Paper presented at the Pervasive Learning 2007. *An International Workshop on Pervasive Learning*, Toronto, Ontario, Canada.

Derycke, A., Vantroys, T., Barbry, B., & Laporte, P. (2008). *E-retail: Interaction of Intelligent Selling Space with Personal Selling Assistant*. Paper presented at the ICEIS 2008, an International Conference on Enterprise Information System, Barcelona, Spain.

Dolog, P., Henze, N., Nejdl, W., & Sintek, M. (2004). *Personalization in distributed e-learning environments*. Paper presented at the Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters.

Dourish, P. (2004). What we talk about when we talk about context. *Personal Ubiquitous Comput., 8*(1), 19-30.

Farmer, J., Lindstaedt, S. N., Droschl, G., & Luttenberger, P. (2004, April 2-3). *AD-HOC - Work-*

integrated Technology-supported Teaching and Learning. Paper presented at the 5th conference on Organisational Knowledge, Learning and Capabilities, Innsbruck, Austria.

Hundebol, J., & Helms, N. H. (2006). Pervasive e-Learning - In Situ Leaning in Changing Contexts.

Ikeda, M., Seta, K., & Mizoguchi, R. (1997). *Task Ontology Makes It Easier To Use Authoring Tools.* Paper presented at the IJCAI.

Janssen, D., Lins, A., Schlegel, T., Kühner, M., & Wanner, G. (2004). *A Framework for Semantic Web Service Retrieval*. Fraunhofer Institute for Industrial Engineering and University of Applied Sciences, Schellingstr.

Jones, V., & Jo, J. H. (2004). *Ubiquitous Learning Environment: an Adaptive Teaching System using Ubiquitous Technology.* Paper presented at the ASCILITE, Perth, Australia.

Keidl, M., & Kemper, A. (2004). *Towards Context-Aware Adaptable Web Services.* Paper presented at the In Proceedings of the 13th International World Wide Web Conference (WWW'04).

Kurti, A., Milrad, M., Alserin, F., & Gustafsson, J. (2006, October 4-6). *Designing and implementing ubuquitous learning activities supported by mobile and positioning technologies.* Paper presented at the the Ninth IASTED International Conference computers and Advanced Technology in Education, Lima, Peru.

Laine, T. H., & Joy, M. (2008). *Survey on Context-aware Pervasive Learning Environments.* Paper presented at the Mlearn 2008, the bridge from text to Context, Wolverhampton.

Lonsdale, P., & Beale, R. (2004). Towards a dynamic process model of context. *Proceedings of Ubicomp 2004 workshop on Advanced Context Modeling, Reasoning and Manage*.

Lyytinen, K., & Yoo, Y. (2002). Issues and challenges in ubiquitous computing. *SPECIAL ISSUE: Issues and challenges in ubiquitous computing, 45*(12).

Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., et al. (2004). *OWL-S: Semantic Markup for Web Services.* Retrieved 15 November, 2008, from http://www.w3.org/Submission/OWL-S

Michael-Spector, J., & Wang, X. (2002). Integrating Technology into Learning and Working: Promising Opportunities and Problematic Issues. *Educational Technology & Society, 5*(1).

Milanovic, N., & Malek, M. (2004). *Current Solutions for Web Service Composition*. Berlin.

Oppermann, R., & Specht, M. (2006). Situated Learning in the Process of Work. In S. Netherlands (Ed.), *Engaged Learning with Emerging Technologies* (pp. 69-89).

OSGi. (2007). *OSGi Alliance*. from http://www.osgi.org

p-LearNet. (2006). *p-LearNet project*. Retrieved from http://p-learnet.univ-lille1.fr/web/guest/home.

Pathak, J., Koul, N., Caragea, D., & Honavar, V. G. (2005). *A framework for semantic web services discovery.* Paper presented at the Workshop On Web Information And Data Management. *Proceedings of the 7th annual ACM international workshop on Web information and data management*, Bremen, Germany

Sharples, M. (2005). Learning As Conversation: Transforming Education in the Mobile Age. In *Proceedings of Conference on Seeing, Understanding, Learning in the Mobile Age* (pp. 147-152), Budapest, Hungary.

Sharples, M. (2006). *Big Issues in Mobile Learning: Report of a workshop by the Kaleidoscope Network of Excellence Mobile Learning Initiative.* LSRI, University of Nottingham.

Sheshagiri, M., Sadeh, N. M., & Gandon, F. (2004). *Using Semantic Web Services for Context-Aware Mobile Applications.* Paper presented at the MobiSys 2004 Workshop on Context Awareness.

Simon, A. (2007). Blurring the lines between learning and working. *Yellow Edge, The Performance Architect.* Retrieved One, from http://www.yellowedge.com.au/Downloads/Issue%201%20-%20Blurring%20the%20lines%20between%20learning%20and%20work.pdf

Thomas, S. (2007, May 13). *Pervasive Scale: A model of pervasive, ubiquitous, and ambient learning.* Paper presented at the An International Workshop on Pervasive Learning, in conjunction with Pervasive 2007, Toronto, Ontario, Canada.

Sivashanmugam, K., Verma, K., Sheth, A. P., & Miller, J. A. (2003). *Adding semantics to web services standards.* Paper presented at the Proceedings of the International Conference on Web Services, ICWS '03, CSREA Press.

Tetchueng, J. L., Garlatti, S., & Laubé, S. (2008). A Context-Aware Learning System based on generic scenarios and the theory in didactic anthropology of knowledge. *International Journal of Computer Science and Applications*, *5*(Special Issue on New Trends on AI Techniques for Educational Technologies), 71-87.

Trichet, F., & Tchounikine, P. (1999). DSTM: a Framework to Operationalize and Refine a Problem-Solving Method modeled in terms of Tasks and Methods. *International Journal of Expert Systems With Applications, Elsevier Science*, *16*, 105-120.

Ullrich, C. (2005). Course generation based on HTN planning. In J. A. & B. B. (Eds.), *Proceedings of 13th annual Workshop of the SIG Adaptivity and User Modeling in Interactive Systems* (pp. 75-79).

UPnP. (2007). *UPnP, Official web site.* from http://www.upnp.org

Vavoula, G. N., & Sharples, M. (2008). *Challenges in Evaluating Mobile Learning.* Paper presented at the mLearn 2008, the international conference on mobile learning, University of Wolverhampton, School of Computing and IT, UK.

Wielinga, B., Velde, W. V. d., Schreiber, G., & Akkermans, H. (1992). The KADS Knowledge Modelling Approach. In R. Mizoguchi & H. Motoda (Eds.), *Proceedings of the 2nd Japanese Knowledge Acquisition for Knowledge-Based Systems Workshop* (pp. 23-42). Hitachi, Advanced Research Laboratory, Hatoyama, Saitama, Japan.

Willamowski, J., Chevenet, F., & François, J. M. (1994). A development shell for cooperative problem-solving environments. *Mathematics and computers in simulation*, *36*(4-6), 361-379.

Yang, S, Lan, B., & Chung, J. (2005, 29 March-1 April 2005). *A new approach for context aware SOA.* Paper presented at the The 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service, 2005. *EEE '05. Proceedings.*

This work was previously published in International Journal of Mobile and Blended Learning, Vol. 1, Issue 3, edited by D. Parsons, pp. 41-69 , copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 4.17
Exploring the Effects of Web-Enabled Self-Regulated Learning and Online Class Frequency on Students' Computing Skills in Blended Learning Courses

Pei-Di Shen
Ming Chuan University, Taiwan

Chia-Wen Tsai
Ming Chuan University, Taiwan

ABSTRACT

Web-based courses have shown to be successful in providing quality distance education. However, due to a national education policy, pure online courses are not permitted in Taiwan. In addition, there exists a lack of appropriate design and delivery of blended learning courses. In this study, the authors conducted a quasi-experiment to examine the effects in applying blended learning (BL) with web-enabled self-regulated learning (SRL) to enhance students' skills of deploying database management system (DBMS). Four class sections with a total of 172 second-year students were taken as four distinct groups. The results showed that students in the SRL and BL groups with 5 online classes had the highest grades for using DBMS among the four groups. Students who received the treatments of web-enabled SRL also outperformed a control group that did not have the benefit of instruction in SRL. The implications of this study are also discussed.

INTRODUCTION

The goals of vocational schools concentrate on developing a highly skilled workforce (Lee & Huang, 1996). Professionals with a vocational degree represent a major portion of the work force in Taiwan (Shen, Lee, Tsai, 2007a). However, vocational education in Taiwan is highly competitive in that it must attract sufficient student enrollments in the face of a continually decreasing birth rate and rapidly increasing number of schools. Schools, facing the high pressure of market competition, often emphasize the proportion of students awarded certificates before they graduate. That is, teaching in this sector usually focuses on helping vocational students to pass the certification examinations (Shen, Lee & Tsai, 2007). The grades on students' certificates and the numbers achieved are the main criteria to evaluate teachers' teaching and students' learning. In this regard, how to help students enhance their professional skills and pass the certificate examinations is a major concern to many teachers in vocational schools in Taiwan.

Web-assisted instruction has been advocated by contemporary educators and researchers (Liu & Tsai, 2008). Asynchronous, web-based educational programs have been shown to be quite successful in providing quality distance education (Overbaugh & Casiello, 2008). However, the policy of e-learning in Taiwan is relatively conservative in contrast with that in the U.S. For example, earning an academic degree entirely through online courses is still not allowed at present. That is, teachers in some nations with conservative institutions and implementations of e-learning, have to adopt a mode of blended learning (BL) rather than pure online learning when implementing e-learning. The effectiveness of BL has already been demonstrated (Liu, Chiang & Huang, 2007; Pereira, Pleguezuelos, Merí, Molina-Ros, Molina-Tomás & Masdeu, 2007; Shen, Lee & Tsai, 2007b), nevertheless, due to limited research on how BL can be conducted effectively using the Internet, it is essential to investigate and develop an appropriate design and arrangement of BL courses for schools and teachers. For example, what frequency of online classes in a BL course is more appropriate to the students, particularly for those with low self-regulatory skills? The authors conducted an experiment to explore the appropriate online class frequency that supports student learning.

Through the Internet, learners are free to access new information without restrictions (Li, Tsai & Tsai, 2008); however, this may also be one of its greatest dangers. There is a continuing debate about the effectiveness of online learning environment designs (Azevedo, 2005; Jacobson, 2005). Online learning differs from didactic presentation, where the student has few opportunities to deviate from the teacher's presentation of the material (Greene & Azevedo, 2007). Moreover, it is indicated that vocational students are more Internet-addicted than students in general (Yang & Tung, 2007). Many vocational students are addicted to shopping websites, online games, and online messengers, and prefer this rather than getting involved in courses, particularly online courses (Shen, Lee, Tsai & Ting, 2008). This addiction to the Internet and the lack of on-the-spot teacher monitoring in web-based instruction makes it even more difficult for students to concentrate on online learning (Shen, Lee & Tsai, 2008). To respond to this challenge, the authors adopt self-regulated learning (SRL) that can help students better regulate and improve their learning.

As more and more institutions of higher education provide online courses, the question arises whether they can be as effective as those offered in the traditional classroom format (Shelley, Swartz & Cole, 2007). However, few studies have discussed effective online instructional methods for vocational students (Shen, Lee, and Tsai, 2007a). Furthermore, we expect that innovative teaching methods and technologies could improve students learning in BL courses. Specifically, this study explores the potential effects of web-based SRL with variations in online class frequency on the development of vocational students' computing skills. Based on suggestions from earlier research, we have re-designed the course and conducted a series of quasi-experiments to examine the effects of web-enabled SRL, varying with frequency of online classes on vocational students' computing skills.

SELF-REGULATED LEARNING

Zimmerman and Schunk (1989) define SRL in terms of self-generated thoughts, feelings, and actions, which are systematically oriented toward attainment of the students' own goals. SRL is an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior in the service of those goals (Winne, 2001; Winne & Hadwin, 1998; Zimmerman & Schunk, 2001). Characteristics attributed to self-regulated persons coincide with those attributed to high-performance, high-capacity students, as opposed to those with low performance, who show a deficit in metacognitive, motivational, and behavioral variables (Montalvo & Torres, 2004; Reyero & Tourón, 2003; Roces & González Torres, 1998; Zimmerman, 1998).

In an SRL environment, students take charge of their own learning by choosing and setting goals, using individual strategies in order to monitor, regulate and control the different aspects influencing the learning process and evaluating their actions. Eventually, they become less dependent on others and on the contextual features in a learning situation (Järvelä, Näykki, Laru & Luokkanen, 2007). Bielaczyc, Pirolli and Brown (1995) incorporated self-explanation and self-regulation strategies in the attainment of the cognitive skill of computer programming. They found that their treatment group, which incorporated self-regulation strategies, outperformed a control group that did not have the benefit of instruction in these strategies. Moreover, Nota, Soresi and Zimmerman (2004) indicate that the cognitive self-regulation strategy of organizing and transforming proves to be a significant predictor of the students' course grades in mathematics and technical subjects in high school, their subsequent average course grades and examinations passed at university.

Previous studies have established that self-regulation skills can help foster learning from any instructional method (see Ertmer, Newby & MacDougall, 1996; Weinstein, 1989; Zimmerman, 1990). In addition, many educators and teachers recognize the importance of SRL in online learning environments. In Lynch and Dembo's (2004) study that investigates the relationship between self-regulation and online learning in a BL context, it is indicated that verbal ability and self-efficacy related significantly to performance, together explaining 12 percent of the variance in course grades. Moreover, it is indicated that successful students in an online course generally used SRL strategies and the effect of self-regulation on students' success was statistically significant (Yukselturk & Bulut, 2007). Therefore, in this study, SRL is applied in this BL course to enhance students' computing skills.

BLENDED LEARNING

Blended learning (BL) is a form of technology-mediated learning that improves learning outcomes through an alternation of face-to-face courses and Internet courses (Lai, Lee, Yeh & Ho, 2005). Marino (2000) discovered that some students experienced difficulty adjusting to the structure of online courses, managing their time in such environments, and maintaining self-motivation. Students may feel frustration in fully online courses, particularly those who are dependent learners, are less self-regulated, and need frequent direction and reinforcement from a visible instructor. These frustrations could be eased when the online course is combined with periodic opportunities for face-to-face interactions (Rovai & Jordan, 2004).

With regard to the effects of BL in previous research, Yushau (2006) shows the positive effect of blended e-learning on students' attitude toward computers and mathematics. It is found that performance, as measured by the final mark of the course under a hybrid teaching method that incorporated both traditional face-to-face lectures and electronic delivery and communication methods, is higher than that of using a traditional teaching method alone (Dowling, Godfrey & Gyles, 2003). In Castelijn and Janssen's (2006) study, their statistical results also indicated that BL students had higher exam scores in a financial management course.

As for the effects of BL on student success rates in learning to program, Boyle, Bradley, Chalk, Jones, and Pickard's (2003) research results indicate a generally positive evaluation of the main elements of the blend, and widespread use of the new online features. Moreover, students in the BL group attained significantly higher average scores than those in the traditional teaching group. Similarly, Pereira et al. (2007) concluded that BL was more effective than traditional teaching. Therefore, BL is applied in this study to help students learn and develop their skills in using application software.

METHODS

Subjects

The subjects in this study were 172 vocational students from two consecutive semesters taking a compulsory course entitled 'Database Management Systems'. Students at this university are expected to spend much more time and effort in mastering a variety of technological skills when compared to those in comprehensive universities in Taiwan. None of students in this study majors in information or computer technology, and the pre-test confirmed that all participants had average or little knowledge of the course content. In addition, there was no student who had taken a web-based course before. We evenly and randomly divided the students into the four experimental groups.

Course Setting

The course is a semester-long, 2 credit-hour class, targeting second-year college students from different major fields of study. The major focus of this course is to develop students' skills in applying the functions of a database management system (DBMS), which is a powerful tool for creating and managing large amounts of data efficiently, robustly and safely, over long periods of time (Матросова, 2007). Students received a study task dealing with the subject of Microsoft Access, one of the popular DBMS applications. Further, according to the school's policy, the course targets helping students to earn a certificate in database applications. That is, students have to take a certification test in Access at the end of the semester.

Experimental Design and Procedure

The experimental design is a factorial pretest - post test design. Four classes were selected from two successive years for this quasi-experiment, with three classes chosen for BL classes and the last one used as a control group. In the first week, the lecturer declared in the three BL classes that the class section would be partially provided with innovative instructional methods mediated on the web as an intervention. Students had the freedom to drop this class section and take another teacher's class section, if preferred. After this declaration, 172 students continued in the three BL classes.

In the first week, students were pre-tested and the results showed that the differences of students' computing skills among the four groups were not statistically significant. That is, students in the four groups had similar levels of computing skills before they received the interventions. Then, participants were randomly assigned to one of the four experimental conditions. The 'SRL and BL with 10 online classes' group (C1, n=44), 'SRL and BL with 5 online classes' group (C2, n=41), 'non-SRL and BL with 5 online classes' group (C3, n=42) are experimental groups, while 'non-SRL and face-to-face' group (C4, n=45) is the control group. C2 and C3 were conducted in the first semester, while C1 and C4 were conducted in the second.

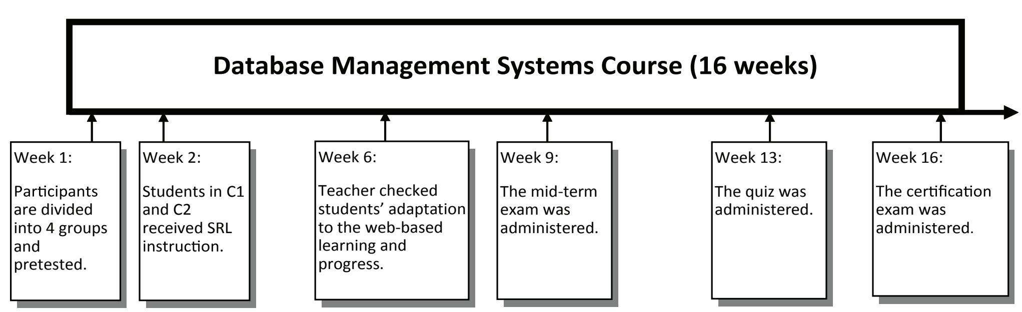
This experiment was implemented in the 'Database Management Systems' course. Students needed to pass the examination to earn the Microsoft Access certificate. The certificate examinations were held immediately after the completion of teaching the course (16th week of the semester). The detailed schedule of the experiment is depicted in Figure 1.

Intervention Concerning SRL

Students in C1 and C2 received instruction in an after-school class teaching SRL strategies. The students were gathered in a classroom and a two-hour lecture was delivered discussing how students could manage study time and regulate their learning. The content of this SRL course was composed of the four processes addressed by Zimmerman, Bonner & Kovach (1996), that is, self-evaluation and monitoring, goal-setting and strategy planning, strategy implementation and monitoring, and monitoring of the outcome of strategy. Students were taught how to implement these four processes to become more self-regulated learners.

In addition to the two-hour lecture, students in SRL groups were required to regularly preview and review the textbook, and practice the skills in using DBMS. The teacher assigned course work to students in the Assignments and Exercises section of the course website and students had to complete the assignments by the required deadline. They were also required to record their learning behavior on the course website every week. The button for submission became unavailable when the time was up. Activities and results of students' learning were captured and recorded on the course website instead of in their notebooks, to prevent falsification of records.

Figure 1. The schedule of the course and skill tests in one semester



Intervention Concerning BL

Networked multimedia technologies and software were applied in the BL classes (C1, C2 and C3). A course website was provided for BL students. The teacher lectured about how to solve simulated computing problems through the Internet or in the classroom. The teacher recorded every lecture session whether in the classroom or via the Internet and later on translated lectures into HTML files with flash, video, and voice. These HTML files were then loaded onto the course website. Students could then preview and review the course sessions on this course website. They could also download the examples and upload their homework to the site.

From the third week, some course work to be delivered online, that is, five or ten non-contiguous weeks in total, was moved onto the website. At the beginning of the semester, the teacher urged students to adjust their learning gradually and smoothly. The remaining weeks' teaching was still conducted in the traditional classroom. In the set of face-to-face classes, the teacher gave lectures and students could ask questions. The mid-term examination, quiz, and certificate examination were all administered in the face-to-face classes.

Evaluation and Qualification

A detailed evaluation of the project was conducted. The authors explored the potential effects of web-based SRL with variations in online class frequency on students' skills in using Microsoft Access. To examine levels of change manipulated by variations in experimental conditions, we first measured students' Access skills as they entered the class. In the first week, students completed two database files in Access as a pretest. The pretest grades showed that the computer skills of almost all the students were low. This confirmed that all participants in the four classes had little knowledge or skill using this software package. Differences in students' skills of using DBMS at the beginning stage among the four classes were not statistically significant. Thus, the researchers ruled out initial differences as a plausible alternative explanation for the differences detected after treatments (Gribbons & Herman, 1997).

The examinations for the certificate in Access were conducted immediately after the course concluded. There were two problems, each consisting of 5 to 7 sub-problems in the examinations. Before testing, students were assigned random seats. All students were tested at the same time. A student's grade came from their correctness and completeness in problem solving. A student could get professional certification using DBMS if his/her grade was higher than 70. Finally, we tested the differences between students' skills in using Access under different conditions.

Design of the Course Website

An open-source Learning Management System (LMS), Moodle, was adopted as the platform for the course website in this study. Teachers who use this LMS can access an array of powerful tools such as assignments, forums, journals, quizzes, surveys, chat rooms, and workshops (Cole, 2005). The course website mainly consisted of five sections: Course Information, Course Content, Course Discussion, Student System, and Assignments and Exercises. Course Information provides course description, syllabus, assignments, grading and course-related information. Course Content includes lectures delivered, conversations that happened in the classroom and the students' exercise files. Students can download the files and listen to audio recordings to review or complete exercises, repeatedly. Teachers may ask questions in the Course Discussion board in order to promote discussion and interaction. Students' personal information and their logs are recorded in the Student System. They could also write their learning journals as blogs in this section. Finally, teachers can assign course work to students in

Assignments and Exercises section and students have to complete and upload the assignments according to deadlines.

RESULTS

A 'one-way ANOVA' was used to compare students' computing skills in using DBMS under different conditions. As shown in Table 1, students from the control group (C4) received the lowest grades among the four groups, and differences in grades among them are significant. However, the difference among the three BL groups (C1, C2 and C3) in students' computing skills is not statistically significant. Further analysis is needed to explore whether the online class frequency influenced students' learning.

The independent samples t-test is used to compare the difference of students' computing skills between 'SRL and BL with 10 online classes' group (C1) and 'SRL and BL with 5 online classes' group (C2). As shown in Table 2, students' average grade for DBMS in C2 (93.83) is significantly higher than that in C1 (84.85) group. This is, a blended course with 5 online classes contributed to better learning effects for vocational students than that with 10 online classes.

Results from Table 3 show that the students' average grade for DBMS in C2 (93.83) is higher than that in the C3 group (86.40). Thus, the effects of web-based SRL on students' skills in using DBMS are positive, and higher than those without SRL intervention.

DISCUSSION AND IMPLICATIONS

Many researchers and educators have highlighted the importance of using technologies to help students learning (Chen, Kinshuk, Wei, Chen &

Table 1. One-way ANOVA: Students' Grades for Using DBMS

Grades	(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
	1	2	-8.981	3.816	.141	-19.76	1.79
		3	-1.556	3.792	.982	-12.27	9.15
		4	19.256(*)	3.727	.000	8.73	29.78
	2	1	8.981	3.816	.141	-1.79	19.76
		3	7.425	3.859	.299	-3.47	18.32
		4	28.237(*)	3.795	.000	17.52	38.95
	3	1	1.556	3.792	.982	-9.15	12.27
		2	-7.425	3.859	.299	-18.32	3.47
		4	20.812(*)	3.771	.000	10.16	31.46
	4	1	-19.256(*)	3.727	.000	-29.78	-8.73
		2	-28.237(*)	3.795	.000	-38.95	-17.52
		3	-20.812(*)	3.771	.000	-31.46	-10.16

** The mean difference is significant at the .05 level*

Table 2. Independent samples t-test: Scores of C1 and C2

Group	n	Mean	S. D.	F	t-value	df	p
C1	44	84.85	17.350	1.274	-2.493	83	.015**
C2	41	93.83	15.749				

*Note. **p < 0.05*

Table 3. Independent samples t-test: Scores of C2 and C3

Group	n	Mean	S. D.	F	t-value	df	p
C2	41	93.83	15.749	.857	2.507	81	.014**
C3	42	86.40	10.834				

*Note. **p < 0.05*

Wang, 2007; Connolly, MacArthur, Stansfield & McLellan, 2007; Liu & Tsai, 2008; Shen, Lee & Tsai, 2007b). There are several advantages in applying a BL approach (Boyle *et al*., 2003). It is also found that many students studying in undergraduate and part-time graduate programs indicated their preferences for retaining some form of face-to-face teaching while at the same time taking advantage of e-learning (Lee and Chan, 2007). Meanwhile, the policy of e-learning in Taiwan is relatively conservative in contrast with that in the U.S. Teachers in this context have to adopt BL if they want to implement e-learning in their courses. Moreover, many vocational students are addicted to shopping websites, online games, and online messengers (Shen, Lee, Tsai & Ting, 2008). In this regard, it is necessary to develop an appropriate design and arrangement for BL courses.

To improve our understanding of this issue, the authors brought in and then tested rigorously a set of hypotheses among four conditions. According to the findings of this study, we believe that our research has made some contributions to e-learning theory in three different ways. Firstly, our research contributed to the existing literature by demonstrating that lower online class frequency in a BL course is more helpful to students learning. Second, this study specifies how teachers help students to regulate their learning by applying web-enabled SRL in a BL course, and further contribute to their learning. Finally, this empirical study demonstrated that a BL course with SRL strategies was more effective in developing students' computing skills than a course in the traditional classroom.

Effects of Online Class Frequency

As the results showed in Table 1, the effects of BL on students' scores for computing skills is significantly higher than those who learned through traditional teaching. Students' computing skills in C1, C2 and C3 were significantly higher than those in C4. The result in this study is similar to Castelijn and Janssen's (2006), Shen, Lee, and Tsai's (2007b), and Yushau's (2006) studies that show the positive effect of blended e-learning on students' learning and attitude toward computers and mathematics.

Moreover, this study attempted to further understand and develop the appropriate design and arrangement of BL courses for schools and teachers. It is shown in Table 2 that a BL course with 5 online classes could result in significantly better learning effects than that with 10 online classes (p = 0.015). It is mentioned that students in the vocational system tend to have lower levels of academic achievement (Lee, 2003), have low confidence and motivation in learning (Su, 2005), have low interest and negative attitude toward their learning (Chen & Tien, 2005), do not adequately get involved in their schoolwork, and do not care so much about their grades (Shen, Lee & Tsai, 2007a). In this specific context, teachers could adopt technologies and teaching websites to help students achieve better learning performance. For example, the BL course and audio-recorded content provide the flexibility and opportunities for students to attend class, review the course content, and practice what they learn at their convenience, particularly before the certificate examinations. However, the online class frequency in a blended course is also one of the critical factors that influences students learning. In the online classes, the physical absence of the instructor and the increased responsibility demanded of learners to effectively engage in learning tasks may present difficulties for learners, particularly those with low self-regulatory skills (Dabbagh & Kitsantas, 2005). Students retreating to the isolation of their computers may avoid school activities and course involvement (Treuer & Belote, 1997). Too many online classes in a BL course may even damage low-achieving students' learning. In this regard, teachers should arrange the appropriate mix of blended classes for students. For example, one traditional class session accompanied the last online class for students to ask questions or for the teacher to check students' progress. With appropriate design, BL could really contribute to students learning.

The Effects of Web-Enabled SRL

With respect to the effects of web-enabled SRL, the results shown in Table 3 indicate that the difference of students' computing skills between C2 and C3 groups is statistically significant (p = 0.014). The success in online courses often depends on students' abilities to successfully direct their own learning efforts (Cennamo, Ross & Rogers, 2002). E-learning should be treated as self-directed learning because the learner attends lectures only to register time, place, subject, and to alter the order of attending lectures (Lee & Lee, 2008). Through the intervention of SRL, the teacher assigned course work to students in the course website and students had to complete and submit the assignments at the required time. The web-enabled SRL helped students regulate their learning behaviors, and further contributed to their learning effects. This result is similar to Chang's (2005), and Yukselturk and Bulut's (2007) studies indicating that self-regulation helped students become more responsible for their learning and contribute to further success.

Based on our findings, we provide suggestions for teachers who teach application software, particularly for those emphasize earning certification. Teachers who wish to stick to traditional methods of teaching, without applying networked multimedia, may no longer be employing a fruitful approach. Students from the control group (C4) received the lowest grades among four groups (see Table 1). In this traditional learning environment, students have neither chance nor channel to review or practice for the tests, and usually ignored the problems of their inadequate skills and knowledge. Therefore, it is suggested that teachers should redesign their courses, design appropriate arrangement of BL courses, and adopt new instructional methods and technologies to fully exploit the benefits of web-based learning environments.

In conclusion, this study explores that the effects of web-based SRL with variations in frequency of online classes on the development of vocational students' computing skills. This study highlights the necessity of applying innovative teaching methods and technologies, and appropriate arrangement of BL courses to help students learn and pass the certificate examinations. Furthermore, this research may provide reference about the intervention of web-enabled SRL and arrangement of online classes in BL courses for schools, scholars and teachers preparing for or presently engaged in implementing e-learning.

REFERENCES

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist*, *40*(4), 199-209.

Bielaczyc, K, Pirolli, P., & Brown, A. (1995). Training in self-explanation and self-regulation strategies: investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction*, *13*(2), 221-252.

Boyle. T., Bradley, C., Chalk, P., Jones, R., & Pickard, P. (2003). Using blended learning to improve student success rates in learning to program. *Learning, Media and Technology*, *28*(2 & 3), 165-178.

Castelijn, P., & Janssen, B. (2006). *Effectiveness of Blended Learning in a Distance Education Setting.* Retrieved July 15, 2008, from http://www.ou.nl/Docs/Faculteiten/MW/MW%20Working%20Papers/gr%2006%2006%20castelijn.pdf

Cennamo, K.S., Ross, J. D., & Rogers, C.S. (2002). Evolution of a web-enhanced course: Incorporating strategies for self-regulation. *Educause Quarterly*, *25*(1), 28-33.

Chen, C.H., & Tien, C.J. (2005). *Market Segmentation Analysis for Taking Skill Test by Students in an Institute of Technology*. Retrieved July 15, 2008, from http://www.voced.edu.au/td/tnc_85.574

Chen, N.S., Kinshuk, Wei, C.W., Chen, Y.R., & Wang, Y.C. (2007). Classroom climate and learning effectiveness comparison for physical and cyber F2F interaction in holistic-blended learning environment. *Proceedings of the 7th IEEE International Conference on Advanced Learning Technologies* (pp. 313-317).

Chang, M.M. (2005). Applying self-regulated learning strategies in a web-based instruction - An investigation of motivation perception. *Computer Assisted Language Learning*, *18*(3), 217-230.

Cole, J. (2005). *Using Moodle: Teaching with the Popular Open Source Course Management System*. O'Reilly Media, CA.

Connolly, T.M., MacArthur, E., Stansfield, M.H., & McLellan, E. (2007). A quasi-experimental study of three online learning courses in computing. *Computers & Education*, *49*(2), 345-359.

Dabbagh, N., & Kitsantas, K. (2005). Using web-based pedagogical tools as scaffolds for self-regulated learning. *Instructional Science*, *33*(5–6), 513-540.

Dowling, C., Godfrey, J.M., & Gyles, N. (2003). Do hybrid flexible delivery teaching methods improve accounting student learning outcomes? *Accounting Education: An International Journal*, *12*(4), 373-391.

Ertmer, P.A., Newby, T.J., & MacDougall, M. (1996). Students' approaches to learning from case-based instruction: The role of reflective self-regulation. *American Educational Research Journal*, *33*(3), 719-752.

Gribbons, B. & Herman, J. (1997). True and quasi-experimental designs. *Practical Assessment, Research & Evaluation*, *5*(14). retrieved

July 15, 2008, from http://PAREonline.net/getvn.asp?v=5&n=14

Greene, J.A., & Azevedo, R. (2007). Adolescents' use of self-regulatory processes and their relation to qualitative mental model shifts while using hypermedia. *Journal of Educational Computing Research*, *36*(2), 125-148.

Jacobson, M.J. (2005). From non-adaptive to adaptive educational hypermedia: Theory, research, and design issues. In S. Chen & G. Magalas (Eds.), *Advances in Web-based education: Personalized learning environments* (pp.302-330). Hershey, PA: Idea Group.

Järvelä, S., Näykki, P., Laru, J., & Luokkanen, T. (2007). Structuring and regulating collaborative learning in higher education with wireless networks and mobile tools. *Educational Technology & Society*, *10*(4), 71-79.

Lai, S.Q., Lee, C.L., Yeh, Y.J., & Ho, C.T. (2005). A study of satisfaction in blended learning for small and medium enterprises. *International Journal of Innovation and Learning*, *2*(3), 319-334.

Lee, P.W.R., & Chan, F.T. (2007). Blended learning: Experiences of adult learners in Hong Kong. In J. Fong & F. L. Wang, (Eds.), *Blended Learning* (pp. 79-87). Retrieved July 15, 2008, from http://www.cs.cityu.edu.hk/~wbl2007/WBL2007_Proceedings_HTML/WBL2007_PP079-087_Lee.pdf

Lee, J.K., & Lee, W.K. (2008). The relationship of e-Learner's self-regulatory efficacy and perception of e-Learning environmental quality. *Computers in Human Behavior*, *24*(1), 32-47.

Lee, L.S., & Huang, J.J. (1996, May 27-29). *Technological and vocational teacher education in Taiwan, R. O. C.* Paper presented at the Taiwan-Australia Conference on Vocational Education and Training, Melbourne, Australia.

Lee, Y.M. (2003). *An Investigation of Taiwanese Graduate Students' Level of Civic Scientific Literacy*. Ph. D. dissertation. Austin, The University of Texas.

Li, K.C., Tsai, Y.T., & Tsai, C.K. (2008). Toward development of distance learning environment in the grid. *International Journal of Distance Education Technologies*, *6*(3), 45-57.

Liu, C.C., & Tsai, C.C. (2008). An analysis of peer interaction patterns as discoursed by on-line small group problem-solving activity. *Computers and Education*, *50*(3), 627-639.

Liu, C.H., Chiang, T.C., & Huang, Y.M. (2007). Assessment of effectiveness of web-based training on demand. *Interactive Learning Environments*, *15*(3), 217-235.

Lynch, R., & Dembo, M. (2004). The relationship between self-regulation and online learning in a blended learning context. *International Review of Research in Open and Distance Learning*, *5*(2). Retrieved July 15, 2008, from http://www.irrodl.org/index.php/irrodl/article/view/189/799

Marino, T.A. (2000). Learning Online: A view from both sides. *The National Teaching & Learning Forum*, *9*(4), 4-6.

Матросова, С.Т.А. (2007). *Computer World*. Retrieved July 15, 2008, from http://venec.ulstu.ru/lib/go.php?id=1566

Montalvo, F., & Torres, R. (2004). Self-regulated learning: Current and future directions. *Electronic Journal of Research in Educational Psychology*, *2*(1), 1-34.

Nota, L., Soresi, S., & Zimmerman, B.J. (2004). Self-regulation and academic achievement and resilience: a longitudinal study. *International Journal of Educational Research*, *41*(3), 198–251.

Overbaugh, R.C., & Casiello, A.R. (2008). Distributed collaborative problem-based graduate-level learning: Students' perspectives on communication tool selection and efficacy. *Computers in Human Behavior*, *24*(2), 497-515.

Pereira, J.A., Pleguezuelos, E., Merí, A., Molina-Ros, A., Molina-Tomás, M.C., & Masdeu, C. (2007). Effectiveness of using blended learning strategies for teaching and learning human anatomy. *Medical Education*, *41*(2), 189-195.

Reyero, M., & Tourón, J. (2003). The development of talent: acceleration as an educational strategy. A Coruña, Netbiblo.

Roces, C., & González Torres, M.C. (1998). Ability to self-regulate learning. In J.A. González Pienda, & J.C. Núñez (Eds.), *Dificultadesde aprendizaje escolar.* Madrid, Pirámide/Psicología.

Rovai, A.P., & Jordan, H.M. (2004). Blended learning and sense of community: A comparative analysis with traditional and fully online graduate courses. *International Review of Research in Open and Distance Learning*, *5*(2). Retrieved July 15, 2008, from http://www.irrodl.org/index.php/irrodl/article/view/192/274

Shelley, D.J., Swartz, L.B., & Cole, M.T. (2007). A comparative analysis of online and traditional undergraduate business law classes. *International Journal of Information and Communication Technology Education*, *3*(1), 10-21.

Shen, P.D., Lee, T.H., & Tsai, C.W. (2007a). Applying Web-Enabled Problem-Based Learning and Self-Regulated Learning to Enhance Computing Skills of Taiwan's Vocational Students: A Quasi-Experimental Study of a Short-Term Module. *Electronic Journal of e-Learning*, *5*(2), 147-156.

Shen, P.D., Lee, T.H., & Tsai, C.W. (2007b, September 15-17). Facilitating Students to Pass Certificate Tests via Blended E-Learning with Self-Regulated Learning: A Quasi-experimental Approach. *WSEAS Proceedings on Multimedia, Internet & Video Technologies.* Beijing, China

Shen, P.D., Lee, T.H., & Tsai, C.W. (2008). Enhancing Skills of Application Software via Web-Enabled Problem-Based Learning and Self-Regulated Learning: An Exploratory Study. *Journal of Distance Education Technologies*, *6*(3), 69-84.

Shen, P.D., Lee, T.H., Tsai, C.W., & Ting, C.J. (2008). Exploring the Effects of Web-Enabled Problem-Based Learning and Self-Regulated Learning on Vocational Students' Involvement in Learning. *European Journal of Open, Distance and E-Learning*, 2008(1). Retrieved July 15, 2008, from http://www.eurodl.org/materials/contrib/2008/Shen_Lee_Tsai_Ting.htm

Su, M.H.M. (2005). A study of EFL technological and vocational college students' language learning strategies and their self-perceived English proficiency. *Electronic Journal of Foreign Language Teaching*, *2*(1), 44-56.

Treuer, P., & Belote, L. (1997). Current and emerging applications of technology to promote student involvement and learning. *New Directions for Student Services*, 78, 17-30.

Weinstein, C. (1989). Teacher education students' preconceptions of teaching. *Journal of Teacher Education*, *40*(2), 53-60.

Winne, P.H. (2001). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman & D. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 153-189). Mahwah, NJ: Erlbaum.

Winne, P.H., & Hadwin, A. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277-304). Hillsdale, NJ: Erlbaum.

Yang, S.C., & Tung, C.J. (2007). Comparison of Internet Addicts and Non-addicts in Taiwanese High School. *Computers in Human Behavior*, *23*(1), 79-96.

Yukselturk, E., & Bulut, S. (2007). Predictors for student success in an online course. *Educational Technology & Society*, *10*(2), 71-83.

Yushau, B. (2006). The effects of blended e-learning on mathematics and computer attitudes in pre-calculus algebra. *The Montana Mathematics Enthusiast*, *3*(2), 176-183.

Zimmerman, B.J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, *25*(1), 3-17.

Zimmerman, B.J. (1998). Developing self-regulation cycles of academic regulation: An analysis of exemplary instructional model. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulated learning: From teaching to self-reflective practice*. New York: Guilford.

Zimmerman, B.J., & Schunk, D.H. (1989). *Self-Regulated Learning and Academic Achievement: Theory, Research, and Practice*. New York: Springer-Verlag.

Zimmerman, B.J., & Schunk, D.H. (2001). *Self-Regulated Learning and Academic Achievement: Theoretical Perspectives*. Mahwah, NJ: Erlbaum.

Zimmerman, B.J., Bonner, S., & Kovach, R. (1996). *Developing self-regulated learners: Beyond achievement to self-efficacy*. Washington, DC: American Psychological Association.

This work was previously published in International Journal of Mobile and Blended Learning, Vol. 1, Issue 3, edited by D. Parsons, pp. 1-16, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Section V
Organizational and Social Implications

This section includes a wide range of research pertaining to the social and organizational impact of Web technologies around the world. Chapters included in this section analyze social marketing, e-government, Web vendors, and Web tourism. The inquiries and methods presented in this section offer insight into the implications of Web technologies at both a personal and organizational level, while also emphasizing potential areas of study within the discipline.

Chapter 5.1
Building Trust in E-Commerce through Web Interface

Muneesh Kumar
University of Delhi South Campus, India, & ESC-PAU, France

Mamta Sareen
University of Delhi, India

ABSTRACT

The emergence of Internet has revolutionalized the way businesses are conducted. The impact of e-commerce is pervasive, both on companies and society as a whole. It has the potential to impact the pace of economic development and in turn influence the process of human development at the global level. However, the growth in e-commerce is being impaired by the issue of trust in the buyer-seller relationship which is arising due to the virtual nature of e-commerce environment. The online trading environment is constrained by a number of factors including web interface that in turn influences user experience. This article identifies various dimensions of web interface that have the potential to influence trust in e-commerce. The empirical evidence presented in the article is based on a survey of the web interfaces of 65 Indian e-Marketplaces. [Article copies are available for purchase from InfoSci-on-Demand.com]

INTRODUCTION

Convergence of Information Technology and telecommunication technology has resulted in emergence of a new economy wherein the buying and selling process is being executed through Internet and other computer networks. This is being termed as e-commerce. The proliferation of Internet technologies into business has fundamentally changed the relationship between suppliers and consumers. It has provided faster access and better knowledge of commodities and prices. The ability to exchange information in both directions between producer and consumer has created a relationship not previously possible. E-commerce is changing the way business is being conducted and eventually all companies will have to make the transition to remain competitive because soon all customers will expect this level of service and it will inevitably become the standard for customer satisfaction (Rust & Kannan, 2003)

The impact of e-commerce is pervasive, both on companies and society as a whole. It is the first mass application of information and communication technologies in the movement towards digital economy. It has broken all man-made boundaries and provided an opportunity for both buyers and sellers to interact among themselves regardless of difference in language, society, culture and tradition.

The rapid growth of e-commerce is now being related to economic development and is often been cited as a driver of economic growth. E-commerce is also been touted as a powerful medium through which less developed economies can exploit the potential of global markets. It, thus, has the potential to impact the pace of economic development and in turn influence the process of human development at the global level. However, the growth in e-commerce is being impaired by the issue of trust which is arising due to virtual nature of e-commerce environment. This virtual nature of e-commerce environment imposes certain constraints on buyer-seller relationship that did not exist in traditional face-to-face transactions. The on-line trading environment basically thrives on 'virtuality' *(Handy, 1995)* and 'user experience' (Marsh, 2000). Since, the transactions in this virtual environment *are conducted through the 'veil' of web interface, trust becomes an important issue. The web interface acts like the only 'contact point' among the buyers and sellers. Hence, there is a need for the web interface to* induce trust in online environment. *The focus of the present article is to identify various trust inducing web dimensions that may enhance the effectiveness of web interface and there by help in inducing trust among the e-commerce players.*

TRUST IN E-COMMERCE

Trust is defined as "the willingness of a party to be vulnerable to the actions of another party based on the expectations that the other party will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer,Davis and Schoorman 1995). In the context of e-commerce, trust may be regarded as a judgment made by the user, based on general experience learned from being a customer/seller and from the perception of a particular merchant. In other words, trust is also seen as a generalized expectancy that the word, promise, or written statement of another party can be relied on (Rotter, 1980).

To date, research on understanding online trust and e-commerce is limited (Grabner-Kräuter and Kaluscha 2003; Yoon 2002; Corritore et al. 2003; Kolsaker and Payne 2002). In their critical reviews of website and/or ecommerce trust, Corritore et al. (2003) and Grabner- Kräuter and Kaluscha (2003) argued that there is a lack a conceptual understanding of online trust and theoretical support for its role in online transactions and relationships. Without trust, businesses are unable to function (Reichheld et al. 2000). Jian, Bisantz, and Drury (2000) and Bailey et al. (2003) claim that trust not only plays a strong role in human-to-human interactions, but also plays a critical role in human-to-computer interactions.

LITERATURE REVIEW

A rich web interface may have a positive impact on trust in the faceless environment of e-commerce. Several studies like Fogg et al 2001; Lee and Kim & Moon, 2000; Neilsen, 1999, 2005 etc. reported evaluations of a list of design features that could potentially appear on the web interface to impact trust. Ang and Lee (2000) stated that if the web site does not lead the buyer to believe that the seller is trustworthy, no business can be conducted. In other words, one key consideration in fostering online trust in e-commerce is to build a trust inducing web interface. Lohse and Spiller (1998) identified four interface design features that affect the effectiveness of the web interface. Their results indicated that features like effective navigation, detailed product descriptions, links, etc. affect the trust in e-commerce activity. Xiling Zhou asserted that poor quality of web interface, lack of proper content in the web sites, unintuitive navigation, etc. can diminish the trust in the concerned company in e-commerce activities. Bailey et al stressed that visual aesthetics and navigation quality of a web site help to assess its trustworthiness in e-commerce activities. Jarvenpa et al, 1999 stated that a web site with trust inducing features functions as a skillful sales person for that company and therefore moderates the disadvantages of an impersonal web site. It is believed that online buyers in e-commerce look for the presence of positive cues about a site's general trustworthiness, as well as for the absence of negative cues. Hence, the e-commerce players by carefully designing their site to set and meet user expectations can influence the trustworthiness of other players.

Arion et al. (1994) asserted that user interface is the point where trust is generated. They stated that trust is a dynamic process, initially based on faith due to the lack of evidence that seeks to reach a certain level of confidence, i.e., where there is conclusive evidence in favor of trusting behavior. In their consideration of computer-supported cooperative work (CSCW) systems, Arion et al. (1994) argued that development of trust in the human-computer interaction need to be supported by the infrastructure/system.

A web interface provides total "user experience". Hence, an effective web interface can make this contact point between the trading partners more meaningful and help in building up trust. The effectiveness of the web interface may also be determined by factors such as the aesthetic appearance of the site, the content and the way the information is presented to the user. The web interface is not just how it looks; it is how easy it is to learn, how well it recedes into the sub consciousness of users, and how well it supports users' tasks. Different authors have suggested various features for making the web interface more effective (Egger, 2003; Neilson, 1999; Wang et al 2007 and others). These features can be classified into three broad categories i.e. ***appeal, content and usability.***

Appeal: It refers to 'attitude' component and the first impression a user gets when accessing a site for the first time Lindgaard (1999) stated that an immediate negative impression may well determine the subsequent perception of the site's quality and usability, whereas one may inherently judge a site by its first impression. Literature from psychology also stresses the important role of a party's first impression, as someone's confirmation bias would entail that all user actions will unconsciously seek to confirm the first impression rather than falsify it (Kahneman & Tversy, 1973; Good, 1988). Fogg et al. (2002) reported, in their large study about how people evaluate the credibility of websites that, almost 50% of all comments made by participants referred to graphic design. They therefore argue that, in the context of online credibility and trust, findings indicate that *looking* good is often interpreted as *being* good, credible and trustworthy. Hence, appeal has largely to do with the site's graphics design and layout. In addition, Demonstrating important clients or providing links towards company's various policies also instills trust among the users (Doney and Canon, 1997).

Usability: Usability is the measure of the quality of a user's experience when interacting with the web site Sweden Canada Link (2001) stated that usability is about making the visit to the website as effective as possible for the users. The focus of usability is on enabling users, whatever their interests and needs, by removing barriers and making the system as easy to use as possible. According to usability expert Jakob Nielsen, usability is a necessary condition for survival on the web. If a web site is difficult to use, people leave. If the homepage fails to clearly state what a company offers and what users can do on the site, people leave. Indeed, ***visual design*** is presented to the user passively, while the user actively needs to ***navigate*** the website in order to access relevant information. Usability is all the more important in the context of online shopping as it is known to be an important condition for the acceptance and adoption of new technologies. The Technology Acceptance Model (TAM), as defined by Davis (1989), holds that usefulness and ease of use are both strong predictors of trust. This model has also been explicitly used to relate trust and e-commerce by researchers like Gefen and Straub (2000), Pavlou (2001), amongst others.

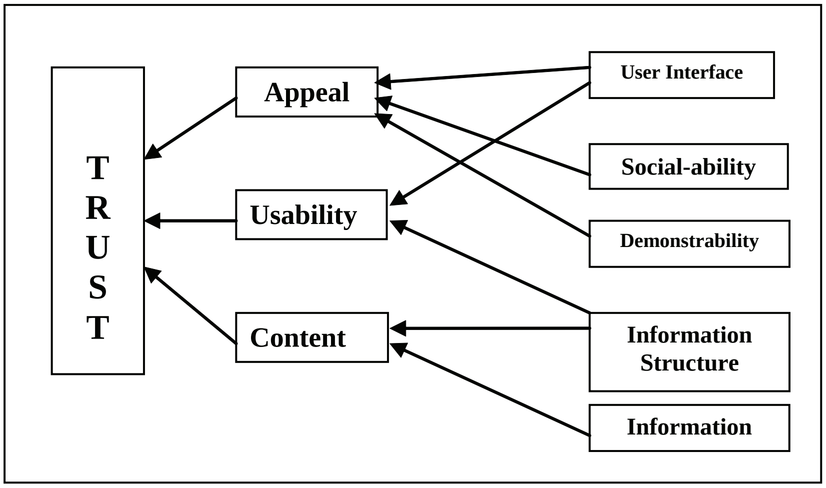
Content: Websites contain information and serve as a medium that predominantly is used for the transfer of information (either technical or not). This plays an important role in the effectiveness of the site. Product information has historically been regarded as a critical element of the content of web interface. A number of researchers have investigated the relationship between web interface and information structure (e.g., Gay et al., 1991: Radha and Murphy, 1992; Mohageg, 1992; Utting and Yankelovich, 1989), concluding that information structure is an essential element of an effective web interface. However, a web site may also contain other information such as detailed and relevant information about the company, its complete offline address, seals of approvals from various trusted third parties, etc. Green (1998) stated that in the e-commerce environment, information plays an important role as business audiences seek more information about products/ services and the company. To be able to convey the information effectively, it is necessary to structure it properly. Correct and detailed descriptions of the products and services offered by the company helps the users to make informed decisions about their transactions. Features that reduce user costs, such as comparisons with competitive products, may also be seen as a sign of honesty and competence (Egger, 2002). In addition, the provision of related content, if relevant, can also be interpreted as the company truly understanding its customers' needs. The credibility of the information has also been observed to be very important about a company's ethical standards.

Various dimensions of the web interface address the appeal, usability and content features of an effective web interface. Various authors have identified a number of such dimensions (Kim 1998; Egger 2003; Wang et al 2006; etc.). Some of these dimensions are categorized as trust inducing dimensions. Kim (1998) had identified four interface design features which contribute to trustworthiness of the web site. Wang et al. (2007) also stressed on certain web dimensions to enhance the richness of web interface. He proposed four broad dimensions of web interface namely, graphic design dimension, Structure design dimension, Content design dimension and Social-cue design dimension that may influence trust. However, the various studies do not include features that are commonly found in any modern e-commerce web site and the e-commerce players are interested in. These include reference to the kind of security policy, privacy policy being followed by the company for online transactions, the technology related policies and procedures followed by the company to address the security, privacy etc. issues, the statements from well known customers, media excerpts, etc. Thus, there is a need for a more comprehensive model relating trust and web interface.

WEB INTERFACE AND TRUST MODEL

In order to incorporate the various gaps in the earlier model relating trust with web interface, the present article proposes five trust inducing dimensions of an effective web interface. These dimensions are: a) User Interface Dimension; b)

Figure 1. Web interface and trust model



Information Structure Dimension; c) Information Content Dimension; d) Demonstrability Dimension; and e) Social-ability Dimension. The relationships between these dimensions and three features identified earlier are exhibited in the Figure 1.

- **User Interface Dimension:** User Interface indicates the appearance and the tools available for accessing the information contained in the web site/portal. Constaine (1995) pointed out that interface is important aspect as the more intuitive the user interface is, the easier it is to use and trust. The website must be recognizable as from the organization. That is, it must be obvious that the look of the site – colors, logos, layout, etc. is consistent with other collateral from the organization. The various features offered by the web site that normally giving the first impression about the company lays the initial foundation of trust building process. Kim and Moon, (1999) reported that the overall color layout and graphical interface influences the trustworthiness of the web site. This dimension aids in addressing the appeal aspect of an effective web interface. Various features that may be useful in enhancing trust levels among the users include: a) Home Page; b) Graphics Interface; c) Links; d) Professionalism; and e) Loading Time.
- **Information Structure Dimension:** The structure dimension defines the overall organization and accessibility of displayed information on the web site. Ease of navigation has frequently been mentioned as a key to promote online trust (e.g., Cheskin/Sapient Report, 1999; Neilsen, 1998). In other words, users must be able to easily locate the information they seek on the web site. This ease-of- use reflects two characteristics of a trustworthy web site: simplicity and consistency. Buyers appreciate simplicity or a clear design of e-commerce web sites because it reduces the perceived risks of deception, frustration, and wasting time. When the structure and design of the web site are consistent, users feel more confident using the site because they can transfer their learning from one sub-site to the next rather than having to learn everything over again for each new page and trust is build (Neilsen, 1998,). For example, broken links, meaningless images, and similar ‘‘hygiene factors’’ may relate to users dissatisfaction with a web site (Zhang et al., 1999). Key

features of structure dimension of a web site may include: a) Navigation; b) Accessibility; c) Functionality; d) Consistency; and e) Learnibility.

- **Information Content Dimension:** This dimension refers to the informational components that can be included on the web site, either textual or graphical. A logically structured web site providing comprehensive, correct, and current product information instills trust among the users (Egger, 2001; Neilsen, 1999). If the information regarding the products is precise, factual and contains links of details required, if necessary, then trust is build (Bhattacharya, 2001). Recent market surveys include that some of the companies are using their web sites as a part of integrated communication strategy to create trust and action (Sheehan and Doherty, 2001). In such cases the website interface plays an important role in e-business transactions. The contents should be displayed as being less complex and more users friendly that enriches the visitor's experience and motivates him/her to visit the site again (G.Chakraborty et al 2003). key features of structure dimension of a web site may include: a) Navigation; b) Accessibility; c) Functionality; d) Consistency; and e) Learn-ability.
- **Demonstrability Dimension:** Several researchers stress the importance of ''demonstrability'' in e-commerce, which is to promote the brand reputation of a company online. It is often seen that various features like offline address of the company, details about real people behind etc play the same role of offering certain clues about the credibility of the company as the physical clues in the brick and mortar business transactions. These clues help the potential trading partners to assess the credibility of the company. Especially displaying the seals of approval from various trusted third parties and the accreditations earned help in building trust levels.
- **Social-ability Dimension:** This dimension relates to embedding social cues, such as face-to-face interaction and social presence, into web site interfaces via different communication media, because a lack of the ''human touch'' or presence may constitute a barrier for at least some consumers to trust online merchants (e.g., Riegelsberger & Sasse, 2002). The effectiveness of a personalization system improves in the long run. Weiner and A. Mehrabian (1968) stated that the choice of language can help create a sense of psychological closeness and warmth. Where as Nass and Steuer (1993) stressed that the use of natural and informal language can impact perceived social presence. Yoon (2002) also showed that web site trust is influenced by consumer familiarity. Every time a customer interacts with the web site, the personalization mechanism collects new data about the user's preferences, so that a more and more satisfactory service can be offered.

METHODOLOGY AND RESULTS

A list of 100 B2B web e-marketplaces operating in India was prepared through the use of various search engines like Google, Yahoo, MSN etc. The criterion for the selection of these e-marketplaces was random. On visiting these web sites, it was found that some of the web portals were merely a directory of sellers and buyers and not actually carrying out e-commerce transactions. Such web portals were excluded from the sample. Hence, a total of 65 B2B web sites/e-marketplaces were selected for the purpose of the survey. Based on the number of elements present, the web site was to be rated on a 5 point Likert scale for each of the trust inducing dimension. Finally, an overall trust rating of the web portal was also to be obtained for each web interface of the select e-marketplace.

Initially, 25 participants were requested to rate six e-marketplaces. The participants were IT savvy in the age range of 30-45 years and came from a variety of background like public sector, business sector, private sector, banking sector and academic sector. The participants were asked to provide their own ratings of each dimension and also the overall rating of the web site based on a questionnaire consisting of 20 questions. Hence, a total of 120 responses were received from each of the 25 participants. Cronbach's Alpha Scale Reliability test was used to test the reliability of the questionnaire and it was found to be 0.8751, which is fairly good degree of reliability.

In order to study the significance of these trust inducing dimensions for the level of trust, linear regression model was used (the limitations of linear regression model in this context may be recognized). The model used the level of trust as dependent variable and each of the trust inducing dimensions as independent variable. The purpose was to find out any redundant dimension included in the dimension. The regression analysis was carried out on the data collected from these 25 participants for each of the six e-marketplaces. The adjusted R^2 ranged from 86% to 92%. This would imply that more than 85% of the variation in the trust ratings could be collectively explained by the five trust inducing dimensions of an effective web interface. The results of regression analysis for the ratings given for one of the e-marketplaces are presented in Table 1.

During the systematic elimination process, all the five trust inducing dimensions were found to be significant. The results showed Demonstrability and Information Content dimension as significant trust inducing dimensions of web interface. This further strengthens the reasons of the present article for the inclusion of demonstrability dimension as a trust inducing dimension for web interface to the model proposed by Wang *et al*. The information structure dimension and the User Interface dimension of web interface also were found to be having significant contribution towards building trust in web interface. This could be because of the fact that in India, complete virtual B2B transactions do not occur very often and they are aided with various offline channels like phone, fax, etc. However, the result assures that all these dimensions of an effective web interface as proposed in the model have a potential to enhance trust among the users. Further, socio-ability dimension, though considered important in B2C e-commerce activities, was not found to be contributing much towards trust levels in e-marketplaces.

In order to identify any bias in the evaluation of the web interfaces of the six e-marketplaces, the author also independently evaluated all the select e-marketplaces. The results of the evaluations were compared with the trust equation obtained of these select e-marketplaces. It was observed that the difference between the two results was not significant. This would imply that the evaluations of the e-marketplaces done by 25 participants and the author held nominal bias. Therefore, evaluation of the remaining 59 e-marketplaces was carried out independently by the author, in the same manner as it was done by the 25 participants.

The regression model used earlier was used on the data so collected from the evaluations of web interfaces of the selected e-marketplaces. The results of the linear regression analysis are presented in the Table 2. As may be observed from Table 2, all the five trust inducing web dimensions were found to be contributing towards trust in the web site.

All the dimensions of effective web interface as identified in the model were found to be good predictors of level of trust. The results were fairly comparable with the results obtained from the linear regression analysis carried out on the data collected through 25 participants with respect to six select e-marketplaces. Thus, it may be concluded that the five trust inducing web dimensions namely Information Content dimension, Demonstrability dimension, Information

Table 1. Trust ratings of an e-marketplace and trust inducing dimensions. Partial results of regression

Model	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. error	Beta	t	Sig
(Constant)	-0.312	0.460	-	-0.677	0.506
User Interface	0.243	0.115	0.242	2.912	0.003
Info. Structure	0.225	0.116	0.265	2.093	0.009
Info. Content	0.327	0.217	0.327	3.209	0.001
Demonstrability	0.338	0.165	0.264	3.049	0.002
Socio-ability	0. 033	0.097	0.003	1.634	0.020

Table 2. Trust ratings of an e-marketplace and trust inducing dimensions. Partial results of regression

Model	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig
(Constant)	-.633	0.232	-	-2.732	0.008
User Interface	0.178	0.114	0.242	2.816	0.008
Info. Structure	0.269	0.102	0.211	2.334	0.011
Info. Content	0.421	0.094	0.382	3.486	0.001
Demonstrability	0.296	0.105	0.261	2.836	0.006
Socio-ability	0.067	0.114	0.060	0.590	0.055

Structure dimension, User Interface dimension and the Social-ability dimension (in that order) positively contribute towards trust building processes in e-commerce.

LIMITATIONS AND SCOPE FOR FUTURE RESEARCH

The main limitation of this article could be the coverage of only e-marketplaces in validating trust in e-commerce. Since, e-commerce involves various other activities, it would be better, if the sample data would contain web interfaces from other segments of e-commerce also. Further, the model proposed in this article has tried to include as many trust inducing dimensions as possible. It may, however, be possible that the article might have overlooked certain issues that might influence the web interface of the e-commerce sites. However, user experience is a very complex issue related to human-computer interface and may vary from individual to individual. Ideally, different individuals must have rated all the e-marketplaces. However, it was not possible for the present scope.

CONCLUSION

E-commerce has the potential to provide a flip to the pace of economic development and provide a unique opportunity to organizations in less developed countries to operate in global markets. This could also have implications on the socio-economic conditions of the less developed countries. Trust has been a major hurdle impeding the growth of e-commerce and the need for enhancing trust cannot be over-emphasized. The article, through validation, identifies various trust inducing dimensions for enriching the web interface and there by inducing trust. Focus on these dimensions, which demonstrate 'correct and concise' information, 'relevant structure' of information and 'usability', would help in inducing trust in the faceless environment of e-commerce. This in turn would improve the user experience and the buyer-seller relationship in online trading environment. This supports the earlier findings of Arion et al. (1994) wherein trust in human-computer interactions was stated to be supported by the system. It may, however, be noted that the trust in buyer-seller relationship is also influenced by a number of factors and the enhancement of features of web interface would need to be viewed as an integral part of an overall trust building strategy of an organization. Thus, by effectively incorporating various features of the trust inducing dimensions on the web interface, the merchant is able to provide a trust worthy platform for the customer across global boundaries to transact among each other. This has helped in the development of a virtual society which is driven by trust in technology and enhances the growth of e-commerce. Professional bodies and business organizations need to play an important role in this regard. Development and adherence to globally accepted standards in this regard can go a long way in enriching the 'user experience' in e-commerce.

REFERENCES

ACM Special Interest group on Computer-Human Interaction Group (1992). *ACM SIGCHI*, Technical Report. New York: ACM.

Arion, M., Numan, J. H., & Pitariu, H. (1994). Placing Trust in Human-Computer Interactions. *Proceedings of 7th European Cognitive Ergonomics Conference* (pp. 352-365).

Ba, S., & Pavlou, P. A. (2002). Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior. *MIS Quarterly, 26*(3)

Bailey, J. P., & Faraj, S. (2000).The Role of Intermediaries in the Development of Trust on the WWW: The Use and Prominence of Trusted Third Parties and Privacy Statements. *JCMC*, 5(3).

Bailey. A. (2005). Consumer Awareness and Use of Product Review Websites. *Journal of Interactive Advertising, 6*(1), 90-108.

Bhattacherjee, A. (2002). Individual Trust in Online Firms: Scale Development and Initial Test. *Journal of Management Information Systems, 19*(1), 211-241.

Chakraborty, G., Lala, V., & Warren, D. (2002). An Empirical Investigation of Antecedents of B2B Websites Effectiveness. *Journal of Interactive Marketing, 16*(4), 51-72 .

Cheskin/Sapient Research and Studio Archetype/Sapient. (1999). E-Commerce Trust Study, Retrieved June 10, 2008 from http://www.sapient.com/cheskin.

Corritore,C. l., Marble, R. P., Wiedenbeck, S., Kracher, B., & Chandran, A.(2005) Measuring Online Trust of Websites: Credibility, Perceived Ease of Use, and Risk. *Proceedings of the Eleventh Americas Conference on Information Systems*, Omaha, USA.

Davis, J., Mayer, R., & Shoorman, F. (1995). *An integrated model of organizational trust, the academy of management review*, *20*(3), 705-734.

Davis F. D. (1989). Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS Quaterly*, *13*(3), 319-340.

Doney, P. M., & Cannon, J. P. (1997). An Examination of the Nature of Trust in Buyer-Seller Relationships. *Journal of Marketing, 61*, 35-51.

Egger F. N. (2003). *From interactions to transactions: Designing the Trust Experience for Business-to-Consumer Electronic Commerce.*

Egger F. N. (2001). Affective Design of E-Commerce User Interfaces: How to maximize perceived trustworthiness. *Proceedings of The International Conference on Affective Human Factors Design*. London: Asean Academic Press.

Egger, F. N. (June 27-29, 2001). *Affective Design of E-Commerce User Interfaces: How to maximize perceived trustworthiness.* Conference on Affective Human Factors Design. Singapore, (pp. 317-324).

Fogg B. J., & Nass C. (1997). Effects of computers that flatter. *International Journal of Human-Computer Studies, 46*, 551-561.

Grabner-Kräuter & Kaluscha. (2003). Empirical research in on-line trust: a review and critical assessment. *International Journal of Human-Computer Studies, 58*(6), 783–812.

Gefen, D., & Straub, D. W. (2000). The Relative Importance of Perceived Ease-of-Use in IS Adoption: A Study of E-Commerce Adoption. *Journal of the Association for Information Systems, 1*(8), 1-28.

Gray, S. H. (1990). Using Protocol Analyses and Drawings to Study Mental Model Construction during Hypertext Navigation. *International Journal of Human-Computer Interaction, 2*(4), 359-377

Green, M. C. (2002). Perceptions of trust in Internet relationships. *Computer-supported social interaction conference*, Oxford, OH

Handy, C. (1995). Trust and the Virtual Organization. *Harvard Business Review, 73*(3), 40-50.

Jarvenpaa, S. L., & Tractinsky, N. (1999). Consumer Trust in an Internet Store: A Cross-Cultural Validation. *Journal of Computer Mediated Communication, 5*(2).

Jian, J-Y., Bisantz, A. M., & Drury, C. G. (1998). Towards an empirically determined scale of trust in computerized systems: Distinguishing concepts and types of trust. *In Proceedings of the Human Factors & Ergonomics Society 42nd Annual Meeting*, (pp. 501-505).

Keen, P. G. W. (1997). Are you ready for 'Trust' Economy. *Computer World, 31*(16), 80.

Kolsaker, A., & Payne, C. (2002). Engendering trust in e-commerce: a study of gender based concerns. *Marketing Intelligence & Planning, 20*(4), 206-214.

Lohse, G. L., & Spiller, P. (1998). Electronic Shopping: Quantifying the Effect of Customer Interfaces on Traffic and Sales. *Communications of the ACM, 41*(7)

Lee, J., Kim J., & Moon J. Y. (April 2000).What Makes Internet Users Visit Cyber Stores Again? Key Design Factors for Customer Loyalty. *Proceedings of the Computer-Human Interaction Conference on Human Factors in Computing Systems*, The Hague, Netherlands, (pp. 305-312).

Lee, J., & Huang, Y. (2005). The Design Factors of the Initial Trust in the Internet environment. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications,* (pp. 993-997).

Lingaard G. (1999). *Does emotional appeal determine perceived usability of web sites*? Retrieved June 8, 2008 from http://www.cyberg.com

Marsh, T., & Wright, P. (2000). Maintaining the illusion of interacting within a 3D virtual space. *Presence 2000: 3rd International Workshop on Presence*, Delft, The Netherlands.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709-734.

McMahon, K. (2005). *An exploration of the importance of website usability from a business perspective*. Retrieved July 1, 2008 from http://www.flowtheory.com/KTMDissertation.pdf

Moon, J.-W., & Kim, Y.-G. (2001). Extending the TAM for a World-Wide-Web Context. *Information & Management, 38*, 217-230.

Nass, C., Steuer, J., Tauber, E., & Reeder, H. (1993). Anthropomorphism, Agency, & Ethopoeia: Computers as Social Actors. *Conference of the ACM/SIGCHI & the IFIP* Amsterdam, the Netherlands (pp. 24-29).

Nielsen, J. (1999). *Trust or Bust: Communicating Trustworthiness in Web Design. Jacob*

Nielsen's Alertbox. Retrieved April 10, 2008 from http://www.useit.com/alertbox/990307.html.

Nielsen, J. (May 16, 1999). *Who Commits the 'Top Ten Mistakes' of Web Design*? Jacob Nielsen's Alertbox. Retrieved August 2, 2008 from http://www.useit.com/alertbox/990516.html.

Neilsen Normen group (2000). *Trust: Design guidelines for e-commerce user experience*. Retrieved September 6, 2008 from http://www.nngroup.com

Sheehan, K. B., & Doherty, C. (2001). Re-weaving the Web: Integrating Print and Online Communications. *Journal of Interactive Marketing, 15*(2), 47-59.

Radha, R., & Murphy, C. (1992). Searching Versus Browsing in Hypertext. *Hypermedia*. *4*(1), 31.

Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: Quality comes to services. *Harvard Business Review, 68*, 105-111.

Resnick, P., Zeckhauser, R., Friedman, E., & Kuwabara, K. (2000). Reputation Systems. *Communications of the ACM, 43*(12), 45-48.

Riegelsberger, J., Sasse, M. A., & McCarthy, J. D. (2003). Trust at First Sight? A Test of Users' Ability to Identify Trustworthy e-Commerce Sites. *Proceedings of HCI2003*, 8-12 243-260.

Rotter J. B. (1980). Impersonal trust, Trustworthiness and gullibility. *American Pschologist, 35*(1), 1-7.

Rust, R. T., & Kannan, P. K. (2003). E-Service: A new paradigm for business in the electronic environment. *Communications of the ACM, 46*(6), 36-43.

Utting, K., & N. Yankelovich (1989). Context and Orientation in Hypermedia Networks. *ACM Transactions on Information Systems, 7*(1) 58-84.

Wang, E., Barrett, S., Caldwell & Gavriel, S. (2003). Usability comparison: similarity and differences between e-commerce and world wide web. *Journal of the Chinese Institute of Industrial Engineers, 20*(3), 258-266.

Weiner, M., & Mehrabian, A. (1968). *Language within Language: Immediacy, a Channel in Verbal Communication*. New York: Appleton-Century-Crofts.

Xiaoni, Z. (2003). Factors Contributing to Purchase Intentions. *Internet Journal of Internet Commerce*, *2*(1).

Yoon, S. J. (2002). The antecedents and consequences of trust in online purchase decisions. *Journal of Interactive Marketing, 16*(2), 47-63.

This work was previously published in International Journal of Information Communication Technologies and Human Development, Vol. 1, Issue 1, edited by S. Chhabra, pp. 64-74, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 5.2
Swift Trust in Web Vendors:
The Role of Appearance and Functionality[1]

Xin Li
University of North Carolina at Pembroke, USA

Guang Rong
Clemson University, USA

Jason B. Thatcher
Clemson University, USA

ABSTRACT

With the growth of product search engines such as pricegrabber.com, Web vendors have many more casual visitors. This research examines how Web vendors may foster "swift trust" as a means to convert casual visitors to paying customers. We examine whether perceptions of Web sites' appearance features (normality, social presence and third-party links) and functionality features (security, privacy, effort expectancy and performance expectancy) positively relate to swift trust in a Web vendor. Using a quasi-experimental research design, we empirically test the proposed relationships. Based on an analysis of 224 respondents, we found appearance and functionality features explained 61% of the variance in swift trust. The article concludes with a discussion of findings and implications.

INTRODUCTION

Trust is a key enabler of e-commerce (Lee & Turban, 2001; Kracher, Corritore, & Wiedenbeck, 2005). Extensive research has found that trust leads to online end users reporting higher levels of purchasing intention, expressing greater loyalty to a Web vendor, and engaging in actual purchasing behavior (Gefen, McKnight, Choudhury, & Kacmar, 2002; Gefen, 2002b; Karahanna, &

Straub, 2003a). Given the extensive support in the literature for ties from online end users trust to their behavior, it is important to direct attention to identifying levers that foster trust and encourage online end users to complete transactions.

Fostering trust has proven problematic among fast-moving online end users. Although search engines direct over 85% of consumers to vendor Web sites, 57% of consumers abandon their shopping carts prior to checkout (DoubleClick, 2004). Perhaps due to the ease of identifying alternatives through engines such as pricegrabber.com, online consumers frequently visit Web sites and fail to complete transactions. Thus, Web vendors are under the pressure to quickly win trust, retain first time visitors and convert them to paying customers.

In this article, we examine how Web vendors may foster swift trust. Swift trust refers to trust formed quickly in a new relationship (Meyerson, Weick, & Kramer, 1996). Unlike other forms of trust which rest on experience or familiarity with a vendor, swift trust develops quickly during a consumer's first exposure to an unfamiliar Web vendor. Although first time visitors may lack experience or knowledge of a Web vendor, their perceptions of the "working conditions" of e-commerce, including appearance and functionality of a Web site (Zhang, von Dran, Small, & Barcellos, 1999), influence their understanding of a vendor's ability to complete transactions and its trustworthiness. For example, Gefen et al. (2003a) suggested that a fair and open Web site that clearly states due process, policies that handle the relationship, and provides clear explanations, engenders consumer trust. Hence, designers emphasize a Web site's appearance and functionality as means to promote swift trust.

Although substantial research has examined the implications of trust in Web vendors, research has left unexplored the influence of appearance and functionality on swift trust in a Web vendor. To gain deeper insight into online consumer behavior, this study examines the following research question:

How do the appearance and functionality of a Web site influence online consumers' swift trust in a Web vendor?

The article unfolds as follows. First, we develop a general model tying appearance and functionality to consumers' swift trust and purchasing intention. In terms of appearance, we examine normality, social presence, and links to third-party assurance providers. In terms of functionality, we examine the roles of security, privacy, effort expectancy and performance expectancy. Next, we present the results of a quasi-experimental study that tests the proposed model. Then, we discuss our findings. The article concludes with implications for research and practice.

LITERATURE REVIEW AND MODEL DEVELOPMENT

Trust is a complex and abstract concept. Although defined differently in many literatures, trust most commonly refers to one's willingness to depend on another based on the expectation that the other has the attributes to be trusted (Mayer, Davis, & Schoorman, 1995). Trust is driven by: 1) risk and uncertainty in relationships, 2) the trusting party's vulnerability, and 3) his/her expectations of the trusted party. When these drivers are present, an individual must extend trust to another in order for a relationship to exist.

E-commerce researchers extended the traditional trust definition and drivers to online shopping and found that they are important in studying online end user behavior (McKnight et al., 2002; Gefen et al., 2003a) (see Table 1). For example, high uncertainty characterizes complex and anonymous online transaction processes. While traditional customers can pay at the counter and take the products home immediately, online end users do not directly observe vendors nor

Table 1. Extending trust definition and drivers to the domain of e-commerce

	Traditional	**E-Commerce**
Trust Definition	The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party (Mayer et al.., 1995 p. 712).	
Prerequisites: Trustor and Trustee in a relationship	Two parties—a trusting party (trustor) and a party to be trusted (trustee)—in a social relationship (Wang & Emurian, 2005)	In the context of online shopping, the relationship is predicated on the exchange of money for goods or services. Online consumer is the trustor. Web vendor is the trustee (Wang & Emurian, 2005).
Drivers of Trust: Risk and Uncertainty	Trust is necessary in risky situations (Kee & Knox, 1970). When the trustor has no control or even no means to monitor the trustee's behaviors, the trustee would have chances to act opportunistically. The higher the risks and uncertainty exist, the more trust is needed to enable behaviors in the context (Ruyter, Moorman, & Lemmink, 2001).	Because of high complexity and anonymity, the e-commerce environment is more risky and uncertain than the traditional market (Wang & Emurian, 2005). Unobservability causes even more information asymmetries than those in traditional markets (Tan & Thoen, 2002).
Drivers of Trust: Vulnerability	Vulnerability refers to the extent to which a trustor would lose to the trustee's opportunistic behaviors. The more a trustor will lose, the more vulnerable s/he is in specific circumstances, and the higher level of trust is needed to justify the participation in such circumstances (Hosmer, 1995; Mishra, 1996).	E-commerce customers are exposed to both financial loss (e.g., credit card theft, defective or below-expectation product, delivery delay and damage, etc.) and privacy loss (e.g., identity theft, spam mails, etc.) (Friedman, Kahn, & Howe, 2000; Gefen, 2002a).
Drivers of Trust: Expectations	Reasonable expectation refers to trustor's belief that the trustee can be relied on. Usually a trustor will form a set of expectations about the trustee, including its competence, benevolence and integrity (Mayer et al., 1995). Based on the expectations, the trustor grants trust.	Online customers form expectations or beliefs of an e-vendor in terms of competence (ability of the vendor to do what the customer needs), benevolence (the vendor's caring and motivation to act in the customer's interests) and integrity (the vendor's honesty and promise keeping) (McKnight et al., 2002; Gefen et al., 2003a).
Outcomes of Trust Trusting Behavior and Relationship	Trust leads to risk-taking behavior. The form of the action depends on the situation (Mayer et al., 1995). Trust also leads to long-term relationships and anticipated future interaction (Mayer et al., 1995; Doney & Cannon, 1997).	Consumers' trust in an e-vendor leads to their risk-taking behaviors, such as following advice, providing information, and making purchases (McKnight et al., 2002). It also plays an important role in developing long-term relationships, like enhancing customer loyalty.

directly view the actual products they purchase. Given the salience of risk and uncertainty in online environments (Wang & Emurian, 2005), understanding how to foster trust becomes essential for understanding how to encourage end users to purchase goods online.

In light of risk and uncertainty, e-consumers accept some degree of vulnerability when they make online purchases. For example, when purchasing a good, a consumer accepts exposure to the chance that an unfamiliar vendor could misuse the credit card information. Consumers who accept cookies also open themselves up to having their activities tracked and information stolen. When consumers perceive higher levels of vulnerability (McKnight et al., 2002), they must extend trust to justify entering a business relationship with a Web vendor.

Consumers extend trust based on their expectations of a Web vendor. Expectations refer to perceptions of a vendor's trusting attributes (McKnight et al., 2002). Typically, online trust is conceived as a set of beliefs and expectations that about a vendor's: a) capability to operate a business and complete transactions (i.e., competence); b) desire to positively interact with consumers (i.e., benevolence); and c) honesty and adherence to widely accepted principles (i.e., integrity) (Mayer et al., 1995; McKnight et al., 2002; Gefen et al., 2003a). When dealing with a new Web vendor, consumers may not have enough direct knowledge or experience. They have to build such expectations based on other sources.

Swift Trust

To retain quick moving consumers, vendors must foster swift trust. Swift trust refers to trust formed quickly in new or transitory relationships (Meyerson et al., 1996). It forms without the benefit of familiarity, past experience, and fulfilled promises. Swift trust is similar to initial trust in that it develops early in a relationship, reflects limited interaction between a trustor and trustee, and rests on cues derived from the environment (Meyerson et al., 1996; McKnight, Cummings, & Chervany, 1998) (see Table 2 for a comparison of forms of trust). However, unlike initial trust, swift trust is developed under time constraints due to environmental pressures or consumers' short attention span (Galleta, 2006). Jarvenpaa and Leidner (1999) argue that swift trust is characterized by short time frames, which limit trustors' ability to develop expectations of trustee. Once individuals form swift trust, they are inclined to pursue future interactions with a trustor.

Swift trust is particularly germane to understanding online consumer behavior. Product search engines usually organize results based on price. Frequently, low-cost vendors use low prices to overcome a lack of name recognition. In this case, online consumers do not have past experiences or direct interactions with an unfamiliar trustee—the Web vendor—suggested by a search engine. Due to their numerous shopping choices, the online consumers may expend limited time and effort exploring an unfamiliar Web site. To "close the deal," Web vendors must foster swift trust, or lose the new customers to competitors.

Table 2. Developed trust, initial trust and swift trust

	Definition	Temporal	Studies	Trust Foundation
Developed trust	Trust built on shared experience and direct interactions between the two parties.	In an established relationship	(Ganesan, 1994; Gefen et al., 2003a)	Familiarity, Past Experiences, Direct Interactions
Initial trust	Trust in an unfamiliar trustee; a trust relationship in which the actors do not yet have credible, meaningful information about, or affective bonds with, each other.	In a new relationship	(McKnight et al., 2002; Koufaris & Hampton-Sosa, 2004)	Disposition to Trust*, Institutional-Based Trust including Situational Normality and Structural Assurances
Swift trust	Trust formed quickly in a new or temporary system, without traditional sources of trust such as familiarity, past experience, and fulfilled promises.	In a new relationship, with time constraints	(Meyerson et al., 1996; Jarvenpaa & Leidner, 1999)	Disposition to Trust*, Imported Trust and Category-Driven Trust based on quick cues

Because individuals lack sufficient time to form perceptions of new trustees from scratch, swift trust is usually "imported from other settings and imposed quickly in categorical forms" (Meyerson et al., 1996, p. 174). To engender swift trust in online shopping, vendors embed cues in the Web site design to facilitate the importing and categorization processes. Some cues provide links for consumers to import trust from other trusted sources (i.e., trust transference [Stewart, 2003]). For example, the presence of Visa and FedEx icons provides signal to consumers about safe payment and fast delivery. Other cues invoke categorization processes to speed up the trust formation. For instance, some vendors utilize page layout and menu sets similar with those known and trusted Web sites to encourage new consumers to categorize them as familiar. Industry certificates or awards are usually placed prominently on the homepage for consumers to label the vendors as professional. More Web sites present BBB and Trust-e seals, so consumers will classify them as reliable vendors. Thus, through embedding cues in Web design, new vendors can successfully foster swift trust among first time visitors.

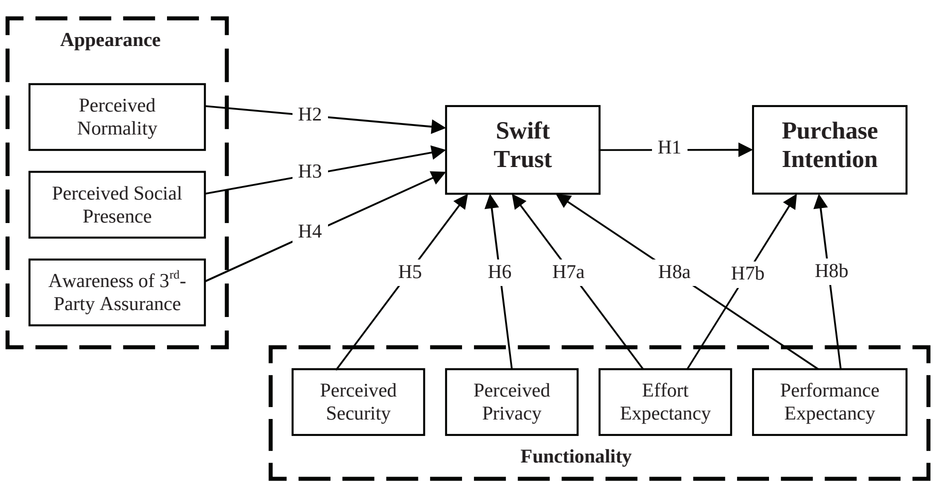
A Model of Swift Trust in Web Vendor

Web design consists of two major dimensions: appearance and functionality (Zhang et al., 1999). Appearance features refer to the look and feel of a Web site. They give the first impression to the Web visitors. Functionality features are more task-related. They refer to a Web site's ability to supply information and conclude transactions. Here, we propose a model (see Figure 1) depicting how a set of appearance and functionality features induce swift trust and purchase intention. In both appearance and functionality dimensions, we examine features that have been previously studied and are often noted as salient to fostering positive beliefs about a Web vendor and its Web site. While they are not exhaustive, the set of features will provide us some initial insight into swift trust formation. In the following sections, we discuss these appearance and functionality features and their relationships with swift trust and purchase intention.

Swift Trust and Purchase Intention

Purchase intention refers to online consumers' willingness to purchase from a Web vendor

Figure 1. A model of swift trust in Web vendor



(Stewart, 2003). Prior research suggests that purchase intention is influenced by consumers' trust in a Web vendor (Stewart, 2003; van der Heijden, Verhagen, & Creemers, 2003; Gefen et al., 2003a). Although formed quickly at the early stage of a relationship, swift trust encourages further interactions and facilitates the development of purchase intention. If consumers grant swift trust towards a Web vendor, they tend to believe that this vendor will protect their privacy, sell high-quality products, and be capable of consummating online transactions. Thus, we hypothesize that as swift trust in a vendor grows, consumers will report higher purchase intentions.

H1: *Swift trust in a Web vendor positively affects purchase intention.*

Appearance Features

A Web site is an electronic storefront of a Web vendor. For a pure e-commerce company, the Web site is the only access point through which consumers may collect information on, and interact with, the goods and services. Because of the Web site's centrality in the consumer-Web vendor relationship, a Web site's appearance is an important source of signals about the vendor's trustworthiness. Prior trust literature has mentioned several appearance features that affect trust (McKnight et al., 2002; Gefen et al., 2003a; Gefen & Straub, 2004). Based on the literature, we believe online consumers will express swift trust when they perceive a Web site's appearance as normal and high on social presence, and they are aware of third-party assurances of trustworthiness. In the following paragraphs, we explain the logic behind these relationships.

Perceived Normality

Perceived normality refers to consumers' perception that a Web site performs in a normal or customary manner, which is consistent with their experience with similar Web vendors (Gefen et al., 2003a). Normality of a Web site signals to new consumers that the vendor will fulfill their expectations of a good shopping experience (i.e., product information can be easily acquired, transactions are completed smoothly and safely, and expected goods and services are received quickly) (Gefen et al., 2003a). If a Web site is perceived as normal, consumers will categorize it as trustworthy and extend swift trust. Alternatively, an unusual Web site design that is not well-known and requires learning new ways to access information may not be trusted.

When visiting a new Web site, consumers assess normality through their quick interaction with the Web site. Specifically, they evaluate their perceptions of: 1) the shopping process, 2) the information required to complete a purchase, and 3) the quality of their interaction with the Web site (Gefen et al., 2003a). For example, most consumers are familiar with one-click shopping. If, on a new Web site, consumers find that they must go through a complicated process and provide unusually detailed information to complete a transaction, or they have a complicated or unpleasant communication with online representatives, they will not trust the Web vendor and fail to complete a transaction. Thus, we anticipate that consumers who perceive normality of a Web site are more likely to form swift trust towards a Web vendor.

H2: *Perceived normality positively affects swift trust in Web vendor.*

Perceived Social Presence

Perceived social presence refers to consumers' perception that there is personal, sociable, and sensitive human contact and/or peer community on the Web site (Gefen & Straub, 2004). A Web site can project a social presence similar to what consumers experience in off-line stores. Gefen and Straub (2004) found that when a Web site is perceived as having a higher level of social presence, it enables more effective interaction and communication between a consumer and Web vendor, and thus, renders consumers more inclined to trust the Web vendor. By projecting social presence similar to off-line stores, a Web site may induce consumers to categorize the Web site as part of a trustworthy group of known, off-line vendors.

In practice, although Web vendors are not able to use salespersons or store assistants to physically interact with customers, they send cues or project social presence through images of high-quality products, happy customers, and live-chat customer service functions. For example, on Overstock.com, service representatives use interactive, pop-up windows to ask online consumers if they need more information, or offer to help complete a transaction. Many Web sites foster virtual communities, where customers can participate in peer discussions, provide product/store reviews, and receive peer assistance (e.g., newegg.com and target.com). If, during the first visit, consumers perceive a Web site as high on social presence, we anticipate they will extend swift trust to a Web vendor.

H3: *Perceived social presence positively affects swift trust in Web vendor.*

Awareness of Third-Party Assurance

Awareness of third-party assurance refers to consumers' knowledge and awareness of third-party institutional mechanisms that ensure a trustworthy environment for Web site operations and business transactions (McKnight et al., 2002). Several institutional mechanisms are widely used in e-commerce, including seal programs (e.g., Trust-e reliability and privacy seal programs, VeriSign security seal program), credit card protections (e.g., Visa card security and protection programs, MasterCard security and credit basics), and customer certified store ratings and feedback systems (e.g., BizRate.com) (Pavlou & Gefen, 2004). Prior research suggests that these mechanisms provide a safe and reliable environment for online transactions. Through prominently placing seals and icons of these mechanisms, a Web site encourages consumers to extend trust based on their experiences with the established institutional structures. Based on the prior studies (McKnight et al., 2002; Stewart, 2003), consumers who are quickly aware of the third-party assurance on an unfamiliar Web site are more likely to extend swift trust towards the Web vendor.

H4: *Awareness of the third-party assurance positively affects swift trust in Web vendor.*

Functionality Features

Online customers' impression of a Web site's functionality, that is, its ability to successfully complete a transaction, influences their trust in the Web vendor. Although deep understanding of functionality requires extensive interaction with a Web site, cues in the Web site design may contribute to the formation of swift trust. Based on prior research, we direct our attention to four perceptions of functionality that may influence swift trust: security, privacy, effort expectancy and performance expectancy (Gefen et al., 2003a; Thatcher & George, 2004).

Perceived Security

Perceived security refers to consumers' perception that a Web site can safely complete transactions (Pavlou & Chellappa, 2001). When considering a transaction with an unknown party, consumers will evaluate the vendor's ability to protect their information, successfully complete a transaction, and offer help resolving a dispute. In the online environment, consumers usually perceive additional security threats than the traditional consumers, such as credit card theft, defective products/services, and delivery failure (Featherman & Pavlou, 2003). Consumers' security concerns are the major determinant of their online perceptions and behaviors (Salisbury, Pearson, Pearson, & Miller, 2001).

Consumers' assessment of the Web site's ability to safely complete transactions contributes to the formation of trust (Zhou, Dai, & Zhang, 2007). For example, Amazon.com provides notices when users sign in using the secure server. It gives customers easy access to the shipping and return policies on its homepage. By doing so, Amazon.com sends signals about its ability to facilitate safe, secure online shopping. Similarly, many Web vendors send signals about security through providing order status, real-time shipping tracking, and return assistance. When first time visitors feel fewer security concerns on a Web site, they will be more likely to report swift trust in the vendor.

H5: *Perceived security positively affects swift trust in Web vendor.*

Perceived Privacy

Perceived privacy refers to the consumers' belief that a Web vendor will protect consumers' personal and financial information (Pavlou & Chellappa, 2001). An online transaction requires consumers to provide more personal and financial information than traditional shopping channels. For example, online transactions require providing billing and shipping addresses, something that is not typically requested by brick-and-mortar stores. Also, e-commerce companies often use cookies to collect information and provide customized recommendations to consumers. Due to demands for consumer information, as well as a Web vendor's ability to collect information without consumers' notice, privacy concerns have become a defining issue for many Web consumers (Malhotra, Kim, & Agarwal, 2004).

When considering a transaction with an unfamiliar store, privacy may be a salient issue for a consumer (Liu, Marchewka, Lu, & Yu, 2005). A Web site may address privacy concerns through prominently featuring links to the privacy notice or policies, by briefly stating how they collect data on consumers, and by illustrating how such data may be used by the firm (Gefen et al., 2003a; Gefen, Rao, & Tractinsky, 2003b). When consumers quickly recognize that a Web vendor complies with the fair information practices and commits to protect its customers' privacy, they are more likely to express swift trust. Hence:

H6: *Perceived privacy positively affects swift trust in Web vendor.*

Effort Expectancy and Performance Expectancy

Within the domain of e-commerce, effort expectancy and performance expectancy have been tied to trust and purchase intentions (Pavlou, 2003; Gefen et al., 2003a). Effort expectancy (also referred to as ease of use [Venkatesh, Morris, Davis, & Davis, 2003]) refers to the degree to which the consumers believe that using a new Web site would be free of effort (Gefen et al., 2003a). Performance expectancy (also referred to as perceived usefulness [Venkatesh et al., 2003]) refers to the degree to which the consumers believe that using a new Web site would enhance their ability to safely, effectively purchase goods (Gefen et al., 2003a). Gefen et al. (2003a) found that customers' effort expectancy in using a Web site influences their trust. Their findings also indicate that trust, effort expectancy, and performance expectancy influence customers' usage intentions towards the Web site. Consistent with prior research, we hypothesize that when a new Web site is perceived as low on effort expectancy, consumers will be more likely to report swift trust as well as intentions to use a Web site. Hence:

H7a: *Effort expectancy negatively affects swift trust in Web vendor.*

H7b: *Effort expectancy negatively affects purchase intentions.*

In a departure from Gefen et al. (2003a), we argue that performance expectancy positively influences swift trust. Because first time visitors lack extensive experience with the tools provided by a Web vendor, expectations about performance reflect how they interpret cues embedded in a Web site. For example, a standard "useful" feature of a shopping site is the ability to search for products. Similarly, a comparison tool can help consumers compare similar products in terms of price, specifications and features in a more effective way. When consumers perceive that a Web site possesses tools that enable searching for and acquiring products, they will be more likely to express positive expectations of performance. Given that swift trust reflects limited interaction with a Web site, we anticipate that the presence of these "useful" tools positively influences swift trust and purchase intention. Hence:

H8a: *Performance expectancy positively affects swift trust in Web vendor.*

H8b: *Performance expectancy positively affects purchase intention.*

METHOD

A quasi-experimental survey design was used to test the research model. We used Overstock.com as our target Web site as it is a pure e-commerce company. Subjects would not confound trust in Overstock.com with their experiences at a real-world brick-and-mortar or click-and-mortar business.

We collected data at three public universities in the United States. Junior- and senior-level college students participated in the study. Although student samples have been criticized in many IS research contexts, evidence suggests that college students are active online shoppers, who do not differ from their working peers (Sen, King, & Shaw, 2006), and thus, represent a useful sample for testing our research model (Suh & Lee, 2005).

Study Procedure

The study consisted of a task and a survey. Participants were assigned a shopping task on Overstock.com. Specifically, subjects were asked to shop for a birthday gift for their significant others. They were given 10 minutes to view the Web pages and research a purchase. No actual purchase was required in the shopping task. Following completion of the task, subjects were asked to complete a survey about their experiences on Overstock.com. The survey collected data on all constructs in the research model and demographic information. Respondents were required to complete the task and survey in 25 minutes, thus engendering time pressure characteristic of situations involving swift trust.

Measurement Development

Where possible, we adapted measures from prior online trust research. Each item was measured using a seven-point Likert-type scale (anchored with 1=strongly disagree, 7=strongly agree). Purchase intention was measured using four items adapted from Pavlou's work (Pavlou, 2003; Pavlou & Gefen, 2004) (see Appendix I for the items).

Swift trust was operationalized as a second-order construct. It comprises three dimensions: competence, benevolence, and integrity. Each dimension was measured using three to four items adapted from McKnight et al. (2002).

Measures of most appearance features were adapted from prior studies. Perceived normality was measured by items adapted from Gefen et al. (2003a). Perceived social presence was measured by items adapted from Gefen and Straub (2004). To measure awareness of third-party assurance, we developed four items that directed attention to the presence of icons and links to trusted third parties such as VeriSign security seal program or credit card partners that appear on Overstock.com.

Functionality features measures were adapted from prior research. Effort expectancy and performance expectancy were adapted from Gefen et al. (2003a). Perceived security and perceived privacy were adapted from Salisbury et al. (2001), Yang and Jun (2002), and Smith et al. (1996).

Pilot Study and Data Collection

We conducted a pilot study to validate the measurement instruments and our quasi-experiment design. The instrument and task were well received by the pilot study participants. Minor changes were made in wording and the order of items. Also, we received feedback that our task was meaningful and induced a feeling of time pressure among our respondents.

Two-hundred eighty-one students participated in this study. We dropped 14 responses due to missing values. Also, we dropped 43 subjects because they reported prior experience with Overstock.com. This yielded a usable sample of 224 respondents that included 118 females and 106 males, with substantial Internet and online shopping experience. T-tests revealed that sample characteristics and responses across research sites were not significantly different. Table 3 presents sample characteristics.

RESULTS

We used partial least square (PLS), a structural equation modeling technique, to evaluate the hypotheses. Although a debate is emerging in the MIS literature on the merits of PLS, in the broader methods literature, PLS is considered a useful tool for theory building, specifically when one seeks to establish predictive validity in a structural model (Chwelos, Benbasat, & Dexter, 2001). As a result, we believe PLS is an appropriate tool for estimating our measurement and structural models. In the following sections, we present the results in two steps: measurement model and structural model.

Measurement Model

Two measurement models were evaluated. Following Agarwal and Karahanna's approach (2000), we first evaluated the measurement model for swift trust—a second order construct. Swift trust is formed by three reflective sub-dimensions: competence, benevolence and integrity. All items yielded high loadings (ranging from 0.855 to 0.948, $p < 0.000$) on the appropriate sub-dimension. Also, each sub-dimensions' internal consistency score (ranging from 0.908 to 0.941) and composite reli-

Table 3. Sample characteristics

Char.	Number	Percentage	Char.	Number	Percentage
Gender			Internet Experience – Years of Use		
Female	118	52.7%	0-3 years	5	2.2%
Male	106	47.3%	4-7 years	103	46.0%
Age			8 years or above	116	51.8%
Under 21	85	37.9%	Internet Experience - Often		
21 - 45	133	59.4%	Once a day or less	16	7.1%
Above 45	6	2.7%	Many times a day	208	92.9%
Internet Level			Online Shopping Experience – in past 2 years		
Beginner	2	0.9%	0-2 times	31	13.8%
Intermediate	34	15.2%	3-10 times	67	29.9%
Proficient	188	83.9%	11 time or above	126	56.3%

Table 4. Swift trust: Means, reliability and average variance extracted[1]

Dimension	Means	StDev	CR	Alpha	1	2	3
1. Competency	5.516	0.949	.952	.931	.832		
2. Benevolence	5.022	1.095	.943	.908	.707	.848	
3. Integrity	5.142	1.010	.958	.941	.698	.797	.852

[1]Diagonal elements in the 'correlation of constructs' matrix are the square root of the average variance extracted. For adequate discriminant validity, diagonal elements should be greater than corresponding off-diagonal elements.

Table 5. Overall measurement model: Means, reliability and average variance extracted[1]

	Means	StDev	CR	Alpha	1	2	3	4	5	6	7	8	9
1. Transaction Intention	5.008	1.348	.961	.944	.928								
2. Trust in Online Store	5.227	0.924	.933	.892	.533	.907							
3. Perceived Normality	5.961	1.028	.959	.935	.340	.538	.942						
4. Perceived Social Presence	4.231	1.351	.964	.953	.364	.396	.070	.918					
5. Awareness of 3rd-party	5.056	1.057	.927	.894	.414	.635	.400	.360	.872				
6. Effort Expectancy	6.221	0.921	.964	.949	.341	.520	.717	.119	.375	.933			
7. Performance Expectancy	5.498	1.144	.957	.940	.566	.530	.501	.338	.392	.583	.921		
8. Perceived Security	4.479	1.324	.960	.944	.413	.596	.316	.361	.545	.274	.349	.926	
9. Perceived Privacy	2.908	1.311	.947	.932	.064	.211	.185	.042	.184	.145	.087	.313	.864

[1]Diagonal elements in the 'correlation of constructs' matrix are the square root of the average variance extracted. For adequate discriminant validity, diagonal elements should be greater than corresponding off-diagonal elements

ability score (ranging from 0.943 to 0.958) were high. The detailed item loadings and reliability scores are reported in Appendix II (a) and Table 4, respectively. A confirmatory factor analysis was performed to assess the convergent and discriminant validity of the measures (see Appendix II (a) and Table 4). Briefly, all items load higher on the corresponding sub-dimensions than on other sub-dimensions, and each sub-dimension has a square root of the AVE (ranging from 0.832 to 0.852) greater than its correlations with other sub-dimensions. Hence, our analysis indicates adequate reliability, convergent, and discriminant validity of the swift trust measurement model.

Next, following Agarwal and Karahanna's approach (2000), we used the factor scores of the above three sub-dimensions as direct measures of swift trust to assess the full measurement model. Results are presented in Table 5 and Appendix II (b). All indicators load on the construct of interest, with loadings ranging from 0.829 to 0.965 ($p < 0.000$). Internal consistency scores range from 0.892 to 0.953. The composite reliability scores range from 0.927 to 0.964. All indicators loaded higher on the construct of interest than other constructs. Every construct has square root of AVE (ranging from 0.864 – 0.942) greater than its correlations with any other constructs. Hence, our analysis indicates adequate convergent and discriminant validity for the overall measurement model.

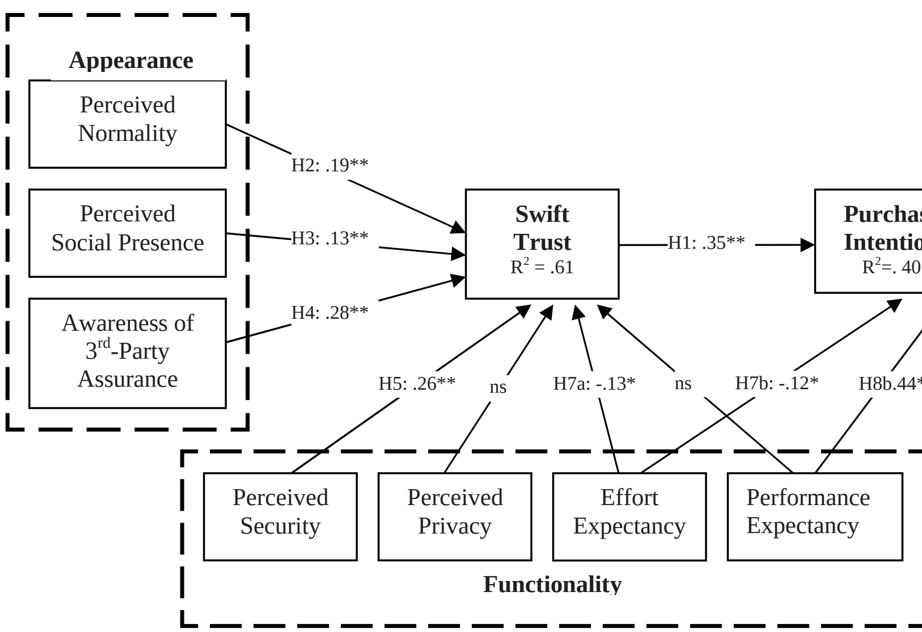
Structural Model

Figure 2 presents the structural model results. Appearance and functionality features explain 61% of the variance in swift trust. Also, swift trust, effort expectancy, and performance expectancy explained 40% of the variance in purchase intention.

Swift Trust

In terms of appearance features, we found perceived normality (H2: .19, $p<.01$), perceived social presence (H3: .13, $p<.01$), and awareness of 3rd party assurance (H4: .28, $p<.01$) positively relating to swift trust. In terms of functionality features,

Figure 2. Research model with PLS results



perceived security (H5: .26, p<.01) positively related to swift trust while effort expectancy (H7a: -.13 p<.05) negatively related to swift trust. Perceived privacy (H6: n.s.) and performance expectancy (H8a: n.s.) were not significantly related to swift trust. Collectively, our results provide evidence that a Web site's appearance and functionality attributes influence swift trust.

Purchase Intention

The overall model explains 40% of the variance in purchase intention. Swift trust (H1: .36, $p < 0.01$) and performance expectancy (H8b: .44, $p < 0.01$) positively relates to purchase intention. Effort expectancy (H7b: -.12, $p < 0.05$) negatively relates to purchase intention.

DISCUSSION

This study examined sources and consequences of swift trust in Web vendors. In the following paragraphs, we discuss our findings and the implications.

Appearance Features and Swift Trust

All of the three appearance features that we examined positively relate to swift trust. The study findings show that perceived normality has a significant positive relationship with swift trust (H2: Supported). Similar to other groups of online consumers, our respondents had significant Internet and e-commerce experience. Among our study participants, similarity with familiar Web sites engendered their swift trust in an unfamiliar Web site. This finding suggests that Web designers should carefully consider how they visually and structurally differentiate their Web site from competitors. If a Web site employs an unusual or novel design, first time visitors may not extend trust to a Web vendor. For new Web vendors, in addition to relying on low price or novelty of product lines/business models, our findings suggest patterning Web sites after successful e-commerce businesses in their market as a means to engender swift trust for first time visitors.

The study also shows that social presence positively related to swift trust in a Web vendor (H3: Supported). An unfamiliar Web site that projects social presence similar to real-world salespersons and customer service representatives engenders feelings of competence, benevolence, and integrity in online consumers (Gefen & Straub, 2004). Overstock.com uses images of satisfied customers and content employees to convey social presence. In addition, when customers check a product, a live assistance chat window frequently pops up and offers to provide more information or help completing a transaction. Further, below the item description, customers can read the product reviews posted by other customers or post their own reviews. All of these features give consumers the feeling like they were shopping at a physical store with helpful service representatives, and they were part of a community of shoppers (Thatcher & George, 2004). Thus, we suggest prominently projecting social presence as a useful means to foster consumers' swift trust.

First-time visitors reported higher levels of swift trust when they were aware of the third-party assurance (H4: Supported). Overstock.com places icons of third-party assurance providers on every Web page—including icons from credit card companies, payment options, VeriSign BBB Online, and BizRate.com. Our findings confirm that third-party icons signal that Overstock.com is operated in a safe and reliable environment, and the transactions in the environment are well protected (McKnight, Kacmar, & Choudhury, 2004). For practice, this finding directs Web designers to carefully consider the placement of third-party assurance icons and seals. If properly sized, placed at directly visible positions, repeated on different pages or different sections, the third-party assurance icons engender consumers' swift trust (McKnight et al., 2004).

Overall, appearance features are salient to understanding swift trust. This is probably because the appearance features are mostly visible to the first-time Web site visitors under time pressure. Our findings suggest that consumer perceptions of normality, social presence, and third-party assurances lead to swift trust in a Web vendor. For Web designers, these findings suggest that the look and feel of a Web site can be used to foster swift trust and convert first-time visitors to paying customers.

Functionality Features and Swift Trust

We found mixed results for functionality features' relationship with swift trust. Perceived security positively related to swift trust (H5: supported). Consumer perceptions of a Web site's capability to perform secure transactions influence their perceptions of the Web vendors' competence, benevolence, and integrity. Fully understanding the security features on a Web site may require time and extensive interaction. However, Web vendors can quickly signal their intention and ability to provide a safe shopping experience to the first time visitors through security cues. For example, Overstock.com provides prominent links to order tracking, shipping and return policies on its homepage. The transparent transaction processes may create a sense of security for online consumers. In addition, the "golden lock" signs on the login and check-out pages, as well as the timeouts and re-login requests, also provide quick signals of a Web site's security. If these security cues are prominently placed, first-time visitors can quickly perceive good protection and make quick inferences about the trustworthiness of a Web vendor. While most Web vendors expend significant efforts in developing security features that consumers may not directly see or understand, we suggest, to foster swift trust, Web vendors should give special attentions to how to signal the presence of the security features to online consumers.

Contrary to our expectations, perceived privacy was not significantly related to swift trust (H6: not supported). This finding may be a function of the task we used in our study. Participants searched for information and assessed features of a Web site; but they did not provide personal information required to complete a transaction. Perhaps due to the short and anonymous visit, our participants were not concerned with privacy issues associated with completing a purchase on Overstock.com. Because swift trust develops prior to making a purchase, we are comfortable with the strength of our research design. However, future research should examine ties between swift trust and an antecedent to privacy concerns. Hence, even though we did not find a relationship between privacy and swift trust, there remains room for future research on privacy and swift trust.

Effort expectancy negatively related to swift trust (H7a: Supported) and purchase intention (H7b: Supported). When using a Web site was perceived as relatively free of effort, consumers were more likely to perceive the Web vendor as competent, benevolent and high on integrity. While consistent with prior online consumer behavior research (Gefen et al., 2003a), our findings confirm the nomological net of relationships between effort expectancy, trust, and purchase intention. For practice, our findings underscore the importance of designing Web sites to be easy to navigate, easy to find product information, and easy to conclude transactions.

Performance expectancy was not related to swift trust (H8a: Not Supported) and it was related to purchase intention (H8b: Supported). Performance expectancy's influence on swift trust may have been washed out by perceived normality, perceived social presence, perceived security, and effort expectancy. While other perceptions refer to specific aspects of a Web site, performance expectancy requires consumers to make global assessments of a Web site. In light of this, it makes sense that perceptions of specific Web site attributes would exert greater influences on swift trust than a global assessment of a Web site. An alternative explanation for this finding may be that it requires consumers' substantial experience to form perceptions of performance expectancy. Because swift trust reflects only limited interaction with a Web site, online consumers may lack the experience to form beliefs about performance expectancy. Thus, it is not surprising that performance expectancy is not an antecedent to swift trust in a Web site. However, purchasing a product rests on general, well-developed perceptions of a Web site's functionality. Hence, performance expectancy, as expected, directly influences purchase intention (Gefen et al., 2003a).

Overall, online consumers report higher levels of swift trust when they report positive perceptions of a Web site's security and low levels of effort to use. The other functionality features may require more interactions to be realized and contribute to trust formation. Although perceptions of the features may evolve over time, Web vendors who embed noticeable signals that they are capable to complete secure, easy transactions are likely to engender consumers' swift trust and purchasing intention in their Web sites.

Swift Trust and Purchase Intention

Swift trust positively relates to purchase intention (H1: Supported). When online consumers quickly form trust, they are more likely to participate in further interaction with the Web site and eventually make actual purchases from the vendor. This finding is consistent with prior trust research (Gefen, 2002b; see Gefen et al., 2003a; 2003b; Pavlou & Gefen, 2004). Our finding also supports the idea that trust is relevant to purchasing intention even when consumers have limited time to familiarize themselves with a Web vendor. Hence, this finding underscores the importance of the previous findings on how appearance and functionality features influence the formation of swift trust in a Web vendor. For practice, although the casual visitors led by search engines may have many alternatives and move fast, Web vendors still can successfully foster swift trust and convert them to paying customers through the proper use of the appearance and functionality features.

CONCLUSION

This study examined swift trust in Web vendors. To induce conditions necessary for the formation of swift trust, we constrained study participants' time to visit a Web site. Also, we dropped responses from study participants with experience with our target Web site—Overstock.com. Based on our study, we found substantial support for appearance and functionality features positively relating to swift trust. Also, we found supports for swift trust, effort expectancy, and performance expectancy influencing purchase intention.

For research, this study provides initial evidence that swift trust has different antecedents than other forms of experience-based trust identified in the e-commerce literature—it was not affected by perceived privacy or performance expectancy. This suggests that future research should examine additional factors that shape the development of swift trust. For practice, our research underscores the importance of fostering swift trust as a means to increase purchase intention. Our findings direct attention to approaches under managerial control such as normality (i.e., using Web site designs similar to known vendors) and social presence (i.e., engendering a sense of community or having interactive help features) that positively influence the formation of swift trust.

In light of findings, it is important to note several limitations and opportunities for extending our research. An important limitation of this study is the use of a single Web site, Overstock.com. We used this approach to control for variance in how a Web site's features and respondents' experience influence swift trust. Overstock.com possesses many favorable features such as recommendation agents, icons, and product descriptions that may foster swift trust. In future research, it would be useful for researchers to examine how manipulating these features would influence consumers' swift trust in other unfamiliar Web vendors.

Also, future research should examine how to foster swift trust in diverse populations of online consumers. In this research, we drew on students who are active online consumers to examine swift trust. Although students are a valid sample for the current study, a pure student sample may include less variance in age, occupation, and Internet experience. For example, research suggests that age influences perceptions of technology among members of the workforce (Morris & Venkatesh, 2000). Hence, future research should examine differences in demographic factors such as age, income or occupation, which may relate to the formation of swift trust in e-commerce.

REFERENCES

Agarwal, R., & Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS Quarterly, 24*(4), 665-694.

Chwelos, P., Benbasat, I., & Dexter, A. (2001). Research report: Empirical test of an EDI adoption model. *Information Systems Research, 12*(3), 304-321.

Doney, P., & Cannon, J. (1997). An examination of the nature of trust in buyer-seller relationships. *Journal of Marketing, 61,* 35-51.

DoubleClick. (2004). *DoubleClick Q3 2004 ad serving trend report.* Doubleclick.com.

Featherman, M., & Pavlou, P. (2003). Predicting e-services adoption: A perceived risk facets perspective. *International Journal of Human-Computer Studies, 59*(4), 451-474.

Friedman, B., Kahn, P., & Howe, D. (2000). Trust online. *Communications of the ACM, 43*(12), 34-40.

Galleta, D., Henry, R., McCoy, S., & Polak, P. (2006). When the wait isn't so bad: The interacting effects of Web site speed, familiarity, and breadth. *Information Systems Research, 17*(1), 20-37.

Ganesan, S. (1994). Determinants of long-term orientation in buyer-seller relationships. *Journal of Marketing, 58*(2), 1-19.

Gefen, D. (2002a). Nurturing clients' trust to encourage engagement success during the customization of ERP systems. *Omega, 30,* 287-299.

Gefen, D. (2002b). Customer loyalty in e-commerce. *Journal of Association for Information Systems, 3,* 27-51.

Gefen, D., Karahanna, E., & Straub, D. (2003a). Trust and TAM in online shopping: An integrated model. *MIS Quarterly, 27*(1), 51-90.

Gefen, D., Rao, V., & Tractinsky, N. (2003b). The conceptualization of trust, risk and their relationship in electronic commerce: The need for clarifications. *Paper presented at the 36th Hawaii International Conference on System Sciences.* Big island, Hawaii.

Gefen, D., & Straub, D. (2004). Consumer trust in B2C e-commerce and the importance of social presence: Experiments in e-products and e-services. *Omega, 32,* 407-424.

Hosmer, L. (1995). Trust: The connecting link between organizational theory and philosophical ethics. *Academy of Management Review, 20*(2), 379-403.

Jarvenpaa, S., & Leidner, D. (1999). Communication and trust in global virtual teams. *Organization Science, 10*(6), 791-815.

Kee, H., & Knox, R. (1970). Conceptual and methodological considerations in the study of trust and suspicion. *Journal of Conflict Resolution, 14*(3), 357-366.

Koufaris, M., & Hampton-Sosa, W. (2004). The development of initial trust in an online company by new customers. *Information and Management, 41,* 377-397.

Kracher, B., Corritore, C., & Wiedenbeck, S. (2005). A foundation for understanding online trust in electronic commerce. *Information, Communications and Ethics in Society, 3*(3), 131-141.

Lee, M., & Turban, E. (2001). A trust model for consumer internet shopping. *International Journal of Electronic Commerce, 6*(1), 75-91.

Liu, C., Marchewka, J., Lu, J., & Yu, C. (2005). Beyond concern—a privacy-trust-behavioral intention model of electronic commerce. *Information and Management, 42,* 289-304.

Malhotra, N., Kim, S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research, 15*(4), 336-355.

Mayer, R., Davis, J., & Schoorman, F. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709-734.

McKnight, D., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research, 13*(3), 334-359.

McKnight, D., Cummings, L., & Chervany, N. (1998). Initial trust formation in new organizational relationships. *Academy of Management Review, 23*(3), 473-490.

McKnight, D., Kacmar, C., & Choudhury, V. (2004). Shifting factors and the ineffectiveness of third party assurance seals: A two-stage model of initial trust in a Web business. *Electronic Markets, 14*(3), 252-266.

Meyerson, D., Weick, K., & Kramer, R. (1996). Swift trust and temporary groups. In: R. Kramer & T. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research,* (pp. 166-195). Thousand Oaks, CA: Sage Publications.

Mishra, A. (1996). Organizational responses to crisis: The centrality of trust. In: R. Kramer & T. Tyler (Eds.), *Trust in organizations,* (pp. 261-287). Newbury Park, CA: Sage Publications.

Morris, M., & Venkatesh, V. (2000). Age differences in technology adoption decisions: Implications for a changing workforce. *Personnel Psychology, 53*(2), 375-403.

Pavlou, P. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International Journal of Electronic commerce, 7*(3), 101-134.

Pavlou, P., & Chellappa, R. (2001). The role of perceived privacy and perceived security in the development of trust in electronic commerce transaction. Unpublished work.

Pavlou, P., & Gefen, D. (2004). Building effective online marketplaces with institution-based trust. *Information Systems Research, 15*(1), 37-59.

Ruyter, K., Moorman, L., & Lemmink, J. (2001). Antecedents of commitment and trust in customer-supplier relationships in high technology markets. *Industrial Marketing Management, 30,* 271-286.

Salisbury, D., Pearson, R., Pearson, A., & Miller, D. (2001). Perceived security and World Wide Web purchase. *Industrial Management and Data Systems, 101*(3-4), 165-176.

Sen, R., King, R., & Shaw, M. (2006). Buyers' choice of online search strategy and its managerial implications. *Journal of Management Information systems, 23*(1), 211-238.

Smith, H., Milberg, S., & Burke, S. (1996). Information privacy: Measuring individuals' concerns about organizational practices. *MIS Quarterly, 20*(2), 167-196.

Stewart, K. (2003). Trust transfer on the World Wide Web. *Organizational Science, 14*(1), 5-17.

Suh, K., & Lee, Y. (2005). The effects of virtual reality on consumer learning: An empirical investigation. *MIS Quarterly, 29*(4), 673-697.

Tan, T., & Thoen, W. (2002). Formal aspects of a generic model of trust for electronic commerce. *Decision Support Systems, 33,* 233-246.

Thatcher, J., & George, J. (2004). Commitment, trust, and social involvement: An exploratory study of antecedents to Web shopper loyalty. *Journal of Organizational Computing and Electronic Commerce, 14*(4), 243-268.

van der Heijden, H., Verhagen, T., & Creemers, M. (2003). Understanding online purchase intentions: Contributions from technology and trust perspectives. *European Journal of Information Systems, 12,* 41-48.

Venkatesh, V., Morris, M., Davis, G., & Davis, F. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425-478.

Wang, T., & Emurian, H. (2005). An overview of online trust: Concepts, elements, and implications. *Computers in Human Behavior, 21,* 105-125.

Yang, Z., & Jun, M. (2002). Consumer perception of e-service quality: From Internet purchaser and non-purchaser perspectives. *Journal of Business Strategies, 19*(1), 19-41.

Zhang, P., von Dran, G., Small, R., & Barcellos, S. (1999). Web sites that satisfy users: A theoretical framework for Web user interface design and evaluation. *Paper presented at the 32nd Hawaii International Conference on System Sciences.* Maui, HI.

Zhou, L., Dai, L., & Zhang, D. (2007). Online shopping acceptance model—a critical survey of consumer factors in online shopping. *Journal of Electronic Commerce Research, 8*(1), 41-62.

ENDNOTE

1 We specially thank the special issue editor for granting extra space to include the appendices.

APPENDIX I: SURVEYS AND MEASURES

Construct	Items
Appearance Features	
Perceived Normality	The steps required at this Web site to search for products and make orders are typical of other similar type Web sites.
	The information requested of me at this Web site is the type of information most similar type Web sites request.
	The nature of the interaction with this Web site is typical of other similar type Web sites.
Perceived Social Presence	There is a sense of human contact in this Web site.
	There is a sense of personalness in this Web site.
	There is a sense of sociability in this Web site.
	There is a sense of human warmth in this Web site.
	There is a sense of human sensitivity in this Web site.
Awareness of Third-Party Assurance	The display of the VeriSign icon on Overstock.com makes me feel that this Web site is trustworthy.
	The display of the BBBOnline icon on Overstock.com makes me feel that this Web site is trustworthy.
	The display of the Visa, MasterCard, American Express, and DiscoverCard icons on Overstock.com makes me feel that this Web site is trustworthy.
	The display of the BizRate.com icon on Overstock.com makes me feel that this Web site is trustworthy.
Functionality Features	
Effort Expectancy (Reverse Coded)	I find this Web site easy to use.
	It would be easy for me to become skillful at using this Web site.
	Learning to use this Web site would be easy to me.
	My interaction with this Web site is clear and understandable.
Performance Expectancy	I find this Web site useful.
	Using this Web site can improve my shopping performance.
	Using this Web site can enhance my shopping effectiveness.
	Using this Web site can increase my shopping productivity.
Perceived Security	I feel secure in providing sensitive information (e.g., credit card number) when transacting with Overstock.com.
	I would feel totally safe providing sensitive information about myself to Overstock.com.
	I would feel secure sending sensitive information to Overstock.com.
	Overall, Overstock.com is a safe place to send sensitive information.

Construct	Items
Perceived Privacy (Reverse Coded)	I am concerned that Overstock.com is collecting too much information about me.
	It bothers me when Overstock.com asks me for personal information.
	I am concerned about my privacy when browsing Overstock.com.
	I have doubts as to how well my privacy is protected on Overstock.com.
	My personal information could be misused when transacting with Overstock.com.
	My personal information could be accessed by unknown parties when transacting with Overstock.com.
Swift Trust in Online Store	
Competence	I believe this online store is effective in assisting and fulfilling my purchases.
	This online store performs its role of e-vendor very well.
	Overall, this online store is a capable and proficient e-vendor.
	In general, this online store is very knowledgeable about the business it operates.
Benevolence	I believe that this online store would act in my best interest.
	If I required help, this online store would do its best to help me.
	This online store is interested in my well-being, not just its own.
Integrity	This online store is truthful in its dealings with me.
	I would characterize this online store as honest.
	This online store would keep its commitments.
	This online store is sincere and genuine.
Purchase Intention	
	Given the need, I intend to transact with Overstock.com.
	Given the chance, I think that I would consider making purchases from Overstock.com.
	I would probably purchase from Overstock.com when I have a need.
	It is likely that I will actually buy products from Overstock.com in the near future.

APPENDIX II: CONFIRMATORY FACTOR ANALYSES: FACTOR LOADINGS & CONVERGENT VALIDITY

(a) First-Order Constructs

	Com	Ben	Int
Com1	0.855	0.568	0.581
Com2	0.948	0.664	0.650
Com3	0.929	0.679	0.678
Com4	0.915	0.666	0.636
Ben1	0.693	0.935	0.755
Ben2	0.664	0.914	0.706
Ben3	0.597	0.914	0.741
Int1	0.639	0.743	0.917
Int2	0.689	0.746	0.947
Int3	0.633	0.706	0.895
Int4	0.616	0.748	0.933

(b) Second-Order Constructs

	Intent	Trust	Normality	Social Presence	Aware-ness	Effort Expect.	Perf. Expect.	Security	Privacy
Intent1	0.936	0.489	0.336	0.309	0.370	0.298	0.475	0.382	0.080
Intent2	0.947	0.565	0.401	0.330	0.439	0.397	0.553	0.397	0.054
Intent3	0.958	0.498	0.344	0.353	0.395	0.350	0.556	0.395	0.084
Intent4	0.885	0.430	0.175	0.365	0.332	0.214	0.522	0.366	0.020
Com	0.479	0.886	0.561	0.267	0.567	0.532	0.480	0.494	0.206
Ben	0.499	0.925	0.449	0.405	0.562	0.445	0.505	0.566	0.214
Int	0.480	0.922	0.463	0.408	0.607	0.445	0.463	0.568	0.158
WN1	0.307	0.507	0.946	0.055	0.356	0.698	0.486	0.281	0.180
WN2	0.347	0.527	0.938	0.115	0.440	0.641	0.462	0.343	0.204
WN3	0.308	0.493	0.955	0.026	0.335	0.697	0.474	0.271	0.138
SP1	0.368	0.382	0.048	0.916	0.344	0.101	0.331	0.332	0.083
SP2	0.334	0.360	0.097	0.888	0.368	0.135	0.333	0.321	0.070
SP3	0.319	0.376	0.070	0.940	0.330	0.093	0.289	0.350	0.036
SP4	0.331	0.338	0.045	0.931	0.306	0.106	0.281	0.305	-0.044
SP5	0.326	0.368	0.064	0.937	0.311	0.115	0.324	0.354	0.041
AW1	0.348	0.612	0.410	0.289	0.893	0.397	0.354	0.500	0.214
AW2	0.307	0.544	0.289	0.274	0.886	0.317	0.364	0.444	0.164
AW3	0.374	0.537	0.393	0.322	0.848	0.328	0.311	0.445	0.119
AW4	0.426	0.524	0.301	0.384	0.875	0.261	0.342	0.519	0.138
PEOU1	0.334	0.520	0.688	0.128	0.368	0.950	0.583	0.273	0.186
PEOU2	0.302	0.474	0.706	0.080	0.361	0.921	0.531	0.233	0.123
PEOU3	0.306	0.444	0.619	0.103	0.280	0.923	0.512	0.217	0.136
PEOU4	0.333	0.505	0.672	0.133	0.391	0.953	0.554	0.300	0.098
PU1	0.482	0.480	0.478	0.271	0.340	0.589	0.836	0.330	0.159
PU2	0.548	0.496	0.463	0.299	0.386	0.512	0.965	0.315	0.041
PU3	0.538	0.500	0.450	0.337	0.373	0.541	0.958	0.317	0.026
PU4	0.526	0.485	0.466	0.343	0.350	0.519	0.936	0.332	0.102
Secu1	0.410	0.542	0.315	0.278	0.451	0.235	0.273	0.897	0.296
Secu2	0.364	0.511	0.251	0.378	0.493	0.217	0.338	0.939	0.300
Secu3	0.389	0.531	0.266	0.341	0.507	0.245	0.314	0.961	0.310
Secu4	0.375	0.620	0.335	0.347	0.567	0.312	0.367	0.924	0.263
Priv1	0.085	0.153	0.237	-0.057	0.174	0.230	0.098	0.218	0.829

(b) continued

Priv2	0.054	0.178	0.190	-0.001	0.179	0.174	0.095	0.273	0.839
Priv3	0.051	0.174	0.160	0.041	0.170	0.132	0.058	0.265	0.878
Priv4	0.061	0.219	0.174	0.019	0.167	0.133	0.095	0.321	0.929
Priv5	0.025	0.187	0.112	0.096	0.154	0.051	0.040	0.280	0.877
Priv6	0.062	0.182	0.104	0.107	0.117	0.057	0.068	0.259	0.855

This work was previously published in Journal of Organizational and End User Computing, Vol. 21, Issue 1, edited by M. Mahmood, pp. 88-108, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 5.3

Understanding Brand Website Positioning in the New EU Member States: The Case of the Czech Republic

Shintaro Okazaki
Universidad Autónoma de Madrid, Spain

Radoslav Škapa
Masaryk University Brno, Czech Republic

ABSTRACT

This study examines Websites created by American multinational corporations (MNCs) in the Czech Republic. Utilizing a content analysis technique, we scrutinized (1) the type of brand Website functions, and (2) the similarity ratings between the home (US) sites and Czech sites. Implications are discussed from the Website standardization versus localization perspective.

INTRODUCTION

Both academics and practitioners have long debated whether advertising messages should be standardized. The proponents of standardization argue that the use of uniform advertising would provide significant cost benefits, thus improving company performance in the short run, while creating a consistent brand image in multiple markets. In contrast, the proponents of localization contend that ignoring the cultural, social, and economic characteristics of particular markets would cause psychological rejection by local consumers, thus decreasing profits in the long run. The debate has also produced a compromised or hybrid approach, which suggests that whether to standardize or localize advertising in a given market is a question of degree, and it is necessary to analyze many factors on a case-by-case basis (Mueller, 1991).

This debate is not limited to traditional media. As multinational corporations (MNCs) integrate their marketing communication with an emergent interactive medium, websites are becoming increasingly important for brand marketing and customer relationship management in multiple markets. This is because the Internet is, by definition, a *glocal* medium, which allows companies to create localized content with global access. In fact, many MNCs have established so-called "global gateway" sites with several language options. Consumers can first choose the language, then seek the information they desire. In this regard, the content of the local sites may need to be adapted to local consumers' tastes and preferences, in terms of design, layout, copy, message, and so forth. (Okazaki and Alonso, 2002).

Okazaki (2005) examined websites created by American MNCs' in four EU member states (i.e., the UK, France, Germany, and Spain), and found a high level of localization in website communication strategy. This research extends Okazaki's exploration into the new EU member states, by conducting a content analysis of the MNCs' websites created in the Czech Republic. Specifically, we address the following questions: (1) What types of brand website functions are used? (2) To what extent are the Czech sites standardized?

SIGNIFICANCE OF THE STUDY

This study will be an interesting addition to the literature of global information technology for two reasons. First, prior research provides little information on how the content created by the most globally diffused information technology, the Internet, has been standardized in foreign markets. Information managers in global markets should be aware of a question of transmitting culturally bound meanings into local sites. Secondly, this study addresses how design features and website functions can be used as a tool to create a universal imagery in global websites. Specifically, this study explores one of the most understudied countries in Europe: the Czech Republic. After joining the European Union, studies on information technology in this new member state is almost non-existent, thus, this research makes a unique contribution to the literature.

ENLARGEMENT OF THE EUROPEAN UNION

In 2004, the enlargement of the European Union increased its member states from 15 to 25, by adding 10 countries: Cyprus, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia, and Slovenia. In 2007, two more countries, Romania and Bulgaria became the member states, making the Union of 27 countries. This drastic expansion changed the way multinational corporations (MNCs) operate their businesses in Europe. Because of these countries' low labour costs and investment incentives (e.g., tax reduction, construction aid, etc.), many firms moved their production facilities from other regions to these new member states. For example, Sheram and Soubbotina (2000) report that "Countries seen as more advanced in market reforms—the Czech and Slovak Republics, Hungary, and Poland—attracted almost three-quarters of foreign investment" in transition economies. In fact, Poland received approximately $6.4 billion in foreign direct investment in 2003, an increase of $360 million over the previous year (MacKay 2004).

As these new EU Member States experience rapid economic expansion, global marketing influences consumers in them more and more. Their product experiences increasingly resemble those of their "Western" neighbours. In this light, it is reasonable to argue that the role of advertising in everyday consumption has also undergone a drastic transition, in both content and executions. For example, in the Czech Republic, advertising spending reached 563 million euros in 2004, while the average annual growth rate over the last 5 years has been 5%. Multinational corporations (MNCs) are the largest advertisers in these countries.

MEDIA USAGE IN EASTERN AND CENTRAL EUROPE

The Czech Republic

In the Czech Republic, television has traditionally been the primary vehicle for advertising, accounting for 46% of the MNCs' marketing budgets. Print media is the second medium with

34%, while outdoor advertising (i.e., billboard) is third, with 8% of total advertising expenditure (OMD Czech, 2005). However, the rapid growth of the Internet has significantly affected this media distribution. According to the Czech Publishers Association, the share of Internet advertising has been estimated at approximately 4%, or 25 million Euro (760 millions CZK), with a growth rate of almost 80% in 2004 (Unie vydavatelů, 2005). The telecommunications, financial, and automobile companies are the heaviest users of the Internet as an advertising medium (Unie vydavatelů, 2005).

In 2005, nearly 30% of Czech households had a personal computer (Czech Statistical Office, 2005). Internet penetration is increasing steadily in the Czech Republic. Nowadays, 35% of the adult population in the Czech Republic uses the Internet, almost twice the number of Internet users in 2000. The Czech Republic has thus clearly outmatched other Central European countries: for example, Bulgaria (16%), Hungary (22%), and Poland (31%). However, it has not yet achieved the levels of Internet penetration in Estonia (51%) or Slovenia (56%). The most dynamic increase is found in older people (GfK, 2006), even though the Internet use remains the domain of younger people. A quarter of Czech citizens have an Internet connection at home.

Searching for information is one of the most frequent activities on the Internet, according to a survey by the Czech Statistical Office (2005). In the most recent quarter to be surveyed, 62% of the Internet users used the Internet to find information about goods and services, 54% used it to find and download professional texts, 38% looked for services related to travel and accommodation, 36% to read and download on-line newspapers and magazines, and 28% to play or download games or music. However, the number of individuals with e-shopping experience increased rapidly between 2005 and 2006 (the survey was carried out in the first quarters of 2005 and 2006): it amounted to 14% in 2006, while a year before it had been only 6% (Czech Statistical Office, 2006). The most popular items in the Internet shopping are electronics, books, journals, textbooks, tickets and travel services, and accommodation. Online shopping is typically used more by men than women, and by the younger generation groups, between 25 and 45 years.

Approximately 12,500 Czech enterprises purchased goods or services via the Internet in 2003, almost 30% of the total number. The value of Internet purchases reached 2.8% of total purchases, and the value of Internet sales reached nearly 2.1% of total sales in these enterprises.

Poland and Hungary

Along with the Czech Republic, Poland and Hungary make up the fastest-growing economic region within the new EU member states. For example, the rapid transformation of the Polish economy is reflected in the accelerated growth of its advertising market. Between 1996 and 1999, average annual growth in advertising expenditure was more than 40%, which can be attributed to the drastic structural changes, and the subsequent economic boom, in this period. With regard to media share, television was the most popular (48%), with print media second (35%). On-line advertising, including websites, remains far behind traditional media, representing approximately 1% of total media spending (Zenith Optimedia, 2004). The telecommunications, financial, and automobile industries are the heaviest users of the Internet for advertising, promotion, and transactions (Agora, 2005).

In 2005, 30% of households in Poland had the technical possibility of the Internet access. In term of quality of connections, the survey found that only 16% of Polish households used a broadband connection. The significant disparity in the Internet infrastructure is between urban areas and the countryside, where the penetration of broadband connections is four times lower than in urban areas (Eurostat, 2005).

The Internet usage by Polish enterprises is below the EU average: in 2005, 87% of enterprises used the Internet connection. The share of broadband the Internet connections was 43%. More than 67% of companies have their own website homepage (Eurostat, 2006). Online purchases have not yet become popular. Only 5% of Polish consumers ordered goods or services via the Internet in 2005 (Eurostat, 2006). According to the survey by GfK (2006), only 4% of the Internet users make a purchase on the Internet once a month or more. A further 6% buy online once every two to three months, while 18% go online sporadically with the intention of buying something. The most frequently purchased items include books, CDs, clothes, and shoes. Less frequently, people buy DVDs and air tickets (GfK, 2006).

In B2B the situation is similar. In 2005, only 9% of the enterprises surveyed ordered products or services via the Internet. Sales via the Internet were lower, with only 4% of enterprises selling via the Internet. In 2005, turnover from e-commerce via the Internet amounted to 1.6% of total turnover (Eurostat, 2005).

Similarly, the Hungarian market has shown a drastic growth in market size and advertising spending. According to the Budapest Business Journal (BBJ, 2004), advertising expenditure in television media reached 213 million euros by 2003. The print media also showed a drastic growth, to spending of 212 million euros. In 2003, the online advertising market expanded by approximately 30%, achieving a 2% share of the total media market. The principal reasons for this growth were an increased number of the Internet users in younger generations, and the rapid proliferation of broadband high-speed connection. The largest online advertisers include car dealers, telecommunication companies, beer makers, and cosmetics firms (BBJ, 2003).

Other EU Member States

In 2005, Slovenia had the highest rate of the Internet usage in the new member states, both for households (48%), and for enterprises (96%) (Eurostat, 2006). The lowest rates of access were in Lithuania, for households (16%), and in Latvia, for enterprises (75%). The largest disparities in the Internet access between households and enterprises were recorded in the Czech Republic, Lithuania, and Slovakia. The number of individuals who have never used the Internet outweighs the number of regular users in the new member states. That differs from the situation in the old member states.

There is also disparity in the presence of companies' websites on the Internet. In January 2005, 62% of enterprises in the EU were equipped with a website, but only 49% in the new member states. The lowest percentages of companies with websites were found in Latvia, Hungary, and Lithuania. Most enterprises use the content of their web presentations mainly to market their products. Less than half use it to display catalogues of their products, services, and prices. One quarter use websites to offer after-sales service to their customers. Apart from the Czech Republic, enterprises in the new member states registered lower rates than the EU average for purchases, sales, and for total sales on the Internet, as a percentage of their overall turnover (Eurostat, 2006).

The e-readiness rankings of the Economist Intelligence Unit can be seen as a complex indicator of the level of ICT of a country's infrastructure. The index is a weighted collection of nearly 100 quantitative and qualitative criteria, which assesses the "state of play" of a country's information and communications technology (ICT) infrastructure, and the ability of its consumers, businesses, and governments to use ICT to their benefit. In the 2006 e-readiness rankings, Estonia achieved the best position of all the new EU member states (27^{th}), whereas Latvia (39^{th}) was lowest. By comparison, ten of the fifteen old EU members were in the top 20.

The Networked Readiness Index, published annually in the Global Information Technology Report, is a similar index. This index captures such aspects as available ICT infrastructure, and

Table 1. Descriptive statistics

	Population[1]	GDP per capita[2]	Advertising spending[3]	Advertising spending as % of GDP[4]	Internet penetration[5]	Internet household penetration[6]	No of local domains[7]	Online spending[8]	Internet Advertising spending[9]
Czech Rep.	10,288.9	73.6	769,186	0.65	50	29	1,502,537	7	22,734
Cyprus	776.0	88.9	89,073	0.54	33.6	37	75846	2	n.a.
Estonia	1,339.9	59.8	107,744	0.79	51.8	46	449,036	n.a.	3,607
Hungary	10,057.9	62.5	1,029,874	0.91	30.4	32	1,176,592	7	21,302
Latvia	3,385.7	48.6	129,961	0.81	45.2	42	132,204	1	7,277
Lithu-ania	2,280.5	52.1	150,07	0.50	35.9	35	240,592	2	3,086
Malta	407.7	71.7	n.a.	n.a.	33	53	20,673	n.a.	n.a.
Poland	3,8101.8	49.7	1,862,672	0.55	29.9	36	5,001,786	6	32,885
Slovakia	5,391.6	57.1	n.a.	n.a.	46.5	27	486,020	0	n.a.
Slovenia	2,010.3	81.9	242,656	0.64	55.5	54	64,284	9	5,484
EU 10	74,040.3	64.6			44	39	9,149,570		

Note: 1. Data in thousands for the 1st of January 2007. Source: Eurostat (2007)
2. GDP (in PPS per capita) in 2005. EU25=100%. Source: Eurostat (2007).
3. Global advertising expenditure 2006. In $US Thousands. Initiative Innovation (2007).
4. Initiative Innovation (2007) and The World Factbook, Central Intelligence Agency (2007)
5. Internet Usage in the European Union. Penetration (% Population) in 2007. Source: Internet World Stats (2007).
6. Percentage of households who have Internet access at home in 2006. Source: Eurostat (2007).
7. Number of local domains based on number of top-level domain in January 2007. Source: ISC Internet Domain Survey (2007).
8. The Internet turnover as percentage of the total turnover of enterprises with 10 or more employees in 2006. Source: Eurostat (2007).
9. Global advertising expenditure 2006. In thousands of $. Initiative Innovation (2007).

actual levels of ICT usage, and its purpose is to understand more thoroughly the impact of ICT on the competitiveness of nations. In this index, Estonia again scored best amongst the new members. Latvia and Poland had the lowest ratings.

STANDARDIZATION VS. LOCALIZATION

The issue of standardization arises from the desirability and feasibility of using a uniform marketing mix (4Ps) across national markets (Szymanski et al., 1993). Advertising has been examined more often than the other elements of this mix (Agrawal, 1995; Zandpour et al., 1994). A *standardized* approach is the use of uniform messages with no modification of headings, illustrations, or copy, except for translation in international markets (Onkvisit and Shaw, 1987). The standardized school of thought argues that consumers anywhere in the world are likely to share the same wants and needs (Elinder, 1961; Levitt, 1983). On the other hand, the *localized* approach asserts that consumers differ across countries, and therefore advertising should be tailored according to culture, media availability, product life cycle stages, and industry structure (Synodinos et al., 1989; Wind, 1986). Combining these two extremes, the third school of thought states that the appropriateness of standardization depends on economic similarity, market position, the nature of the product, the environment, and organizational factors (Jain, 1989).

In the 1970s, empirical evidence showed a high degree of localization, due to both increasing nationalistic forces, and various well-publicized advertising blunders in the 1960s (Agrawal, 1995). This trend reversed, to favour standardization, in the 1980s, and went along with a drastic rise in the number of multinational advertising agencies (Yin, 1999). During this period, a series of content analysis studies attempted to identify cross-cultural differences between Japanese and U.S. advertising (Hong et al., 1987; Madden et al., 1986; Mueller, 1987).

In the 1990s, localization seemed to remain popular among MNCs operating in various regions of world markets. Harris (1994) found that 69% of 38 MNCs (19 American and 19 European) standardized their advertising campaigns to some extent throughout the EC markets, whilst the rest of the sample localized. Interestingly, only 8% of the sample used totally standardized advertising, providing "little evidence of any widespread practice of standardized pan-European advertising campaigns" (Harris, 1994). Kanso and Nelson (2002) found that 62% of 193 firms (both American and non-American subsidiaries) in Finland and Sweden use localization, and place a strong emphasis on local cultures. Similarly, Samiee et al. (2003) found that MNCs operating in Southeast Asia tend to localize advertising. They examined 113 firms in Hong Kong, PRC, Taiwan, and Singapore, and found that both environmental and economic factors were the primary drivers of this tendency.

WEBSITE POSITIONING AS GLOBAL INFORMATION MANAGEMENT

Although these issues have been debated for decades in traditional media, a new stream of research has emerged recently, on the standardization versus localization of global websites in multiple markets. With the rapid expansion of the Internet, and the resulting connections between local, regional, and international markets, an increasing number of MNCs are shifting from off-line to on-line marketing. This frequently entails creating a diverse range of websites in multiple markets (Donthu and Garcia, 1999). By 2001, more than 36 million domains for commercial websites had already been established: these "dot coms" are projected to attract an astonishing $6.8 trillion in business by 2004 (Forrester, 2002; Internet Software Consortium, 2001).

Such numbers incline observers to see the Internet as a door to the "global village wonderland", as advocated by Levitt (1983): that is, an entity that creates an environment for more standardized marketing communication in world markets. Product-based websites are replacing such shopping venues as mail-order catalogues and television-based home shopping, and also offer a new format for global advertising among culturally and linguistically diverse groups (Pastor, 2001). An increase in the quantity and quality of product/brand information on the Internet is generating extraordinary consumer interest, which extends beyond physical and political boundaries (Donthu and Garcia, 1999). Accordingly, Roberts and Ko (2001) asserted that websites, with their ability to uniformly blend textual and visual content, constitute the best communication medium in which to develop brand images.

One roadblock that MNCs face involves localized websites: primarily, the need to satisfy the linguistic requirements of a diverse population of potential customers (Warden et al., 2002). According to Quelch and Klein (1996), establishing localized relationships with international consumers is best achieved by creating regional Web content. However, creating regional commercial websites may not be cost-effective if, to elicit return visits, a company is obliged to update information continuously. Such intense website maintenance on a regional level can jeopardize consistent brand strategies, by eliminating the "advantage of centralized management of a firm's Websites" (Warden et al., 2002).

In a pioneering study, Okazaki and Alonso (2002) examined Japanese websites in Japan,

Table 2. Network and e-readiness statistics

	Networked Readiness Index [1]	Networked Readiness Index (Rank) [1]	e-readiness rankings [2]	e-readiness rankings, general index [2]	e-readiness rankings, Connectivity index [2]	e-readiness rankings, Business Environment index [2]	Enterprises selling via Internet 2005 in % [3]	Enterprises availability of the Internet 2005 in % [3]
Czech Republic	0.36	32	32	6.14	4.90	7.39	13	92
Cyprus	0.36	33	n.a.	n.a.	n.a.	n.a.	4	85
Estonia	0.96	23	27	6.71	6.60	7.81	8	90
Hungary	0.27	37	32	6.14	4.80	7.34	4	78
Latvia	-0.03	52	39	5.30	3.95	7.21	1	75
Lithuania	0.08	44	38	5.45	4.65	7.28	6	86
Malta	0.51	30	n.a.	n.a.	n.a.	n.a.	16	90
Poland	-0.09	53	32	5.76	4.30	7.28	5	87
Slovakia	0.19	41	36	5.65	4.05	7.35	7	92
Slovenia	0.34	35	28	6.34	5.90	7.45	12	96

Note: n.a. = not available.
1. Global Information Technology Report 2005-2006
2. Economist Intelligence Unit (2006).
3. Eurostat (2005).

Spain, and the U.S.A., and found that cultural dimensions (power distance, uncertainty avoidance, masculinity – femininity, individualism-collectivism, and long-term orientation) and communication style (high context versus low context) were the primary drivers of cross-cultural differences in MNCs' website communication strategies. Focusing on more operational aspects, Okazaki (2005) examined American brands' website standardization in France, Germany, Spain, and the UK. He argued that the progress of the EU enlargement and economic integration via the euro provided firms with an incentive to use a uniform website communication across the EU member states. However, the findings were mixed, in that the level of standardization of American brands' websites in the European countries was low, compared to their respective home-country (American) websites. On the other hand, differences within the EU were minimal: the websites created within the European markets were somewhat "regionalized", especially for durable and industrial goods.A summary of prior research on website content analysis is shown in Table 3.

COMMUNICATION IN THE GLOBAL WEBSITE ENVIRONMENT

What is the primary factor influencing MNCs that operate in European markets? They now face more and more pressure to generate more comprehensive marketing strategies on the web. Among the various forms of the online environment, websites have been one of the most popular platforms, allowing consumers to see, consult, and obtain product-related information at any time, anywhere. Such websites can be seen as a new form of global marketing communications, offering opportunities to strengthen effective relational marketing in international markets (Robert and Ko, 2002). The creation of a localized URL in Europe, therefore, may be a necessary strategic move, because cultural and linguistic barriers are

Table 3. Prior research on website content analysis

Year	Authors	Countries examined	Unit of analysis	Analyzed content	Sample size	Statistical design
1999	Ju-Pak	US, UK & S.Korea	Product-based websites	Information content, creative strategies, design	110 (EE.UU.) 100 (UK) 100(S.Korea)	Chi-square, ANOVA
1999	Oh, Cho & Leckenby	US & S.Korea	Target ads	Information content, creative strategies	50 for each country	Chi-square
1999	Yoon & Cropp	US & S.Korea	Brand websites	Information content, emotional appeals, cultural aspects	20 for each country	Chi-square
2000	Lee & Sego	US & S.Korea	*Banners*	Information content, emotional appeals, colours, etc.	252 in total	Chi-square
2000	Chung & Ahn	US & S.Korea	*Banners*	Information content, "call-to-action" messages, demographics, etc.	251 (EE.UU.) 221 (S.Korea)	Chi-square
2000	Yoon	US	Product-based websites	Information content, celebrity endorsement, etc.	200 in total	Chi-square, ANOVA
2002	Okazaki & Alonso	Japan, Spain & EE.UU.	Product-based websites	Information content, cultural values, creative strategies	20 for each country	Chi-square, ANOVA
2002	Dou, Nielsen & Tan	Canada, Denmark & Malaysia	Commercial websites	Communication systems, transactional functions, etc.	150 for each country	ANOVA
2002	Zahir, Dobing & Hunter	26 countries	National portals	Linguistic aspects, design, colours, Hofstede's cultural dimensions, etc.	26 portals	Descriptive stat
2003	Robbins & Stylianou	16 countries	Corporate websites	Design, presentation, links, security, information content, financial content, corporate information, etc.	90 in total	ANOVA
2005	Okazaki	US, UK, France, Germany & Spain	Brand websites	Brand website functions, similarity ratings	244	ANOVA, discriminant analysis, multiple regression

perhaps the most difficult obstacles to overcome in marketing communications across European nations (Kahle, Beatty, and Mager, 1994). Such localization, however, could cost a great deal. Hence, MNCs may intuitively favour standardization, given the benefits associated with offline marketing standardization, such as consistent brand image and corporate identity, cost savings, and organizational control. Furthermore, websites seem to be an effective medium for establishing a global brand image, by offering consistent textual and visual information to international consumers. Unfortunately, there seems to be a lack of empirical research regarding the standardization versus localization issue in the online environment, leaving important questions unanswered.

What are the determining factors in international marketing communications on the web? In a recent criticism of the slow progress of international advertising research, Taylor and Johnson (2001) argue that the standardization debate should "focus on what executions can be standardized and when they can be standardized". Following this suggestion, this study intends to fill this gap, by identifying to what extent MNCs have adopted a standardized approach for their websites created in European markets. In order to ensure cross-national data equivalency, we examined only the websites created by America's top brands for the UK, France, Germany, and Spain. These countries differ importantly in terms of cultural and linguistic characteristics, but are relatively homogeneous in socioeconomic conditions and technological infrastructure, and have online markets of a reasonable size.

Furthermore, 3.3%, 6.5%, and 8.1% of the world's online population consist of French, German, and Spanish speaking consumers, respectively, compared to 35.2% of English speakers (Global Reach, 2003). Therefore, these four countries represent an important segment of world online consumers. On the other hand, the languages spoken in the new EU member states, such as Polish or Czech, account for a very small portion of the online population. In fact, the impact of these countries, on both the world economy and the world online population, is negligible (Table 4). Thus, an important question arises: is it worthwhile for MNCs to consider local adaptation in such new markets? Or is it better to use a standardization approach in these markets, to take advantage of cost savings and efficient website maintenance? To address these questions, this study will examine websites created by MNCs for the Czech Republic.

CONCEPTUAL FRAMEWORK

Figure 1 shows the conceptual framework for this study. These concepts are essentially based on the matrix proposed by Quelch and Klein (1996), who suggested two primary models of website: the communication model and the transaction model. Originally, their matrix was not intended to be a theoretical model for formal testing, but since then it has been used as a conceptual base (e.g., see Dou, Nielsen, and Tan, 2002). In our modified matrix, communication and transaction feature form two ends of one axis, which should be balanced with the other axis, consisting of fact and image. The resulting four quadrants need to be effectively combined to achieve the desired level of website standardization. The components in each quadrant can be considered the most relevant programmes for website brand communications.

The extent of website standardization should be determined on the basis of the two major roles of global online programmes: (1) to enhance worldwide transactions by establishing a localized relationship, and (2) to develop a standardized brand image, using the appropriate combination of content, graphics, backgrounds, and multimedia effects in all the MNC's websites in different languages (Roberts and Ko, 2003). In the following section, each principal feature of our proposed model is therefore analyzed in the light of these perspectives.

METHODOLOGY

This study adopts content analysis as a research methodology. This method has been widely used in cross-cultural research (Brislin, 1980), as well as in the Internet research (McMillan, 2000; Okazaki and Alonso, 2002).

Company selection. A website content analysis was performed, to examine the degree of website standardization of American brands' websites created for the Czech market. Methodological recommendations from prior research were adopted, to establish a high reliability (Dou et al., 2002; Okazaki and Alonso, 2002; Philport and Arbittier, 1997). To create a dataset, a ranking of "The 100 Top Global Brands" from *BusinessWeek (2002)* was used. Only brands with America as country of origin (by the classification of *BusinessWeek*) were chosen to match home versus host country website pairs. In total, 66 brands were found, of which 34 brands had websites in the Czech Republic. Here, it is important to note that these

Table 4. World online population and language use

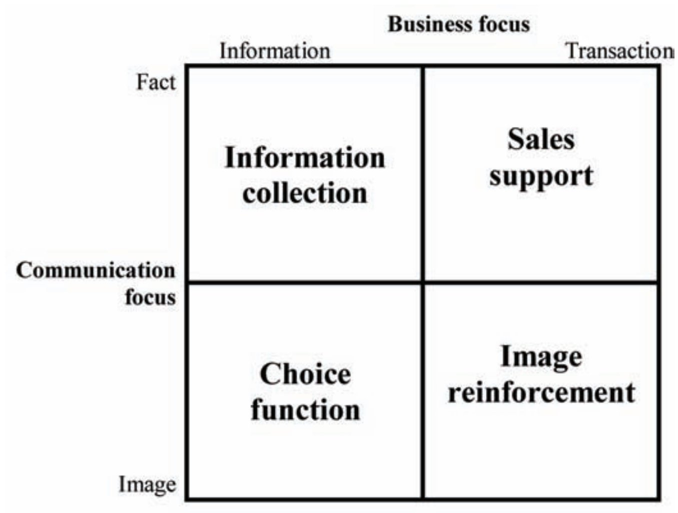
Language type	**Internet access** [1]	**% of world online pop.**	Total pop.[1]	GDP[2]	**% of world economy**	GDP **per capita** [3]
English	238.5	35.2	508	n.a.	n.a.	n.a.
Non-English	439.8	64.8	5,822	n.a.	n.a.	n.a.
European Languages (non-English)	238.1	35.1	1,218	12,968	30.3	n.a.
Czech	4.0	n.a.	12	121	n.a.	10.0
Dutch	13.2	2.0	20	575	n.a.	28.5
Finnish	2.8	n.a.	6	142	n.a.	23.6
French	22.7	3.3	77	1517	4.2	19.7
German	44.4	6.5	100	2,679	5.8	26.8
Greek	2.0	n.a.	12	189	n.a.	15.8
Hungarian	1.6	n.a.	10	96	n.a.	9.6
Italian	24.1	3.6	62	1,251	3.6	20.1
Polish	6.9	n.a.	44	359	n.a.	8.1
Portuguese	19.3	2.8	176	1,487	3.6	8.4
Romanian	2.4	n.a.	26	108	n.a.	4.2
Russian	18.4	2.7	167	822	1.8	4.9
Scandinavian languages (total)	13.5	2.0	20	550	1.3	27.9
Serbo-Croatian	1.0	n.a.	20	n.a.	n.a.	n.a.
Slovak	1.0	n.a.	6	47	n.a.	8.7
Slovenian	0.7	n.a.	2	22.9	n.a.	10.9
Spanish	54.8	8.1	350	2500	8.9	7.1
Turkish	4.6	n.a.	67	431	n.a.	6.4
Ukrainian	0.9	n.a.	47	115	n.a.	2.3
TOTAL EUROPEAN LANGUAGES (non-English)	238.1	35.1	1,218	12,968	33.9	n.a.
TOTAL ASIAN LANGUAGES	201.7	29.7	n.a.	n.a.	n.a.	n.a.
TOTAL WORLD	648.7		6,330	41,400	n.a.	n.a.

Note: 1 US$ in million; 2 US$ in billion, 3 US$ in Thousand.
Source: Global Reach (2003)

firms are considered as representative of American firms doing business in the Czech Republic, because their Internet presence can be considered as the initial step of market entry mode.

Coding categories. Next, a detailed coding sheet was first developed, with detailed operational definitions. The variables include (1) brand website functions, and (2) similarity ratings (Table 5). With regard to the former, the coding categories were adopted from Okazaki (2005), who suggested 23 website communication functions. Similar categories have been suggested in the past (e.g., Ghose and Dou, 1998; Leogn, Huang, and Stanners, 1998). Each variable was measured on a categorical dichotomy as to the existence of each function on a website. Values of “1” and “0” were assigned for answers of “Yes” and “No”, respectively. For example, websites that had appropriate information associated with “job/career development” were assigned “1” for this

Figure 1. Conceptualization of website program standardization



attribute. Those that did not were assigned "0". These values were considered dependent variables in the analysis. The similarity rating refers to the degree of similarity between home-country and host-country websites. This "similarity rating" measure was also adopted from Okazaki (2005), and was originally inspired by Mueller (1991). The textual and visual components of websites created for local markets were assessed for the extent to which they were similar or dissimilar to those created at home. The similarity rating was coded for each pair of websites (i.e., U.S.A.-Czech sites) on a five-point semantic scale, ranging from "very different" (coded as 1) to "very similar" (coded as 5), with an intermediate scale point "not determinable" (coded as 3). The components included copy, headlines, text, layout, colour, photographs associated with the product, with human models, and with background scenes, illustrations, charts, graphs, and interactive images.

Coding instrument. All coding instruments were originally prepared in English, and then translated into Czech, using the "back translation" technique. Each typology was supplemented with additional examples, to give a better illustration. The unit of analysis was determined as the homepage, which has been considered a central gate to Web-based communication. This is appropriate, given the primary objective of the study: to identify major differences in the main text, pictures, and graphics. We examined the first and second levels of websites, because it is practically impossible to scrutinize every detail of an entire site. The existence of online brand communications was primarily determined by the main menu or index provided on the homepage. For instance, if the menu included a link labelled "corporate information", the site was coded as having this variable. The only exception occurred when analyzing direct or indirect online transactions, because in some cases these functions may not be listed on the main index. In this case, the coders were asked to examine the submenu of the websites.

Coder training and reliability. Following the recommendations by Kolbe and Burnett (1991), two bilingual Czech judges, both of whom were unaware of the purpose of the study, were hired and first trained to grasp the operational definitions of all the variables. During the training period, the coders practised independently, by examining 20 randomly chosen websites from non-American firms. Then, the coded results were compared,

Table 5. Measurement schemes

Measure	Coding categories	Scale type
Brand website functions	Global/Local site options, Corporate information, Corporate news release, Product/Brand news release, General product information, Brand specific information, Investor relationships, Direct online transaction, Indirect online transaction, Office/Store locator, Country/Language option, Search engine, Jobs/Career development, Promotion/Prizes/Sweepstakes, Education/Training, Culture/Entertainment, Client registration/Log-in, Guest book/Customer feedback, E-mail alert, FAQs, Free download, Sitemap, Links	Nominal scale (Yes=1, No=0)
Similarity ratings	Company logo, Company logo placement, Major copy, Major copy placement, Major headline, Major headline placement, Major text, Layout in top half / right half, Layout in bottom half / left half, Colour in top half / right half, Colour in bottom half / left half, Major photograph 1 (product), Major photograph 2 (human model), Major photograph 3 (background scene), Major illustrations, Major chart or graph, Interactive image 1 (flash as opening), Interactive image 2 (pop-ups), Interactive image 3 (animated banners), Interactive image 4 (layers, pop-unders, etc.)	Interval scale (1=very different, 5=very similar)

and differences were reconciled through discussion. An inter-judge reliability was calculated using the reliability index suggested by Perreault and Leigh's reliability index (I_r) (1989). Various researchers consider this estimation method to be the best among available alternatives (e.g., Kolbe and Burnett, 1991).

As Table 6 shows, the majority of the resulting indexes far exceeded a widely accepted minimum .80???, and was thus deemed satisfactory. It was recognized that there would be a potential loss of information in similarity evaluation between American and Czech sites, because non-native English speakers had analyzed American websites. However, it was accepted that such potential bias was minimized by the coders' extensive preparation: the subjective interpretation of textual information was minimal, since the coders were responsible for examining only *major* copy, headlines, and text on the websites. Otherwise, they were instructed to objectively measure the similarity of non-textual information.

RESULTS

Table 7 summarizes the frequency distribution of brand website functions. For the sake of comparison, the information provided in Okazaki's (2005) previous exploration was used as a reference with regard to the U.S., UK, French, and German markets. This comparison should help our understanding of MNCs' website standardization in existing versus new EU member states.

The Chi-square analysis detected significant differences in 8 categories: global/local site options, general product information, investor relations, online purchase, email contact, promotion/prizes/sweepstakes, culture/entertainment, and guest book/customer feedback. It appears that American MNCs tend to apply a different website communication strategy in the Czech market because, in prior research, Okazaki (2005) found significant differences in only 3 of 23 variables, suggesting that the frequency of the usage of brand website functions in the UK, France, Germany, and Spain was relatively uniform. In observing the frequencies of brand website functions in the Czech sites, email contact was used more frequently, but the other tools were used much less than in the other markets.

Next, in order to capture the relationships between the brand website functions and country domain, we performed a multiple correspondence analysis via optimal scaling technique. This method is appropriate for nominal variables, from which a multidimensional map can be created. We used the existence of the brand website functions

Table 6. Intercoder reliability

Measure	**Coding categories**	**Perreault & Leigh's *Ir***
Brand website functions	Brand specific information	0.82
	Client registration/Log-in	0.91
	Corporate information	1.00
	Corporate news release	0.97
	Country/Language option	1.00
	Culture/Entertainment	0.97
	Direct online transaction	1.00
	Education/Training	0.94
	E-mail alert	1.00
	FAQs	1.00
	Free download	0.94
	General product information	0.97
	Global/Local site options	0.94
	Guest book/Customer feedback	1.00
	Indirect online transaction	1.00
	Investor relations	0.97
	Jobs/Career development	0.97
	Links	0.91
	Office/Store locator	1.00
	Product/Brand news release	1.00
	Promotion/Prizes/Sweepstakes	0.97
	Search engine	1.00
	Sitemap	1.00
Similarity ratings	Text	0.91
	Major photograph: product	0.75
	Major photograph: model	0.56
	Major photograph: background scene	0.92
	Major illustrations	0.82
	Major chart or graph	
	Layout in top half / right half	0.93
	Layout in bottom half / left half	0.98
	Headline placement	0.74
	Headline	0.69
	Copy placement	0.73
	Copy	0.91
	Company logo placement	0.95
	Company logo	0.96
	Colour in top half / right half	0.93
	Colour in bottom half / left half	0.91

Table 7. Results of brand website functions

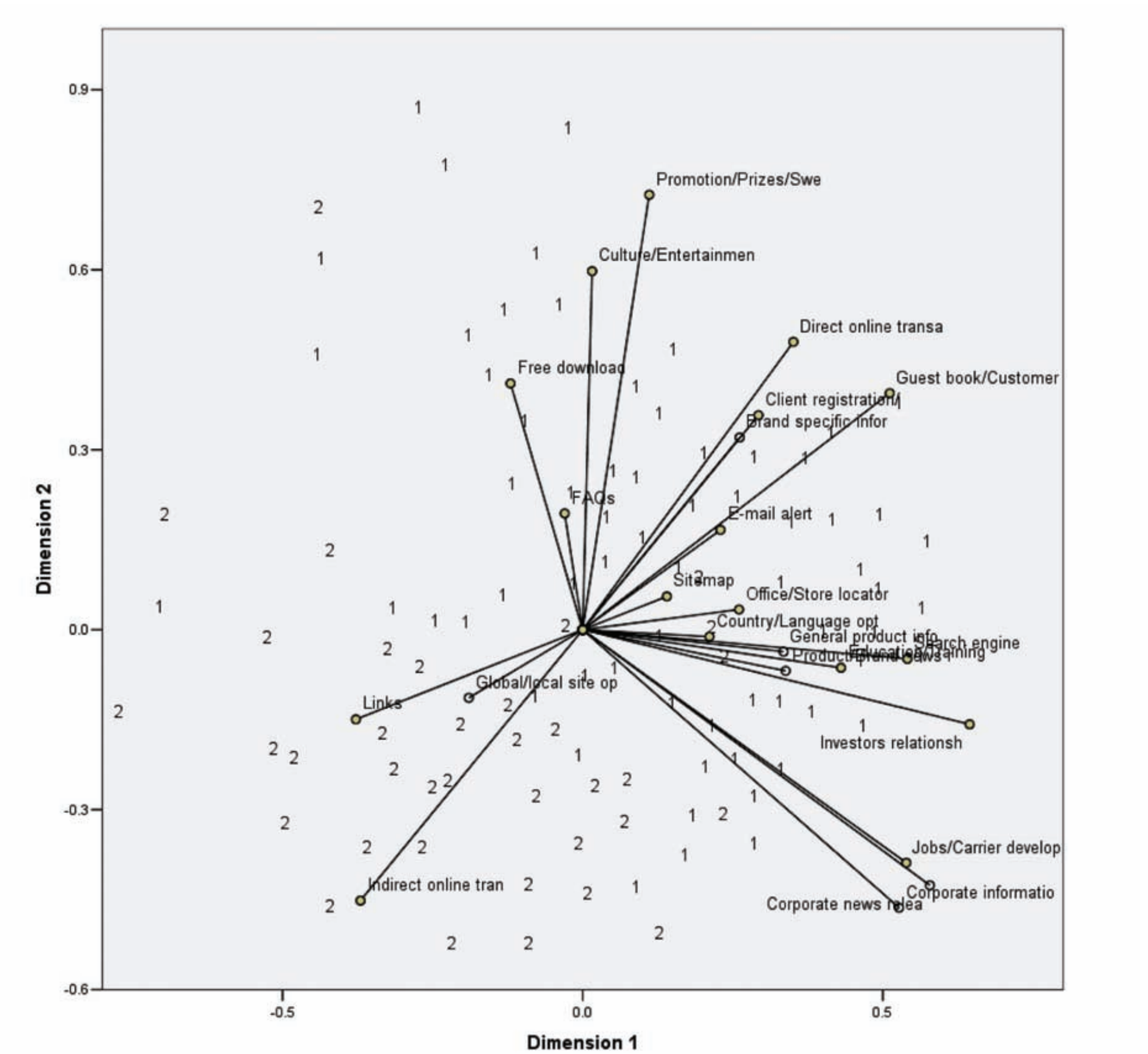
1. Brand website features	US (n=66)	UK (n=57)	France (n=49)	Germany (n=57)	Czech (n=34)	*p*
Global/Local site options	37.9	84.2	65.3	75.4	**64.4**	.000
Corporate information	89.4	86.0	87.8	84.2	**86.1**	.956
Corporate news release	53.0	54.4	55.1	59.6	**55.6**	.983
Product/Brand news release	51.5	49.1	53.1	52.6	**55.6**	.990
General product information	80.3	84.2	83.7	78.9	**47.2**	.000
Brand specific information	75.8	68.4	73.5	68.4	**69.4**	.858
Investor relationships	45.5	26.3	16.3	22.8	**11.1**	.001
Online purchase	71.2	42.1	42.6	43.9	**25.0**	.000
Email contact	22.7	31.6	28.6	24.6	**77.8**	.000
Office/Store locator	33.3	33.3	32.7	26.3	**13.9**	.346
Country/Language option	62.1	57.9	71.4	61.4	**52.8**	.536
Search engine	68.2	57.9	55.1	50.9	**52.8**	.442
Jobs/Career development	62.1	47.4	46.9	54.4	**61.1**	.374
Promotion/Prizes/Sweepstakes	56.1	63.2	44.9	47.4	**11.1**	.000
Education/Training	39.4	33.3	26.5	24.6	**25.0**	.331
Culture/Entertainment	47.0	57.9	53.1	42.1	**13.9**	.001
Client registration/Log-in	51.5	36.8	38.8	36.8	**41.7**	.398
Guest book/Customer feedback	78.8	82.5	75.5	77.2	**22.2**	.000
E-mail alert	25.8	15.8	20.4	19.3	**25.0**	.483
FAQs	18.2	22.8	16.3	19.3	**8.3**	.647
Free download	19.7	26.3	26.5	28.1	**36.1**	.643
Sitemap	45.5	42.1	44.9	36.8	**38.9**	.905
Links	4.5	12.3	8.2	3.5	**19.4**	.053

Note: The data of the US, UK, France and Germany are based on Okazaki (2005)

(yes or no) as descriptive variables, and the type of country domain (U.S.A. or Czech Republic) as classification variables. Figure 2 shows the resulting biplot component loadings. As clearly seen, the majority of brand website functions are more closely associated with U.S. sites (represented by "1"), while only a limited number of brand website functions are associated with Czech sites (represented by "2"). Specifically, global/local site options, links, and indirect online transactions are concentrated in the lower left quadrant (which U.S. sites appear to dominate), but the rest of the brand website functions are concentrated in the upper right quadrant (where Czech sites appear to dominate).

Lastly, Table 8 summarizes the results of the similarity ratings. A higher similarity rating indicates a higher degree of standardization. As the results clearly show, the similarity between the American and Czech sites was notably higher, especially in logo, copy, and colour. In comparison with the other sites, headlines and major photographs also exhibit higher similarity. On this basis, it appears clear that the American MNCs tend to create highly standardized websites in the Czech Republic.

Figure 2. Multiple correspondence analysis



DISCUSSION

This study attempts to examine the website communication strategy used by American MNCs in a new EU member state, the Czech Republic. We performed a content analysis of 34 Czech sites created by America's top brands in terms of brand website functions, and similarity between the home and host sites. The findings indicate that American MNCs appear to standardize their Czech sites. Given that the Czech Republic is a growing market that attracts more and more foreign direct investment, this case could be considered indicative of the general tendency in the other new EU member states.

First, American MNCs tend to use general product information less frequently in the Czech market than in the other EU markets, which suggests two possible scenarios. First, they have not yet commercialized their products in this market, and therefore dispose of little information. In this case, the primary objective of their websites would

Table 8. Results of similarity ratings

Components	UK (n=57)	France (n=49)	Germany (n=57)	Czech (n=34)
Company logo	4.51	4.61	4.39	**4.58**
Major copy	1.43	1.36	1.14	**4.00**
Major headline	1.23	1.17	1.02	**3.47**
Major text	2.84	2.53	2.68	**2.32**
Layout	3.35	3.32	3.29	**3.88**

be to provide a preliminary information platform in a new market. Second, they might have needed to localize product information to a great extent, especially in the local language, with more adapted product usage. However, it appears that the most logical conclusion would be that the lack of online product information is because American firms are still in the very early mode of market entry in the Czech Republic.

The second scenario seems very unlikely when we address the following question: Why would large multinational firms devote resources to extensive adaptations of Czech websites? The country's population is only 10 million and, according to the most recent Eurostat (2006), only 19% of Czech homes had the Internet access in 2005, including 5% with broadband connections. As much as 63% of the population has never used the Internet. Only one in six people who used the Internet (5.5%) bought anything online in 2005, and these purchases were limited largely to electronic goods (2.1%), books (1.6%), and clothing (1.1%). Therefore, if the total market in a given product category is currently only 100,000 or so, and if adapting the website is only going to improve website effectiveness by 5% to 10%, is there any incentive to adapt and then to continue managing that adaptation? Consistent with this argument, our findings indicate that online purchase functions are rarely used in the Czech sites. This suggests that American MNCs may have neither distribution channels nor local investors in the Czech market. Similarly, the much less frequent use of guest book/customer feedback indicates that American brands are less willing to offer personalized contact to the local Czech consumers, probably because of the unavailability of local outlets, representatives, or staff. In contrast, in the Czech Republic, they offer general email addresses more frequently than in the other countries, as an alternative contact mode for general inquiries.

Third, by the same token, culture/entertainment and promotion/prices/sweepstakes are used much less in the Czech market than in the others, because these elements need to be matched to local consumer tastes, and require more personalized content. It would make little sense to offer presents or incentives when the companies actually have no local sales activities.

Finally, a surprisingly high level of similarity ratings for both textual and visual components indicates a lack of any cultural adaptation of websites to the Czech market. This contrasts with Okazaki's (2005) findings regarding American MNCs' website strategy in the UK, France, Germany, and Spain, where clearly localized websites have been created in the existing EU member states. This finding is consistent with the frequency of brand website functions: the Czech sites use far fewer brand website functions with highly standardized textual and visual components.

It is clear that American MNCs consider the EU a single market, and one that is strategically dissimilar to their home market. If we observe only the websites created in the "older" member states, there seems to exist a "regionalization" strategy across Europe, in that the level of similarity ratings among the European samples was relatively uniform. This may be due to the close geographical proximity of the three countries, which would, logically, provide more opportunities for personal interaction and the accumulation of greater knowledge. However, in the case of the Czech Republic, website adaptation has not yet advanced, probably due to many unknown factors: in particular, specific information regarding local consumers' tastes and preferences.

To make our findings more objective, we should recognize a few limitations. First, the current study examined only one country that has recently joined the European Union. Future research should expand this study into other new EU member states, especially Poland and Hungary, because these two countries, along with the Czech Republic, are the most economically developed regions. Because of the extreme scarcity of research related to these countries, any such extension will contribute significantly to the literature. Second, while content analysis could provide useful information regarding the manifest content, the findings should be treated with caution. The findings by no means explain practitioners' "true" intentions in website communication strategy. In this regard, it will be interesting to extend this study in the future, by conducting a questionnaire survey of advertisers and marketers who are actually responsible for the new EU member states.

LIMITATIONS

While this study makes significant contributions to the global information management literature, some important limitations must also be recognized. First, content analysis is, by definition, an observational method that examines only manifest content. Our findings have little or nothing to do with marketers' "true" intentions on global website positioning. Second, our unit of analysis was limited to the menu and submenu of the homepages. However, it is possible that a more localized strategy might have been observed in further links. Finally, we examined only one of the new EU member states, thus, any generalization of our findings should be treated with caution.

FUTURE RESEARCH DIRECTIONS

First, future extensions should examine websites created for the other new EU member states, such as Poland, Hungary, Latvia, Lithuania, and Malta. Information technology management for website positioning in these countries is virtually unknown, and analyzing brands' websites positioning in these countries should help us to draw more generalisable implications. In particular, researchers are planning to examine Polish and Hungarian websites in the next stage because, in these countries, the total online as well as offline advertising spending is substantial, in comparison with the other new EU member sates.

Second, in furthering our explorations, content analysis methodology should be improved. Specifically, we need to examine the level of standardization or localization at deeper levels of websites. While this study scrutinized the first level of websites or homepages, some may claim that much information was lost by ignoring the second and third levels of websites. For example, the lack of direct online transactions need not necessarily mean that the website does not have a link to the online shopping sites of different companies. This was the case for consumer electronics, in that computer or office machine products are sold on "general" e-commerce (or even auction) sites. More specific coding instructions should be established, to enable the coders to improve their analysis with a higher level of inter-coder reliability.

Third, we also should conduct a survey that targets foreign subsidiaries' managers. It will be interesting to compare the findings of this paper with the managers' perceptions. In particular, several questions appear of special interest. For example, are their websites created or controlled locally or globally? What level of control do senior executives of foreign subsidiaries actually have of their electronic commerce planning and executions?

Finally, in an attempt to capture a clearer picture of global website positioning in multiple markets, more collaboration will be needed by researchers in information systems management and other disciplines: in particular, marketing management. Needless to say, a higher level of international cooperation is necessary to conduct more objective and reliable data collection in multiple markets.

REFERENCES

Agora (2005). Advertising market.

Agrawal, M. (1995). Review of a 40-year debate in international advertising. *International Marketing Review*, *12*(1), 26–48. doi:10.1108/02651339510080089

ARBOmedia. (2005a). Standardní reklamní rok přinese médiím 17,8 miliard.

ARBOmedia. (2005b). Standardní reklamní rok přinese médiím 17,8 miliard".

Brislin, R. W. (1980). Translation and Content Analysis of Oral and Written Material. In H. C. Triandis and Berry, J. W. (eds.), *Handbook of Cross-cultural Psychology* (Vol. 2, pp. 389-444), Boston: Allyn and Bacon.

Budapest business journal (2003). Web-based advertising making ground. September 16, 2003,

Budapest business journal (2004). Online advertising boom. August 18, 2004,

BusinessWeek (2002). The 100 Top Brands. August 5, 95-99.

Czech, O. M. D. (2005). Odhady reklamních výdajů.

Czech Publishers Association. (2006). Internetová reklama v roce 2005 přesáhla 1 miliardu Kč. Press release,

Czech Statistical Office (2006), *The survey on ICT usage in households and by individuals in the Czech Republic in 2006* (in Czech).

Donthu, N., & Garcia, A. (1999). The Internet Shopper. *Journal of Advertising Research*, *39*(5), 52–58.

Dou, W., Nielsen, U. O., & Tan, C. M. (2002). Using corporate websites for exporting marketing. *Journal of Advertising Research*, *42*(5), 105–115.

Economist Intelligence Unit (2006), *The 2006 e-readiness rankings*.

Elinder, E. (1961). How International Can Advertising Be?" in S. W. Dunn (ed.), *International Handbook of Advertising*, NY: McGraw-Hill, 59-71.

Eurostat (2005), *Europe in figures: Eurostat yearbook 2005*.

Eurostat (2006), *Community survey on ICT usage in households and by individuals.*

Factum, T. N. S. (2004). Firemní weby a marketing.

Factum, T. N. S. (2005). Češi a reklama v roce 2004.

Forrester (2002), *Forrester's Global eCommerce Predictions For 2004*.

GfK. (2006). Increasingly more kids use the Internet. One in three Polish Internet users willing to buy music on the Internet. Press release, Ghose, S., & Dou, W. (1998). Interactive Functions and Their Impacts on the Appeal of Internet Presence Sites. *Journal of Advertising Research*, *38*(2), 29–43.

Harris, G. (1994). International Advertising Standardization: What Do the Multinationals Actually Standardize? *Journal of International Marketing*, *2*(4), 13–30.

Hong, J. W., Muderrisoglu, A., & Zinkhan, G. M. (1987). Cultural differences and advertising expression: a comparative content analysis of Japanese and U.S. magazine advertising. *Journal of Advertising*, *16*(1), 55–62, 68.

Internet Software Consortium. (2001), *Distribution of Top-Level Domain Names by Host Count, Jul 2002*.

Jain, S. C. (1989). Standardization of International Marketing Strategy: Some Research Hypotheses. *Journal of Marketing*, *53*(January), 70–79. doi:10.2307/1251525

Ju-Pak, K. H. (1999). Content dimensions of Web advertising: a cross-national comparison. *International Journal of Advertising*, *18*(2), 207–231.

Kanso, A., & Nelson, R. A. (2002). Advertising Localization Overshadows Standardization. *Journal of Advertising Research*, *42*(1), 79–89.

Kolbe, R. H., & Burnett, M. S. (1991). Content-Analysis Research: An Examination of Applications with Directives for Improving Research Reliability and Objectivity . *The Journal of Consumer Research*, *18*(September), 243–250. doi:10.1086/209256

Koudelova, R., & Whitelock, J. (2001). A cross-cultural analysis of television advertising in the UK and the Czech Republic. *International Marketing Review*, *18*(3), 286–300. doi:10.1108/02651330110695611

Leong, E. K. F., Huang, X., & Stanners, P. J. (1998). Comparing the Effectiveness of the Web site with Traditional Media. *Journal of Advertising Research*, *38*(5), 44–51.

Levitt, T. (1983). The Globalization of Market. *Harvard Business Review*, *61*(May/June), 92–102.

Mackay, S. (2004). Parliament Approves New Government. *Executive Perspectives: Poland*, *Price Waterhouse Coopers.*

Madden, C. S., Caballero, M. J., & Matsukubo, S. (1986). Analysis of information content in U.S. and Japanese magazine advertising. *Journal of Advertising*, *15*(3), 38–45.

McMillan, S. J. (2000). The Microscope and The Moving Target: The Challenge of Applying Content Analysis to The World Wide Web. *Journalism & Mass Communication Quarterly*, *77*(1), 80–98.

Mueller, B. (1991). An analysis of information content in standardized vs. specialized multinational advertisements. *Journal of International Business Studies*, First Quarter, 23-39.

OBP. (2001). Advertising in Poland.

Oh, K. W., Cho, C. H., & Leckenby, J. D. (1999). A Comparative Analysis of Korean and U.S. Web Advertising. *Proceedings of the 1999 Conference of the American Academy of Advertising*, Gainesville: University of Florida, 73-77.

Okazaki, S. (2005). Searching the Web for global brands: How American brands standardise their websites in Europe. *European Journal of Marketing*, *39*(1/2), 87–109. doi:10.1108/03090560510572034

Okazaki, S., & Alonso, J. (2002). A content analysis of Web communication strategies: Cross-cultural research framework and pre-testing. *Internet Research: Electronic Networking . Applications and Policy*, *12*(5), 380–390.

Onkvisit, S., & Shaw, J. J. (1987). Standardized International Advertising: A Review and Critical Evaluation of the Theoretical and Empirical Evidence. *The Columbia Journal of World Business*, *22*(Fall), 43–55.

Optimedia, Z. (2004). Poland's advertising market in 2004.

Perreault, W. D., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *JMR, Journal of Marketing Research*, *26*(May), 135–148. doi:10.2307/3172601

Philport, J. C., & Arbittier, J. (1997). Advertising: Brand communications styles in established media and the Internet. *Journal of Advertising Research*, *37*(2), 68–76.

Quelch, J. A., & Klein, L. R. (1996). The Internet and International Marketing. *Sloan Management Review*, *38*(Spring), 60–75.

Roberts, M. S., & Ko, H. (2001). Global Interactive Advertising: Defining What We Mean and Using What We Have Learned. *Journal of Interactive Advertising*, 1 (2),

Root, F. R. (1994), *Entry Strategies for International Markets*. Heath, Washington DC: Lexington.

Sheram, K., & Soubbotina, T. P. (2000), *Beyond Economic Growth: Meeting the Challenges of Global Development*, Washington D.C.: World Bank. Taylor, Charles R., P. Greg Bonner and Michael Dolezal (2002). Advertising in the Czech Republic: Czech perceptions of effective advertising and advertising clutter. In Charles R. Taylor (ed.), *New direction in International Advertising Research* (Vol 12, pp. 137-149), San Diego, CA: Elsevier.

Synodinos, N., Keown, C., & Jacobs, L. (1989). Transitional Advertising Practices: A Survey of Leading Brand Advertisers. *Journal of Advertising Research*, *29*(2), 43–50.

Szymanski, D. M., Bharadwaj, S. G., & Varadarajan, P. R. (1993). Standardization versus Adaptation of International Marketing Strategy: An Empirical Investigation. *Journal of Marketing*, *57*(October), 1–17.

Taylor, C. R., & Johnson, C. M. (2002). Standardized vs. specialized international advertising campaigns: What we have learned from academic research in the 1990s. *New Directions in International Advertising Research*, *12*, 45–66. doi:10.1016/S1474-7979(02)12019-9

Unie vydavatelů (2005). Internetová reklama v roce 2004 dosáhla 760 milionů Kč.

van REPEN, Erica, Rik Pieters, Jana Fidrmucova and Peter Roosenboom. (2000). The Information Content of Magazine Advertising in Market and Transition Economies. *Journal of Consumer Policy*, *23*, 257–283. doi:10.1023/A:1007138219753

West, D. C., & Paliwoda, S. J. (1996). Advertising adoption in a developing market economy: The case of Poland. *International Marketing Review*, *13*(4), 82–101. doi:10.1108/02651339610127266

World Advertising Resource Center (WARC) (2004). *European Marketing Pocket Book.*

Zandpour, F., Campos, V., Catalano, J., Chang, C., Cho, Y. D., & Hoobyar, R. (1994). Global Reach and Local Touch: Achieving Cultural Fitness in TV Advertising. *Journal of Advertising Research*, *34*(5), 35–63.

ADDITIONAL READING

Batra, R., Myers, J. G., & Aaker, D. A. (1996). *Advertising Management,* Prentice Hall, Englewood Cliffs, NJ.

De Mooij, M. (1998). *Global marketing and advertising: Understanding cultural paradox*. Sage Publications, Thousand Oaks, CA.

Duncan, T., & Ramaprasad, J. (1995). Standardized multinational advertising: The influencing factors. *Journal of Advertising*, *24*(3), 55–67.

Ghose, S., & Dou, W. (1998). Interactive functions and their impacts on the appeal of Internet presence sites. *Journal of Advertising Research*, *38*(3), 29–43.

Ha, L., & James, E. L. (1998). Interactivity reexamined: A baseline analysis of early business Web sites. *Journal of Broadcasting & Electronic Media*, *42*(4), 457–469.

Hwang, J. S., McMillan, S. J., & Lee, G. (2003). Corporate Web sites as advertising: An analysis of function, audience, and message strategy. *Journal of Interactive Advertising*, *3*(2).

Jain, S. (1989). Standardization of international marketing strategy: Some research hypotheses. *Journal of Marketing*, *53*(1), 70–79. doi:10.2307/1251525

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Sage Publications, Newbury Park, CA.

Laroche, M., Kirpalani, V. H., Pens, F., & Zhou, L. (2001). A model of advertising standardization in multinational corporations. *Journal of International Business Studies*, *32*(2), 250–265. doi:10.1057/palgrave.jibs.8490951

Lynch, P. D., Kent, R. J., & Srinivasan, S. S. (2001). The global Internet shopper: Evidence from shopping tasks in twelve countries. *Journal of Advertising Research*, *41*(3), 15–22.

Okazaki, S. (2004). Do multinationals standardise or localise? The cross-cultural dimensionality of product-based Web sites. *Internet Research: Electronic Networking . Applications and Policy*, *14*(1), 81–94.

Okazaki, S., & Alonso, J. (2003). Right messages at the right site: on-line creative strategies by Japanese multinational corporations. *Journal of Marketing Communications*, *9*(4), 221–239. doi:10.1080/1352726032000129908

Okazaki, S., & Alonso, J. (2003). Beyond the Net: Cultural Values Reflected in Japanese Multinationals' Web Communication Strategies. *Journal of International Consumer Marketing*, *16*(1), 47–70. doi:10.1300/J046v16n01_04

Tharp, M., & Jeong, J. (2001). Executive insights: The global network communications agency. *Journal of International Marketing*, *9*(4), 111–131. doi:10.1509/jimk.9.4.111.19939

Warden, C. A., Lai, M., & Wu, W. Y. (2003). How world-wide is marketing communication on the World Wide Web? *Journal of Advertising Research*, *43*(5), 72–84.

This work was previously published in Encyclopedia of Data Warehousing and Mining, edited by J. Wang, pp. 885-890, copyright 2005 by Information Science Reference (an imprint of IGI Global).

Chapter 5.4
WEB 2.0, Social Marketing Strategies and Distribution Channels for City Destinations:
Enhancing the Participatory Role of Travelers and Exploiting their Collective Intelligence

Marianna Sigala
University of the Aegean, Greece

ABSTRACT

During the last decades, the use of Web 2.0 applications for the generation, dissemination, and sharing of user-generated content (UGC) and the creation of new value added services are enormous. Web 2.0 tools have tremendously changed the way people search, find, read, gather, share, develop, and consume information, as well as on the way people communicate with each other and collaboratively create new knowledge. UGC and Web 2.0 are also having a tremendous impact not only on the behaviour and decision-making of Internet users, but also on the e-business model that organizations need to develop and/or adapt in order to conduct business on the Internet. Organizations responsible to market and promote cities on the Internet are not an exception from these developments. This chapter aims to inform city tourism organizations responsible for the development of city portals about (a) the use of the major Web 2.0 tools in tourism and their impact on the tourism demand and supply; and (b) the ways and practices for integrating the use of Web 2.0 into their e-business model and e-marketing practices.

INTRODUCTION

During the last years, the number and use of numerous Web 2.0 tools, whereby Internet users produce, read and share multimedia content (User Generated Content, UGC), is mushrooming (eMarketer, 2007a). It is estimated (eMarketer, 2007b) that 75.2 million USA Internet users currently use UGC, and this is expected to increase to 101 million by 2011. eMarketer (2007c) also found that over 25 million USA adults regularly share advice on products or services online.

DOI: 10.4018/978-1-60566-134-6.ch011

The Web 2.0 technologies and applications (e.g. tags, RSS, blogs, wikis, podcasts, etc.) are considered as the *tools of mass collaboration*, since they empower Internet users to collaboratively produce, consume and distribute information and knowledge. In other words, Web 2.0 tools do nothing more than realizing and exploiting the full potential of the genuine concept and role of the Internet (i.e. the network of the networks that is created and exists for its users). This has tremendously changed the way people search, find, read, gather, share, develop and consume information, as well as on the way people communicate with each other and collaboratively create new knowledge (Sigala, 2008). UGC and Web 2.0 technologies are also having a tremendous impact not only on the behavior and decision-making of Internet users, but also on the e-business model that organizations need to develop and/or adapt in order to conduct business on the Internet (Bughin, 2007).

The tourism industry is not an exception from such developments. On the contrary, as information is the lifeblood of tourism, the use and diffusion of Web 2.0 technologies have a substantial impact of both tourism demand and supply. Indeed, more than ¼ of Internet users have used a weblog to review information about a destination or travel supplier in the last 12 months (Harteveldt, Johnson, Epps & Tesch, 2006), many new Web 2.0 enabled tourism cyber-intermediaries have risen challenging the e-business model of existing online tourism suppliers and intermediaries who in turn need to transform their e-business model and e-marketing practices in order to survive **(**Adam, Cobos & Liu, 2007). As the Internet plays an important role for the e-marketing of city destinations (Sigala, 2003; Yuan, Gretzel, & Fesenmaier, 2006), Web 2.0 tools and applications also create both threats and opportunities for organizations developing and maintaining destination management systems and portals. In this vein, this chapter aims to inform city tourism organizations responsible for the development of city portals about: a) the use of the major Web 2.0 tools in tourism and their impact on tourism demand and supply; and b) the ways and practices for integrating the use of Web 2.0 into their e-business model and e-marketing practices.

WEB 2.0 TOOLS IN TOURISM: USE, IMPACT AND APPLICATIONS IN CITY MARKETING

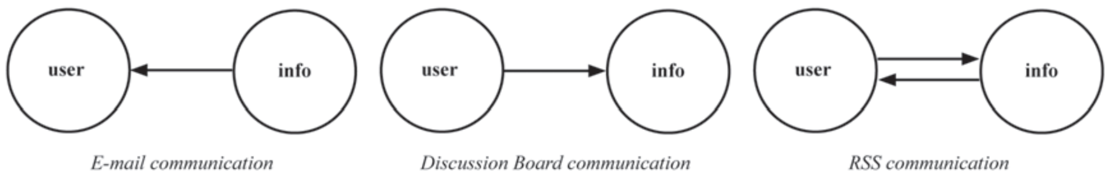
RSS (Really Simple Syndication)

Definition, Features and Use

RSS allow users to subscribe to a webpage for receiving new content, e.g. subscribe to receive online distributions of news, blogs, podcasts etc, and so, RSS allows the creation of links and interactive communication amongst other Web 2.0 applications and users. This is done either through a news aggregator (similar to an email inbox) or a news reader (a web-based environment) (Winer, 2005). By doing so, one does not have to visit each individual website that he/she is interesting to read any new information, but rather the RSS feeds all new updated information to the users' RSS reader. RSS readers enable Internet users to gather and read all new information that is customized to the user's profile within one consolidated message. Many free RSS exist on the Internet, e.g. FeedDemon, NewsGator, Rojo, software on the website of Google™, MyYahoo®, etc.

RSS allows new communication and interaction modes with information (Figure 1) (Farmer, 2004). In e-mail, the control of the communication channel is held entirely by the instigator of the communication. Consequently, e-mail communication is characterized at times by flame wars, antisocial behavior and feelings of intrusion by the participants, while the information artifact is transitory, unfixed and not archived except in individual instances. In discussion boards, information artifacts are fixed, frequently archived and can be interacted with through threading and comments,

Figure 1. Communication types



but for accessing information, the user must deliberately visit a dedicated online area. In contrast, in RSS, both the communicator and the reader of information have control of the communication process, i.e. the former sends information only to those that users have selected to aggregate the RSS feed, while the later select from where and how (e.g. summaries, titles, or full entries) to receive communication. Further, as most RSS aggregators are either integrated with or stand-alone desktop/web applications, readers need only to check the aggregator for new items.

Impact on Tourism Demand

RSS feeders have a tremendous impact on the way consumers search and read information nowadays. RSS has the following benefits for users:

- saves a lot of time spent on information searching;
- provides users with consolidated personalized information;
- is less obstructive and more personalized to users' interests than other Internet based communication, and so, RSS entice subscribers to visit the related websites, thus helping in building website traffic and visitation;
- RSS boosts viral marketing and online word-of-mouth as users tend to forward items in RSS feeds to their friends, family and co-workers, much like the 'forward this message' feature in eNewsletters.

Business Applications for City Marketing

As RSS is an information distribution technology that is characterized as a demand pull rather than a supply push model, many tourism destination organizations have adopted and incorporated RSS feeds in their websites in order to communicate with their potential and current travelers in a less disrupted and personal way. Some examples of RSS include:

- Keep a communication with their travelers such as sending them Newsletters and/or updates of the programme of cultural events organized in the city.
- RSS helps organizations to enhance their Search Engine Optimization by creating inbound links to a company's website and by informing search engines whenever new content is uploaded on a website, so that they can index it.
- RSS is used for syndicating content to other Websites expanding the original website's readership and reach.

For example, the destination marketing organization of Las Vegas has included an RSS on its website (visitlasvegas.com), whereby users can subscribe to feeds that automatically notify them of current travel specials. Almost all of the information (e.g. events' news, weather updates, special offers, etc.) on the official city website of Dublin created by the city tourism board (visitdublin.com) is available to any traveler and/or other website through RSS. RSS are offered for free for anyone for reading and/or enriching his/her own web site, provided that the latter follows the proper format, terms and conditions and attribution, e.g. attribution such as *"Content provided by Dublin Tourism"*. In such a way, visitdublin.com aims to enhance readership of its content, continuous personal communication with its customers, viral marketing, and search engine optimization through content syndication and incoming links.

Blogs (or Weblogs)

Definition, Features and Use

Weblogs began as personal writing spaces that store and update regularly multimedia content (in reverse chronological order) and links of interest to the author. Thus, blogs are used for recording its author's journey and sharing it with others by using links, RSS, trackbacks, comments, taglines, archives, permanent links, blogrolls, etc. (Blood, 2000). Weblogs are defined as a ''… site consists of dated entries'' (Blood, 2000), whereby entries are episodic or conversational in a diary or ''story telling'' format. Motivated by different reasons (Forrester Research, 2006), such as documenting one's life, providing commentary and opinions, expressing deeply felt emotions, articulating ideas through writing, and forming and maintaining community forum, weblogs (or blogs) are a *"new form of mainstream personal communication"* (Rosenbloom, 2004, p. 31) for millions of people to publish and exchange knowledge/information, and to establish networks or build relationships in the world of all blogs. Indeed, blogging tools enabling between-blog interactivity are building up the ''blogosphere'' whereby social networks among bloggers are created. Du and Wagner (2006) identified the following characteristics of blogs:

- **Personalized:** blogs are designed for individual use and their style is personal and informal. Blogger.com offers a ''team blog'' collaborative feature enabling also multi-person weblog.
- **Web-based:** blogs are easy to access and frequently maintain by simply using a web browser.
- **Community-supported:** Weblogs can link to other weblogs and websites (e.g. photos, videos, web-texts), enabling the synthesis and linkage of ideas from different users, and so, stimulating meta-knowledge, i.e. the generation of new knowledge through sharing.
- **Automated:** Blogging tools are easy to create and maintain without the need of HTML programming skills and knowledge; so, bloggers can solely concentrate on the content.

Du & Wagner (2006) also identified 3 types of blogging tools and features: type I features of blogging (such as text, diary, hyperlinks, user friendly editor) provide easy-to-use and learn tools for editing, presentation, publishing and interlinking of content. Such blogs are heavily used by those that solely seek a channel of self-expression. Type II blogging tools (such as Indexed archive, search, "permalink": a permanent URL for each weblog entry enabling referencing of specific past entries like other web-source. ''Trackback'': a reverse hyperlink tracking the referrer weblogs ''making these formally invisible connections visible'', categorisation & syndication) are used by bloggers who wish to easily share rich media (e.g. videos, pictures etc), to have a sophisticated content management system and to enable between-blog commenting or hyper linking, e.g. through ''permalink'' or trackback (Blood, 2000). Nowadays, the emerging Type III of blogging tools provide improved content distribution and between-blog connectivity (e.g. ''pingback'', alert of other bloggers' comments or new posts), as well as integrated applications for further enhancing social networking and community building such as the following examples:

- Workflow or project management (e.g. Lycos Circles workflow for a party, from invitation to management of responses and to travel directions),
- Polls, Intrasite messaging (e.g. ModBlog allows users to track friends' newest entries, or to know who are the most ''recent visitors''),
- Web invitation, picture/music sharing (e.g. MSN Space picture/music sharing, and remote posting of updates via email or mobile devices).

Impact on Tourism Demand

Numerous examples of general and-or (content or user) specific blogs have been created in the tourism industry, such as tripadvisor.com, hotelchatter.com, bugbitten.com, placeblogger.com, realtravel.com, travelpod.com, igougo.com, gazetters.com (a B2B blog for travel agents). Many travelers and tourists also develop and maintain their own blogs for sharing their travel experiences with others and distributing their feedback (reviews) of travel suppliers for achieving fun, social recognition, prestige and-or self-expression. Due to the unbiased information shared in blogs based on first-hand authentic travel experiences, many travelers tend to use and trust blogs' information for searching for travel information, tips and selecting travel suppliers and destinations. Blogs have the power of the impartial information and the electronic word-of-mouth that is diffusing online like a virus. Hence, blogs are becoming a very important information source for international travelers for getting travel advice and suggestions. Moreover, when reading others' travel experiences through weblogs, this also creates to the reader the willingness to travel and visit the same destination or suppliers. Indeed, research has shown that UGC at blogs has a similar AIDA effect to users as paid advertisements have. The latter is because blog's content can (Lin & Huang, 2006): 1) **A**ttract the attention, eyeballs of other Internet users and increase traffic on a website, 2) create **I**nterest to users who can now seek more and additional information, 3) develop someone's **D**esire to also visit a destination and/or buy the product and 4) foster an **A**ction (e.g. book a hotel or organize a trip to a destination). Of course, it should be noted that the power of blogs can also be negative, i.e. spread a bad experience of a tourist to million of online Internet users. Therefore, it is very important that tourism companies authorize a public relations staff as the responsible representative of the company to first scan and read blogs' content and then respond to formally any positive and negative users' comments. Guidelines and corporate policies for responding to UGC should also be established. Nowadays, many blogs take the form and are presented in a video format (Vlog, video blog). The first travel website to implement vlogs exclusively is endlesseurope.com. Due to the multimedia features of video content and the intangible nature of the tourism product, it is argued that vlogs are going to have a much greater impact and influence on travelers' decision making and evaluation of alternative tourism products and suppliers.

Business Applications for City Marketing

Blogs generate and distribute a plethora of UGC related to travelers' experiences, suppliers' and destination reviews, travel tips and advices. City destination management organizations can exploit and use such content for:

- monitoring and influencing electronic word-of-mouth;
- conducting an easy, free, timely and reliable market research about travelers' preferences, feedback and profile;
- communicating with current and prospective travelers in a very personal and informal way;

- gathering travelers' feedback and responding to customers' complaints;
- enhancing search engine optimization: blogs are becoming very important tools affecting information search since their links, content and popularity can dictate the position of a company on a search engine search.

There are many search engines that one can use for identifying and locating weblogs. The most popular one is technorati.com, which also provides statistics about the online activity of weblogs, e.g. about the popularity of a blog and its potential influence on search engines results.

Apart from exploiting others' blogs, many tourism suppliers and destination management organizations have also adopted a pro-active strategy by creating and incorporating blogs on their own websites. For example, Marriott has created its own blog on its website (www.blogs.marriott.com), while Starwood has created a blog to communicate with its Preferred Guests and enhance their loyalty through the website www.thelobby.com. Company initiated and moderated blogs can offer the following benefits: solicit and gather feedback from customers; conduct free online market research; become recognized as an expert on a specific topic; communicate and update your customers with latest news; and use others' customers' suggestions for helping customers select and evaluate products such as what amazon.com is doing by allowing users to upload books' reviews on its website. For example, Eurostar has initiated a blog (www.voiceofacity.com) whereby it has commissioned local Parisians to post blogs for creating a travelers' guidebook with a truly ground-roots feel. The destination organization of the city of Los Angeles has created a blog supporting the sharing of bloggers' experiences and insights on their adventures of the diversity of Los Angeles' arts and culture (http://blog.experiencela.com/). The fully Web 2.0 enabled official portal of Holland features a blog capability (i.e. the triplog) enabling Dutch travelers and locals to share and post the experiences they lived in Holland (http://us.holland.com/blog.php?sf=e.pagerank&so=DESC&sel=popular). In this vein, blogs are becoming a useful tool for enabling local communities to get more involved in destination marketing, communicate and blur with tourists. As a result, blogs and web 2.0 tools can support and foster community participation in city tourism development and marketing practices. Community participation can ensure a better blend between locals and tourists reducing any inter-cultural conflicts, creating social relations and respect and understanding amongst different cultures as well as enabling multi-stakeholder understanding and communication in tourism decision making and activities.

Social Networking–Collaborative Networking

Definition, Features and Use

Social networking websites enable users to create their profile and invite others with similar profile to take part in their online community. The most popular websites such as myspace.com and bebo.com reflect the willingness of Internet users to transform websites as a gathering place of people with similar profiles.

Impact on Tourism Demand

Many social networking websites have been created in the tourism industry allowing travelers and prospective travelers to network with one another based on shared interests or attributes, such as tripmates.com, gusto.com, triporama.com, triphub.com, traveltogether.com and wayn.com. Travelers log into websites and create a personal profile detailing their travel experience and interests, then network with others o share travel advice and stories, and even plan trips together. Hence, social networking websites have a tremendous impact on how tourists nowadays create, organize and consume tourism experiences. Many tourists nowadays prefer to have the reassurance of other users sharing similar profile with them that the trip, the travel company, destination and/or the itinerary that they have scheduled is a good one and it matches their preferences and tastes. Many tourists also wish to use the Internet for collaboratively organizing a group trip with their friends. Websites such as tripup.com, traveltogether.com and travelpost.com enable tourists to create an itinerary, e-mail and share it online with others, who in turn can edit, modify and enhance it, post it back to others for further comments and / or invite and read other travelers' comments and advices on the trip they organized in order to finally achieve a consensus and proceed to a group booking.

Business Applications for City Marketing

Since sharing travel experiences in a social website can significantly inspire travel and boosts one's willingness to visit a destination or supplier, several tourism websites are incorporating social networking tools in their e-business models. For example, existing cyberintermediaries, such as Yahoo!® Trip Planner has adopted a collaborative trip organizing and booking tools. The official website of the city of Philadelphia has also features a collaborative trip planning tool (www.planit.pcvb.org), that potential travelers can use for organizing their itinerary in Philadelphia with friends as well as soliciting feedback and comments from other travelers and locals.

Lufthansa created and operates its own social networking website, named as Jetfriends, for enabling its young flyers to share flight experiences and indoctrinate them into the Lufthansa brand and frequent flyers' club (http://www.jetfriends.com/jetfriends/kids/). Sheraton also re-organized and re-designed its website (which is nowadays titled as the Sheraton Belong Neighborhood), whereby Sheraton's guests can subscribe to the website, upload their experiences, stories, pictures, videos and comments for sharing them with other website visitors and users. The social networking of Sheraton's website enables potential travelers to organize and book their holiday and hotel experience at any Sheraton property that matches their profile and preferences by reading and reflecting on the comments and first-hand experiences written by previous Sheraton customers. The impact of social networking features for persuading potential travelers to select a particular hotel and/or destination is very powerful, because through social networking websites, travelers can search website content based on keywords and stories contributed by other travelers that may be more relevant and make more sense to them than keywords and experiences being created and pushed by the website developers themselves. For example, on Sheraton's website one can search an hotel experience based on the comments and tags contributed by a previous guests referring to "*nice walking in beaches nearby Sheraton hotels*" or "*relaxing family holidays in Sheraton properties*" rather than using the Sheraton's search engine to find hotel based on its location, facilities etc.

Nowadays, many city destination organizations have also incorporated social networking features into their e-business model and strategy in order to further enhance their communication with customers and take benefit of the electronic word-of-mouth that they can create. For example, the official website of Los Angeles invites any cultural and art organization-institution to become a partner with experiencela.com in order to share and distribute related content on their website. Experiencela.com has also created a cooperation e-business strategy with the social networking websites clickr.com and myspace.com. Specifically, the destination management organization of LA created a special webpage for LA on flickr.com (http://www.flickr.com/groups/21164279@N00/pool/) and myspace.com (http://www.myspace.com/experiencela) in order to enable the users of the formers that are also funs and travelers of LA to share personal photographs, comments and stories about LA. The city of LA has realized that such UGC can crucially drive traffic to their website, boost their search engine optimization strategy, instil travelers' desire to visit LA and use customer intelligence for providing reliable and timely advice and suggestions for trips to potential travelers to LA.

Tagging (Social Search and Tag clouds)

Definition, Features and Use

Tagging represents a new way for categorizing information. Users tag a piece of content (e.g. an audio, a picture, a word) with a meaning (a word or phrase) and then this information is categorized in categories based on this meaning. Community tagging is a bottom-up, grass-roots phenomenon, in which users classify resources with searchable keywords. The tags are free-form labels chosen by the user, not selected from a controlled vocabulary. Tagging is also known as consumer-generated taxonomy. Forrester (2006) defines tagging as 'the act of categorizing and retrieving Web content using open-ended labels called tags'. Tagging provides customer value, because it allows them to assign their own word or words to mark products and content online in order to categorize content that they find relevant, i.e. such as what bookmarking allows users to do. Words that users choose for categorizing website and content then become a navigation shortcut that a person can use to browse and search content throughout that site.

Tagging is used not only for saving and sorting a user's content but also for sharing content with others. Websites with tagging capabilities can also allow users to share their personal tags and navigation ways with other users. Moreover, some tag enabled websites enable users not only to share their tag navigations, but also their profile. In this way, users can see who has tagged something, and try to search and find information based on the search behavior of users with similar profiles and mental maps with them (personalized social searching). In this vein, tagging has a great effect on how search engines identify and present information results in keyword searches to users.

Flickr.com represents the first wide-spread of tags, whereby users can add their own tags to any photo they wish to share, aggregate pictures into photosets, create public or private groups, search photos by tags and easily add flickr-stored photos to a blog. Nowadays, there are numerous websites enabling tagging and searching based on tags, e.g. del.icio.us, a bookmarking service, Technorati, a blog cataloging site, and digg, a gathering place for tech fans. These sites create clickable "tag clouds" for resources, groupings of tags arranged alphabetically, with the most used or popular keywords shown in a larger font. In this way, these websites present other websites that users think are important or relevant to them. Many such sites make use of RSS to notify interested users of changes and new developments, e.g. in flickr.com, RSS feeds can be attached to individual tags, or to photos and discussions. In addition to RSS, flickr.com and other social networking sites typically offer functions such as search (for users and tags), comments (and comment trails), and APIs (application program interfaces) for posting to or from the tools, that can be used in combination with blogs. An interesting use of RSS that is combined with tagging is the Flashcard exchange, where, one can view or subscribe to all flashcards posted for learning Spanish (or other languages).

Although the tagging process is by no means simply technical—a way of categorizing resources—it has a strong social dimension as users of the website find common interests and create on-line communities. It represents another example of the fuzziness separating consumers and creators on the Web. A contribution to a tagging site, seen by other users, may cause additional tags or comments to be added, automatically building and updating and thus ultimately defining a resource. Instead of one person making a judgment about a blog entry, photo, or other resource, a consensual classification is created. In effect, a text or object identifies itself over time. This creation of "folksonomies" (as the user defines how to sort information which in turn defines how others search and find information) can be seen as a democratic implementation of the Semantic Web. Thus, for some, tagging helps and boosts the creation of the semantic web (Web 3.0), whereby web content and search is directly related to its meaning for the users.

Impact on Tourism Demand

Several websites offer the capability for users to sort, share and categories travel related content based on tagging, e.g. flickr.com (for pictures sorting and sharing), travbuddy.com (for travel experiences sharing), travelistic.com (for tagging video content). However, although more and more users are using collaborative tools for identifying and sorting travel content, tagging is still an emerging technology: only 5% of USA online leisure travelers—slightly more than 5 million of the 114 million USA adults who travel for leisure and go online regularly—tag Web pages or other content on sites like del.icio.us or Flickr™ (Epps, 2007). Moreover, the social capabilities of tagging for community building and social collaboratively construction of concepts' meaning and of travel experiences could have numerous innovative applications in tourism as well. For examples, travelers may be enabled and offered the opportunity to build structural tags in a text using XML for creating word groups or simply finding appropriate keywords to describe a travel experience. This would offer additional options to Internet users to collaboratively develop travel itineraries and search of travel information with others sharing a similar profile with them.

Business Applications for City Marketing

Because of the power of the folsconomy to provide enhanced user-value and influence the results/page ranking of search engines' search, many tourism firms nowadays include and consider tagging when designing their websites and e-business strategy. For example, Thomson's website provides an affiliate link to deli.ci.ous.com (http://www.thomson.co.uk/), so that its users can tag and sort its website content through this technology.

However, tourism firms may adopt different strategies regarding the way they use and incorporate tagging into their websites (Epps et al., 2007). For example, Triporama.com has launched its tagging system, titled "Triporama Bookmarks", which allows its website users to download a free software in their PC for tagging content in their own words from anywhere on the Web to their Triporama group trip plan, which they can then share with their travel companions. Such a solution provides differentiation customer value for Triporama, because as travelers visit many different Websites while planning a trip, Triporama's bookmarking tool lets travelers collect, label, and share the content they have found on the Web with other members of their group. Users also have the option to make their tags publicly available, while Triporama also aims to edit and curate these public tags into features like "top 10" lists to give other users ideas for planning their own group trips. When redesigning its website, Sheraton introduced its "Vacation Ideas" feature whereby guests are invited to write stories about their hotel stay, users give Sheraton their consent for publishing their stories online and the entire story becomes a tag. A tag cloud is created, titled the "Buzz Barometer", whereby word occurring most frequently in stories appear in bigger fonts, while based on the number of stories shared containing different words a "Vacation Ideas page" is created (http://www.starwoodhotels.com/sheraton/index.html). For example, by clicking on the "Beach" guide brings up the five hotels whose stories mention "beach" most frequently (weighted by the number of stories relative to the size of the property). By making storytelling the method by which the tags are created, Sheraton has made tagging so friendly and easy for its guests that they do not even know that they are tagging content. Sheraton benefits from this tagging strategy because:

- It helps first time website users: Vacation Ideas gives travelers, not knowing where they want to go and/or not familiar with the Sheraton brand, a more creative, user friendly and understandable way to search and book hotels than the customary destination-based and company pushed search.
- It helps Sheraton to build long lasting relations with its guests by maintaining a close relationship with the travelers before, during and after the trip. This is because the website provides guests having stayed at a Sheraton's hotel to return to the Sheraton's website in order to contribute, solicit or read other travelers' stories.
- It improves organizational learning, since Sheraton gets insights into Sheraton's hotel properties and customers' experiences. Instead of conducting expensive and time consuming research, Sheraton uses tagging as a simple and reliable way for gathering customer feedback and intelligence about its products and services. Based on the customer knowledge that is gained, operation managers can improve organizational processes, while marketers get to know what and how customers think and talk about the brand, in order to better position the Sheraton brand in the market and enhance the guest experience at the hotels in ways that reinforce guests' perspective of the Sheraton brand.

Yahoo!® Travel introduced new tagging features into its Yahoo!® TripPlanner that enable users to tag their own or others' Trip Plans with suggested or custom tags, which are later analyzed and used by Yahoo! for identifying and feeding recommendations on the Yahoo! Travel home. Users can set their preferred level of privacy at the level of the Trip Plan (private, shared with invited friends, or visible to any user). Users are provided around 30 tags (e.g. ("budget," "luxury", "weekend", "honeymoon") to choose from for labeling trip plans from the style of the trip, however, Yahoo! monitors the tags used most frequently by its users for augmenting its list of proposed tags. By using taxonomy-directed tagging, Yahoo!® eliminates many of the inherent problems of folksonomies created when users label similar things differently using synonyms or different forms of the same word (e.g. lodging, accommodation, hotel etc.). Tagging has helped Yahoo!® to: a) make its content (750,000 Trips Maps, photos, users' comments etc.) more useful, accessible, searchable and understandable to its users; and b) to gather, analyze and use customers' intelligence (where they live, where they have traveled, and what content they have viewed) and further refines that knowledge through the lens of the tags they use to search in order to create targeted, personalized recommendations for destinations and deals sold through its vertical search website Yahoo!® FareChase. In other words, tagging helps to further refine the collaborative filtering process that Yahoo! uses in order to provide personalized recommendations and suggestions to its users. Personalized suggestions for cross and up sales can significantly drive and enhance booking and sales levels as well as provide additional functional and emotional value to website users that in turn enhances customer loyalty.

In the same vein, the official destination portal of Holland uses tagging technology in order to provide travelers an easy way to search the website content and its multimedia information (video, photos etc.) (http://us.holland.com/). Actually, tags are used as an user constructed and defined search engine rather than providing a search engine designed by the website developer that reflects a top down business defined search process. Tag clouds appear on the left with different font sizes to reflect words used more or less frequently, while "Top 10" suggestions for each tag (e.g. restaurant) are also constructed and updated continuously when new content and UGC is shared on the website. The portal also provides users the possibility to comment each others' contributions and comments as well as to tag the Holland's webpage content by using different social booking technology such as Digg and Furl.

Overall, it becomes evident that city tourism organizations should consider including tagging into their websites, as tagging can help them overcome the following issues (Epps, Harteverldt & McGowan, 2007):

- Very frequently websites do not speak the same language and they do not use the same terminology as their users. City destination organizations should consider using tagging in order to make their website content more accessible, understandable and appealing to its users. For example, the marketer of a city might promote as the major value of the destination its easy accessibility by air transportation, however, travelers may perceive as the most valuable feature of the destination the fact that it is "a safe city to walk around".
- Tagging can help and further enhance keyword search by supporting nuanced, adjective-based searches. Tagging also enables social search whereby users can see who has tagged something, how credible or relevant its suggestions are based on his/her profile and his/her evaluation by other users.
- Tagging helps organize and display user-generated content uploaded on websites. As more and more city tourism organizations invite their users to upload and share their UGC (e.g. reviews, itineraries, photos, videos, and podcasts), they later struggle to make this UGC relevant and accessible to their users and tags can help in addressing the latter.
- City tourism organizations can gather reliable and timely customer intelligence and feedback regarding the image of their destination, the mental maps of their travelers and how they view and perceive their destination etc. Such customer knowledge can be later used for marketing campaigns as well as for improving the products and services of the city as a destination.
- Customer information gathered through social tagging can also be used for improving search engine optimization campaigns. For example, words used frequently as tags by travelers can be used as metatags-metadata for building the portal's website as well as for spending money on keyword sensitive search engines such as Google™ AdWords.

However, when deciding whether and how to use tagging, city tourism organizations responsible for the development of the city portal should also decide the process, the policy and the way for developing their tagging system regarding the following issues: a) does the company edit the tags incorporated by users? This is important specifically if tags are uploaded with spelling mistakes or they include anti-social and embarrassing words. In other words, editing and a clear policy may be required in order to protect the consistency, the ethos and the good image of the website; b) are the tags and taggers' profiles made publicly available for everyone? What consent and agreement are required to take from the users and how the privacy policy of the website should be amended to incorporate the former?; c) is a software going to be purchased to manage the tagging process or is this going to be done manually? Are the required and skilled labor sources available? and d) how tags are going to be created? Are the tags going to be provided by the users or are the tags restricted by the website owner? Forrester (2006) recommends that companies use *taxonomy-directed* tagging, as it makes the tagging process more efficient and easier to use and it promotes consistency among tags. This is because when users create a tag, they can choose from existing suggested tags, or they can add their own.

Wikis

Definition, Features and Use

Wiki is a piece of server software permitting users to freely create and edit (hyperlinked) content via any browser and without the need to have access to and know to use any programming language. A wiki is a collaborative website whose content can be edited by anyone who has access to it. Wiki features include easy editing, versioning capabilities, and article discussions. So, wiki technologies enable users to add, delete, and in general edit the content of a website. Wiki users-creators are notified about new content, and they review only new content. As a result, such websites are developed collaboratively through their users, and a wiki becomes a collaboratively expandable collection of interlinked webpages, a hypertext system for storing and modifying information—a database, where each page is easily edited by any user with a forms-capable Web browser client. Neus and Scherf (2005) defined wiki as web content management systems allowing collaborative creation, connection and edition of contents, while Pereira and Soares (2007) defined wiki as a shared information work space that facilitates access to information content, organizational communications, and group collaboration. In other words, wikis represent another way of content publishing and communication as well as for group collaboration. In this vein, wikis and blogs have some similarities but they differ regarding the notification of new content, editing format, and structure. In other words, *'a wiki can be a blog, but a blog does not have to be a wiki'*.

Impact on Tourism Demand

The most popular wiki is the famous online encyclopedia, titled wikipedia.com, that is created and continually updated by its users. In tourism, wikitravel.org represents a wiki based effort of Internet users to collaboratively create and continuously update an online global travel guide including world-wide destinations. The number of readers, creators and content at wikitravel.org are continuously mushrooming. At wikitravel.org, one can find guides for any destination irrespective of its size and/or geographic location, as well as create a guide for any destination that he/she wishes. Wikitravel.com is further enriched with other web 2.0 tools and technologies such as maps, tags, podcasts etc.

Business Applications for City Marketing

Many tourism organizations take the opportunity to promote and create links to their websites through wikitravel.com in order to create and drive traffic to their own websites (http://wikitravel.org/en/London). Many other destination management organizations exploit and incorporate the wiki technology in their website portals in order to enable its users (travelers and locals) to collaboratively create and share their perceptions and mental images and opinions about their destination. For example, the National Library of Australia has included a wiki on its portal (http://wiki.nla.gov.au) inviting users to share their understanding and knowledge of local Australian dances as well as negotiate their meaning and create metaknowledge by synthesizing different views and perspectives. The National Library of Australia has also developed a wiki and social network website (www.pictureaustralia.org) whereby users can share their pictures about Australia and tell their story. In this way, the Library aims to help democratize history and establish a collective memory of places and events around the country. Ancient Times website (http://ancient.arts.ubc.ca/community.html) includes several collaborative tools, such as a wiki, blog and an arts metaverse enabling any user and history student to collaboratively develop and negotiate the meaning and construction of old cities and destinations, such as Giza and Athens. These cultural guides can significantly enhance the appeal and the interpretation of the cultural artifacts of historical cities and destination by providing several edutainment services and benefits to their users/visitors (Sigala, 2005a). Other wiki applications can also be provided on the city portals in order to boost website loyalty, repeat traffic, and travelers' desire to visit the destination. For example, a destination organization can design and incorporate a wiki on its portal for enabling potential travelers and locals to exchange and collaboratively develop recipes of local dishes and food.

Podcasting and Online Video

Definition, Features and Use

Podcasting refers to the uploading of audio and video files by users on websites. The most popular website for sharing such content with others is youtobe.com. Podcasting represents repositories of audio (podcasts) and video (vodcasts) or "video podcasting" materials that can be "pushed" to subscribers, even without user intervention, through RSS aggregate feeds of audio and video content facilitating users to search the latest services. Podcasting-capable aggregators or "podcatchers" are used to download podcasts. These files can also be downloaded to portable media players that can be taken anywhere, providing the potential for "anytime, anywhere" learning experiences (mobile learning). Podcasting's essence is about creating content (audio or video) for an audience that wants to listen when they want, where they want, and how they want. Podcasting differs from webcasting. A podcast has a persistent site, capable of synchronizing with a portable multimedia device, e.g. an MP3 player or iPod, whereas webcasting is streamed from the internet and requires the user to be connected to the internet while playing or viewing the webcast files. Webcasting is closely related to real-time downloading and synchronous broadcasting. Podcasting adds spatial flexibility to the temporal flexibility that webcasting offers and affords itself for creating personally-customizable media environments. Podcasting offers customer value in terms of the flexibility possibilities to hear personalized content whenever and whatever device one wishes, e.g. one can download the "Economist"'s or "CNN"'s personalized news' and press releases to his/her iPod and listen to its favorite news while he/she driving at work. As podcasting does not rely on the visual senses, it allows users to carry out other tasks while listening.

Impact on Tourism Demand

Tourism experiences are intangible. One cannot experience, feel and try a travel experience before he/she buys and before he/she travels to a destination. As a result, the purchase risk of a travel—tourism experience is high and it is difficult to persuade a user for the qualities of a tourism service. Due to its multimedia features, podcasting helps users to better and easier evaluate travel alternatives by experiencing in someway a travel experience before they decide to buy and consume it and/or travel to a destination. This is because audio and video files of hotels, destinations, and other travel products created and uploaded for sharing by other users are considered as more unbiased information and not staged experiences produced by the supplier aiming to promote his/her own product as the best one. Podcasting has also been used as mobile guides for travelers, e.g. Virgin Atlantic provides through its website free podcasts-guides of cities whereby they fly to.

Impact on Tourism Supply

Many tourism suppliers are using Podcasting as a marketing, information and customer communication tool. For example, Jumeirah hotel uploads podcasts on its website for delivering and updating its potential guests about what is happening in its properties at every single day, and-or delivering to website users and potential buyers the experiences of VIPs that have stayed at their property. Tate Gallery enables their visitors that have experienced their paintings and exhibition to record themselves, upload their audio-video on the Tate Gallery website, and which others can later download and use them as a mobile interpretation guide while visiting the gallery. Orbitz.com provides podcasting of many destinations that travelers can download to their MP3 players and use them as guides while visiting the destination. In a similar way, MGM Grand Hotel Las Vegas has lauded online video on its website under the title "Maximum Vegas" in order to better illustrate to its potential guests the experience and services of its hotel and gaming resort. Similarly, city destination organizations should consider enhancing the content and marketing appeal of their website portals by enabling podcasting opportunities, i.e. either allow users to share content or push their own created podcasting content (e.g. http://www.visitlondon.com/maps/podcasts/, Podcasts at the official portal of London). For developing podcasts, city tourism organizations can outsource this function to companies such as soundwalk.com, podtrip.com and heartbeatguides.com that specialize in the development and dissemination of destination podcasts.

METAVERSES: MASSIVELY MULTIPLAYER ONLINE ROLE PLAYING GAME (MMORPG)

Definition, Features and Use

Metaverses are three dimensional virtual worlds whereby Internet users collaboratively play "online MMORPG games" with others. However, these platforms are wrongly perceived as "simple games" and "virtual" worlds, since they frequently represent an extension to our physical day-to-day world to which users add new socio-economic and political situations. MMORPG are games that are played by numerous players (e.g. millions of users) and they could be considered as an intermediate step from 'computer' to 'ambient' era. Some of these games (e.g. World of Warcraft) develop around a theme defining the goals of the game, while other games, such as SecondLife.com, there.com, cokemusic.com, habbohotel.com and http://play.toontown.com/about.php, encourage a free-style of playing, allowing the users to make what they want out of it. Metaverse environments are internet-based 3-D virtual world whereby their users, called residents, can interact with each other through motional avatars (an internet user's representation of him/herself) providing an advanced level of a social network service. Although it is difficult to measure the size and growth of such games, it is estimated that the market for massively multiplayer online games is now worth more than $1bn in the West world (Book, 2003). For example, one can simply consider the size of and growth of Second Life®® itself. Second Life® has more than 5 million users, while about half a billion US$ are being transacted every year on Second Life's® website (as reported on SecondLife.com on April 2007).

Impact on Tourism Demand

Tourists and travelers participate in such games either for fun and-or for ways of expression of oneself and for achieving satisfaction through task—accomplishment, self-actualization and creation—design of something new. For example, many people dream and try to become and excel on a profession that they could not achieve in their real life, others try to design a new product and service hoping that other players will adapt and pay for it and so they can gain money and/or head hunters would spot their talents and recruit them in their real or virtual companies.

Business Applications for City Marketing

Many tourism and travel related companies have already created their representative offices and headquarters in metaverse environments such as SecondLife.com. Embassies (e.g. that of Sweden), Tourism Authorities (e.g. that of Maldives) of many countries and many tourism companies (e.g. TUI, Burj Al Arab Hotel, Marriott, Costa Cruises) have created their offices and companies on islands of SecondLife.com for boosting their marketing practices such as enhancing customer communication and education about their products/services, building brand reputation and user communities, and achieving word of mouth (WOM) and advertising. Hyatt used residents of Second Life® and exploited their intelligence and knowledge for designing a new hotel concept, named as Aloft; architects and guests were involved in designing the hotel providing their feedback, preferences and specialist knowledge (read the related blog at http://www.virtualaloft.com/). As a result of the popularity of the new hotel, the first Aloft hotel will open and operate in real life in New York in January 2008. Apart from collaborative new product development, a firm can further exploit the social intelligence gathered and generated at SecondLife.com and other metaverse environments in order to conduct market research and to test new product ideas and new advertising campaigns, e.g. Toyota first tested the campaign of its new brand Scion on SecondLife® and then, widely broadcasted the new campaign in real life. Other companies, use SecondLife® for recruiting and identifying new talents e.g. CNN does head hunting of new journalists online.

Many destinations are also moving into the futuristic world of virtual reality and metaverse, as many city and country destination organizations create their virtual destinations. Netherlands Tourism Board recently opened a national tourism board in SecondLife® (http://us.holland.com/secondlife.php), the city of Galveston launched a virtual replica of itself in SecondLife® (http://www.galveston.com/secondlife/), providing their visitors with the chance to become part of an interactive community (Figure 10 and 11). The aim is to provide digital travelers the chance to take guided virtual tours, learn about the history, culture and daily life of the destination, and interact with new virtual friends from around the world. Tourism Ireland has also launched the world's first tourism marketing campaign in SecondLife® (http://dublinsl.com/index.php) including the sponsorship of a range of events and activities, including concerts, fashion shows, and photographic exhibitions, in Second Life's® replica city of Dublin. Dublin's representation in SecondLife.com is the first place-location in Ireland that the Tourism Board created its representation in Second Life®. Similiar to the Dublin creation, Amsterdam in Second Life® comes complete with Dutch signs, canals, trams and a lot of attention to detail. Overall, when investigating the impact of SecondLife.com on its residents' behavior, it becomes evident that historical landmarks and buildings such as Tour Eiffel, London Bridge, Ajax Football stadium etc., have a great effect in building virtual communities of people spending a lot of time on dwelling them. Since it is apparent that real world modern-day cities and their landmark attraction are probably the most effective at driving and retaining visitor traffic, city tourism organizations should exploit this inherent advantage and exploit their cultural and heritage assets in metaverse environments for boosting their city brand name, recognition and promotion.

Mash-Ups

Definition, Features and Use

Mash-ups describe the seamlessly combination of two or more different sources of content and-or software for creating a new value added service to users. Many mash-ups enrich their services with some geographical content such as Google™ Maps; e.g. *The New York Times Travel Section's "36 Hours In…"* mash-up, which allows users to search the *"36 Hours in …"* story archive from a Google™ Map. For example, when visiting the website traintimes.org.uk, one can see on real time where trains are located and when they will arrive at destinations, since the website combines information from Google™ maps, and information from the British rail website about train time tables, delays etc.

There are several mash-up applications in tourism such as new cyberintermediaries including mapping services (e.g. earthbooker.com, tripmojo.com, reservemy.com) and meta-search engines such as farecompare.com. Other examples include: www.43places.com that combines Flickr photos, RSS feeds and Google™ Maps with tagging and user-generated content, allowing users to share their favorite destinations; www.randomdayout.co.uk combines a number of data sources to create a mapped itinerary, using Virtual Earth (Microsoft's equivalent of Google™ Maps). An innovate example related to destination marketing management is the case of the city of Pennsylvania (http://www.visitpa.com): based on a project amongst Google™ Earth, Carnegie Mellon University, NASA, the Pennsylvania Tourism Office and the National Civil War Museum, virtual tourists would have the chance to view Pennsylvania's Civil War trails online. More sophisticated examples of mash-ups are the "Marco Polo" function on triptie.com and the "Trip Planner" function on Yahoo!® Travel, which allow users to integrate content from other websites into the user's own itinerary planning toolkit on the host website.

Impact on Tourism Demand

Travel decisions are complex and involve the searching, comparison and combination of several information located in many different websites. For example, a decision to travel to a destination requires various and a plethora of information about weather conditions, exchange rates, travel and accommodation alternatives and prices, attractions etc. As a result, tourists increasingly demand and expect to combine and cross-check information from different sources, so that they can better and easier make a holistic decision. For example, tourists may not be able to clearly understand where a hotel may be located when the description of the hotel websites states that the hotel is located on the beach, near the beach etc. Tourists easily get confused from different descriptions found in different websites. On the contrary, mash-up websites empowered with maps (e.g. earthbooker.com) enable users to see where exactly a hotel or other attraction is located (sometimes even locate the exact orientation and view of a hotel room and then decide whether to book this room at this hotel).

Business Implications for City Marketing

Mash-up applications have empowered the rise of new cyber- and info- intermediaries offering new sophisticated information services (e.g. flightcompare.com search and compare all flight information from different websites in order to provide comparable flight information within one webpage to its users). Moreover, many tourism suppliers and organizations also enrich their website content with maps in order to make it more user friendly and useful to their visitors, e.g. the official website of London and Dublin use

Google™ Maps with geotags for enabling tourists to identify points of interest, hotels etc. Moreover, many companies leave their software as an Open Application Programme Interface (API), so that users can create limitless combinations of their services. For example, backstage.bbc.co.uk represents BCC's services and opportunities offered to its users, who are enabled to take content from the BBC, re-structure it and present it the way they prefer. Enabling user innovation is another way that companies aim to exploit on users' creativity and intelligence instead of investigating solely on company's R&D efforts.

FUTURE TRENDS

It has become evident from the above mentioned analysis that the two major impacts of web 2.0 and its UGC on consumer behavior and marketing practices are: 1) the electronic word-of-mouth that is created; and b) the opportunities to build and maintain customer communities for enhancing the practices of Customer relationship Management and social marketing. Exploiting web 2.0 for city marketing can have a tremendous effect on the marketing effectiveness, since, as the following analysis and discussion illustrates, both previous issues significantly affect consumer loyalty and purchasing behavior.

Web 2.0 and Electronic Word-of-Mouth (WOM)

Word-of-mouth (WOM) is very important in tourism and in services in general, since objective information about a service experience cannot be easily provided before one buys and consumes the services themselves. Services are intangible and so they are difficult to be tested, tried and evaluated before buying them. Consumers also tend to rely more on consumer reviews when purchasing high involvement products (Park, Lee & Han, 2007), such as several travel products e.g. a honeymoon trip, an adventure travel etc. In this vein, tourism decisions are very complex and risky. Indeed, the literature about information search in the tourism field has recognized the important role of WOM in travel planning and decision making (Hwang, Gretzell, Xiang & Fesenmaier, 2006; Murphy, Moscardo & Benckendorff, 2007). WOM has been found to be one of the most influential information sources for travel (Morrison, 2002). Research has also shown that those with past experience with a specific travel destination and that engage in digital word-of-mouth communication are most likely to be the most preferred and the most influential source of information in the pre-trip stage of travel decision making (Crotts, 1999).

To make travel decisions easier, travelers need to reduce the inherent information complexity of travel decisions as well as the risk related to the service firm (i.e. is that a good and reliable company), the service risk (i.e. is that a service that fits my preferences and needs?) and the purchasing risk (i.e. is that a trustworthy booking and buying channel to use for buying a travel service?). To achieve that, consumers use recommendation-based heuristics and other users' feedback to reduce uncertainty, eliminate the related risks as well as filter and process the plethora of information that must be processed when making decisions (Olshavsky & Granbois, 1979). WOM-based information is heavily used and trusted by consumers for taking travel decisions, because it is seen as more vivid, easier to use, and more trustworthy as it is based on actual experience and typically provided without direct benefits (Smith, Menon & Sivakumar, 2005).

As demonstrated in the above mentioned analysis of UGC in Web 2.0 websites, electronic WOM can take different names and forms such as virtual opinion platforms, consumer portals, social networking, blogs' comments, tag words, podcasting, virtual communities and online feedback mechanisms (Armstrong & Hagel, 1997; Bellman, 2006; Sigala, 2008). Users of Web 2.0 websites and tools may post their own experiences, videos, share their opinion, give advice, or look for answers to their questions. Consumers also perceive electronic WOM to be a reliable source of information (Gruen, Osmonbekov & Czaplewski, 2006). Dellarocas (2003) identified three different characteristics of online WOM relative to traditional WOM: 1) electronic WOM is larger in scale (both in terms of quantity and people impact) due to the Internet's low-cost and networking features; 2) electronic WOM is a powerful and reliable market research tool giving organizations the ability to monitor on real time their operations; and 3) it is difficult to convey contextual cues (e.g. facial expression) through the Internet and peer review websites for example, and so not knowing or seeing who the information provider is makes, it is harder to interpret the subjective information in online interaction. To address this problem, websites often display demographic or other data about reviewers (for example, the length of membership, their location, etc.) in order to help build credibility and trust. Websites may even provide the possibility to users to upload and share their own feedback and evaluation (by incorporating each review into a rating of the reviewer) about the quality of the reviews written by other members. Moreover, because Web 2.0 enables users to identify and use personalized and contextual information (e.g. look at what others' with similar profiles are saying), electronic WOM is considered as both more relevant and unbiased than traditional WOM, whereby one cannot easily track and relate the content with the profile of its original messenger.

Smith, Menon and Sivakumar (2005) claim that consumers prefer such peer recommendations over other forms of input, while Amis (2007) advocated that social network sites have as much influence on consumers as television and more than newspapers. Statistics actually provide evidence of consumers' reliance on electronic word-of-mouth. More than 80% of web shoppers said they use other consumers' reviews when making purchasing decisions (Forrester, 2006). eMarketer (2007d) reports that nearly 6/10 consumers prefer websites with peer-written reviews, and that websites with reviews experience greater conversion rates.

Overall, Dellarocas (2003) summarized organizations' benefits of electronic WOM in the following: brand building; customer relationship management; customer acquisition; addressing customer complaints; market research; product development; quality control and supply chain quality assurance activities.

Web 2.0, Customer Relationship Management (CRM) and Social Marketing

The major aim of CRM is to personalize business services and products as well as develop a 1:1 communications and long lasting relation with profitable customers (Sigala, 2005b). eCRM also requires the development of customers' communities for providing loyal customers with functional, emotional and social benefits and value (Sigala, 2006). The previous section provided practical examples illustrating the way in which web 2.0 applications and tools enable the formation and development of customer virtual communities. By identifying and reviewing the limited related studies that have been conducted so far, the following analysis further supports the capability of web 2.0 to build virtual communities of users and enhance the community benefits (functional and emotional/social) to its users.

Ying and Davis (2007) and, Lento, Welser, Gu and Smith (2006) illustrated how blogs create and maintain strong online communities through their social ties tools such as blogrolls, permalinks, comments and trackbacks. Indeed, many authors (e.g. Lin, Su & Chien, 2006; Ying & Davis, 2007) have started to apply social network analysis for measuring and illustrating the social bonds, networking and communication structures created within the blogsphere. Li and Stronberg (2007) summarized blogs' benefits for firms as follows: search engine optimization; e-word-of-mouth (eWOW); improved brand perception and visibility; instantaneous client feedback; market research and insight; increased sales efficiency; and reduced impact from negative user-generated content. Damianos et al. (2007) advocated that social bookmarking generates social influence and bonds as well as creates value by: enabling resource management, information sharing and discovery, expert finding, and social networking; providing teams with a place to share resources; forming and supporting social networks around interest areas; and feeding expertise finding and user profiling. Awad and Zhang (2007) discussed the marketing benefits of eWOW generated in online review communities and debated firms' efforts and strategies addressing it. By examining the communication tools and social cues of myspace.com, Dwyer (2007) demonstrated the impact of social networking sites on developing customer interrelations and communities. In their study of videos' tags on Del.ici.ous, Paolillo and Penumarthy (2007) found that social tagging can generate community benefits such as: easy retrieval (as users use words they can remember and have useful meaning to them); contribution and sharing; attract attention; opinion expression; play; and self-presentation. Thus, since tagging can be used for providing functional services, creating social ties, market research on users' opinions and interests, and WOW, social tagging's ability in creating user communities is evident. Forrester (2006) demonstrated web 2.0's ability to generate customer and business value in different processes: customer service (e.g. community self-service savings); sales (e.g. community loyalty and sales reduces commissions and price competitions); marketing (e.g. credibility of eWOW); production (e.g. co-design reduces waste); and R&D (e.g. community input raises success rate).

A significant amount of literature also highlights the business benefits from developing virtual communities particularly in the area of CRM and social marketing. Analytically, Wang and Fesenmaier (2004) illustrated that virtual tourist communities are useful for managing customer relations by: attracting customers through in-depth, focused and member-generated content; engaging customers through social interactions; and retaining customers through relation building with other members. Online communities also build customer value (Wang & Fesenmaier, 2004) by generating users with all types of relational benefits namely functional, social, hedonic and psychological (Gwinner, Gremmler & Bitner, 1998). Kim, Lee and Hiemstra (2004) provided evidence of the impact of virtual communities on travelers' loyalty and product purchase decision making. Andersen (2005) explored the use of on-line brand communities for developing interactive communication channels and establishing social and structural bonds with devoted users. Jang, Ko and Koh (2007) showed that online brand communities posses and develop features - such as, quality and credibility of information, service quality, member interaction and leadership, brand reputation and (intrinsic and extrinsic) rewards for members' activities- that in turn, contribute to increased users' brand loyalty, commitment and sales. Erat, Desouza, Schafer-Juger and Kurzawa (2006) discussed how different types of communities of practice (e.g. B2C, C2C) can be used for acquiring and sharing customer knowledge in order to improve business processes and performance. Beyond collecting customer knowledge, online communities can also be used for co-operating with customers for New Product Development (NPD) and innovation (Rowley, Teahan & Leeming, 2007). A plethora of cases and research studies (e.g. in Lagrosen, 2005; Pitta & Fowler, 2006) reflects the possibility to use virtual communities for NPD as well.

Table 1. Web 2.0 extended CRM implementation

	Low market integration	High market integration
High customer integration	**Many-to-one** Target: clients' networks Active customers' involvement *e.g. Lonelyplanet.com, Sheraton.com*	**Many-to-many** Co-exploitation of customers' profiles with other network partners *e.g. mash-ups, earthbooker.com, flightcompare.com*
Low customer integration	**One-to-one** Target: individual customers	**One-to-many** Ecosystems of partners offering a seamless experience to individual clients (cross-selling, products' bundling) *e.g. travelocity.com*

Proposed Models for Exploiting Web 2.0 in Enhancing Marketing Communication and CRM

The previous analysis and industry examples illustrate that web 2.0 tools and applications have a twofold impact on the way CRM is implemented: 1) web 2.0's networking and connectivity capabilities provide enormous opportunities to communicate and co-operate with customers and industry partners in many different directions (e.g. many-to-one, many-to-many) (Table 1) the social intelligence and knowledge created collaboratively in web 2.0 platforms (i.e. the user-generated content) can be exploited in different ways for identifying, developing, enhancing and maintaining relations with profitable customers (Table 2).

Table 2. Exploiting social intelligence for managing and enhancing customer relationships through their lifecycle

Phase	Type of customer information/ intelligence	CRM implementation activities
Acquisition	Of the customer information: transaction and personal data	Create brand awareness and recognition amongst customers and virtual communities by building and supporting electronic word-of-mouth
		Develop brand reinforcement and trust by educating and informing customers about the brand, its services, functionalities etc
		Use customer intelligence in order to identify and target new customers, e.g. clone the profile of existing product-service users, use the connections and recommendations of existing customers etc.
		Use customer intelligence to understand how customers use the service, what functionalities they prefer or not
		Use customer intelligence for profiling customers
Retention	For the customer information: relationship and product data	Use customer intelligence for enhancing customer service and transactions
		Use customer intelligence for personalizing services and products
		Build and develop community of customers-users
		Use customer intelligence for innovation & NPD
Expansion	For the customer information: relationship and product data	Use customer intelligence for cross selling, e.g. suggest compatible products based on other users' purchases
		Use customer intelligence for up-selling
		Use customer intelligence for developing affiliation and loyalty programmes
Win back	By the customer information: feedback and monitoring data	Use customer intelligence (feedback, reviews etc) for identifying pitfalls and faults
		Use customer intelligence and communities for handling customers' complaints
		Use customer intelligence and communities for monitoring and managing the firm's reputation, status and prestige

In other words, CRM cannot anymore be considered as synonymous to one-to-one communication and personalized service at an individual basis. Web 2.0 augment CRM practices and implementation to include various forms of communications with clients and business partners. Following Gibbert, Leibold and Probst (2002), Table 3 reflects a two dimensional matrix, whereby the vertical axis represents how firms integrate customers into their value chains and the horizontal axis represents the integration of business partners into the firm's value chain. Companies can use web 2.0 technologies to communicate and enable dialogues and interactions not only between them and their customers, but also between customers themselves (C2C), between business partners, among all of them etc. When engaged in two directional communication both customers and partners, firms can involve the former in their value chain in order to create customer value and benefits. For example, as explained earlier, when customers communicate with other customers in virtual communities, customers provide social and emotional support to others as well as functional benefits (e.g. free consultancy in trip planning). Also, when co-operating and sharing content and applications with other businesses (e.g. in mash-up websites), firms can collaborate with and integrate other partners in their value chain in order to provide additional services to their clients, e.g. a holistic tourism product-services such as a dynamic packaging.

Moreover, in developing successful relationships with profitable clients, firms need to understand and manage all phases through which relations are evolved, as each phase is characterized by differences in behaviors and orientations and so, it requires different CRM approaches. Theory and practical evidence has shown that customer relations evolve over three major distinct phases related to the customer lifecycle (see Sigala, 2008): initiation, maintenance and retention or termination. Hence, all CRM implementation models reflect practices that collect and use three forms of customer information / intelligence in order to manage each relational phase. "Of-the-customer" information includes customers' personal and transaction data for understanding and measuring their profile, e.g. sales, profitability, purchasing patterns, preferences. "For-the-customer" information refers to product, service and firm information perceived as useful by clients for making more informed decisions. "By-the-customer" information reflects customer feedback (e.g. customer complaints, suggestions, reviews) used for new product development or business improvement. As illustrated previously, Web 2.0's user-generated content mushrooms these three types of customer information and provides firms with several opportunities not only to collect, but also to get access to such types of customer intelligence. In other words, Web 2.0 platforms can be exploited as a free and real time market research and intelligence tool. Table 2 summarizes how firms can exploit web 2.0 tools and platforms for collecting and analyzing this customer intelligence for augmenting and supporting their CRM practices.

Overall, it becomes evident that web 2.0 enabled CRM reflects a cultural shift from product *'designing for customers'* to *'designing with'* and *'design by'* customers. For firms to achieve such a cultural shift, crucial organizational changes should also take place. Importantly, the role of marketers should be changed from being sales people to becoming community builders and perceiving customers not as targets to identify and sell, but as partners to collaborate with. Firms should also realize that they should use customer intelligence not only for learning about their customers and identifying new target markets (opportunistic behavior), that they should also use customer intelligence for learning and improving processes and products with their customers as well as with different business partners (partnership relation). In other words, firms derive and realize maximum benefits when they exploit web 2.0 tools for establishing and maintaining cocreation and co-learning adaptable and flexible ecosystems with their customers and business partners (Table 3).

CONCLUSION

Internet users and travelers are nowadays empowered to create and synthesize in their own way the travel content that they also wish to distribute and share it with others through users' controlled distribution channels. In this vein, Web 2.0 technologies enable Internet users to become the co-producers, the co-designers, the co-marketers and the co-distributors of tourism experiences and services as well as the co-entrepreneurs of new tourism products and new e-business models. As the diffusion of Web 2.0 applications becomes wide and consumers incorporate them within their daily and professional life, travelers expect tourism firms and organizations to provide similar Web 2.0 enabled services. The previous analysis aimed at identifying and illustrating the business implications created for tourism and hospitality enterprises as well as strategies and tactics that they can adopt for eliminating threats while exploiting the arising opportunities. Therefore, as Web 2.0 is here to stay, it is evident that unless a city tourism organization adopts and incorporates Web 2.0 tools into its e-business model and strategies for marketing and managing its destination, the competitiveness of the latter is threaten. Nevertheless, in order to be successful, the adoption and use of any web 2.0 tools should be accompanied with appropriate organizational and cultural changes within the firm regarding the roles, job descriptions and tasks of its staff, users and business partners. Further research is required in order to understand and examine how firms are achieving and trying to implement such organizational changes when incorporating web 2.0 into their e-business model.

REFERENCES

Adam, J., Cobos, X., & Liu, S. (2007). *Travel 2.0: Trends in Industry Awareness and Adoption*. New York University and PhoCusWright Inc.

Amis, R. (2007, May). You can't ignore social media: How to measure Internet efforts to your organization's best advantage. *Tactics*, 10.

Andersen, P. H. (2005). RM and brand involvement of professionals through Web-enhanced brand communities: Coloplast case. *Industrial Marketing Management*, *34*(3), 285–297. doi:10.1016/j.indmarman.2004.07.007

Armstrong, A., & Hagel, J. (1997). *Net gain: Expanding markets through virtual communities*. Boston, MA: Harvard Business School Press.

Awad, N. F., & Zhang, S. (2007, January). *Stay out of my forum! Evaluating firm involvement in online ratings communities*. Paper presented at the 40th HICSS, Waikoloa, Big Island, Hawaii.

Bellman, S., Johnson, E., Lohse, G., & Mandel, N. (2006). Designing marketplaces of the artificial with consumers in mind. *Journal of Interactive Marketing*, *20*(1), 21–33. doi:10.1002/dir.20053

Bickart, B., & Schindler, R. M. (2001). Internet forums as influential sources of consumer information. *Journal of Interactive Marketing*, *15*(3), 31–40. doi:10.1002/dir.1014

Blood, R. (2000). Weblogs: A history and perspective. Retrieved October 15, 2007, from http://www.rebeccablood.net/essays/weblog_history.html

Book, B. (2003, July). *Traveling through cyberspace: Tourism and photography in virtual worlds*. Paper presented at the conference Tourism & Photography: Still Visions - Changing Lives in Sheffield, UK.

Bughin, J. R. (2007, August). How companies can make the most of user-generated content. *The McKinsey Quarterly*, Research in Brief, Web Exclusive.

Crotts, J. (1999). Consumer decision making and prepurchase information search. In Y. Mansfield & A. Pizam (Eds.), *Consumer behavior in travel and tourism* (pp. 149-168). Binghamton, NY: Haworth Press.

Damianos, L., Cuomo, D., & Griffith, J. Hirst, D., & Smallwood, J. (2007, January). *Adoption, utility, and social influences of social bookmarking*. Paper presented at the 40th HICSS, Waikoloa, Big Island, Hawaii.

Dellarocas, C. (2003). The digitization of word-of-mouth: Promise and challenges of online feedback mechanisms. *Management Science*, *49*(10), 1407–1424. doi:10.1287/mnsc.49.10.1407.17308

Du, H. S., & Wagner, C. (2006). Weblog success: Exploring the role of technology. *International Journal of Human-Computer Studies*, *64*, 789–798. doi:10.1016/j.ijhcs.2006.04.002

Dwyer, C. (2007, January). *Digital relationships in the 'MySpace' generation: Results from a qualitative study*. Paper presented at the 40th HICSS, Waikoloa, Big Island, Hawaii. eMarketer (2007a). *Web 2.0 sites draw more visitors*. Retrieved May 2, 2007, from http://www.eMarketer.com.

eMarketer (2007b). *UGC users outnumber creators*. Retrieved July 2, 2007, from http://www.eMarketer.com.

eMarketer (2007c). *The rising roar of word-of-mouth*. Retrieved June 29, 2007, from http://www.eMarketer.com.

eMarketer (2007d). *Reviews boost e-commerce conversions*. Retrieved May 25, 2007, from http://www.eMarketer.com.

Epps, S. R. (2007). *Demystifying tagging for travel sellers*. Forrester Research Report.

Epps, S. R., Harteveldt, H. H., & McGowan, B. (2007). *Executive Q&A: Social tagging for ebusiness. Answers to E-Business professionals' common questions about social tagging*. Forrester Research.

Erat, P., Desouza, K., Schafer-Jugel, A., & Kurzawa, M. (2006). Business customer communities and knowledge sharing: studying the critical issues. *European Journal of IS*, *15*, 511–524.

Farmer, J. (2004). Communication dynamics: Discussion boards, Weblogs and the development of communities of inquiry in online learning environments. In R. Atkinson, C. McBeath, D. Jonas-Dwyer, & R. Phillips (Eds.), *Beyond the comfort zone* (pp. 274-283). *Proceedings of the 21st ASCILITE Conference.*

Forrester Research. (2006). *Social computing*. Retrieved September 14, 2007, from http://www.forrester.com.

Gibbert, M., Leibold, M., & Probst, G. (2002). Five styles of CKM, and how smart companies use them to create value. *European Management Journal*, *20*(5), 459–469. doi:10.1016/S0263-2373(02)00101-9

Gruen, T. W., Osmonbekov, T., & Czaplewski, A. J. (2006). eWOM: The impact of customer-to-customer online know-how exchange on customer value and loyalty. *Journal of Business Research*, *59*, 449–456. doi:10.1016/j.jbusres.2005.10.004

Gwinner, K. P., Gremmler, D. D., & Bitner, M. J. (1998). Relational benefits in services: The customer's perspective. *Journal of the Academy of Marketing Science*, *26*(2), 101–114. doi:10.1177/0092070398262002

Harteveldt, H., Johnson, C. A., Epps, S. R., & Tesch, B. (2006).*Travelers embrace social computing technologies. Guidelines for travel e-commerce and marketing executives and managers*. Cambridge, MA, USA: Forrester Research.

Haven, B. (2007). *Making podcasts work for your brand*. Cambridge, MA, USA: Forrester Research.

Hwang, Y., Gretzel, U., Xiang, Z., & Fesenmaier, D. (2006). Information search for travel decisions. In D. Fesenmaier, H. Werthner & K. Wöber (Eds.), *Destination recommendation systems: Behavioral foundations and applications* (pp. 3-16). Cambridge, MA: CAB International.

Jang, H. Y., Ko, I. S., & Koh, J. (2007, January). *The influence of online brand community characteristics on community commitment and brand loyalty*. Paper presented at the 40th HICSS, Waikoloa, Big Island, Hawaii.

Kim, W. G., Lee, C., & Hiemstra, S. J. (2004). Effects of an online virtual community on customer loyalty and travel product purchases. *Tourism Management*, *25*(3), 343–355. doi:10.1016/S0261-5177(03)00142-0

Lagrosen, S. (2005). Customer involvement in NPD: A relationship marketing perspective. *European Journal of Innovation*, *8*(4), 424–436. doi:10.1108/14601060510627803

Lento, T., Welser, H. T., Gu, L., & Smith, M. (2006). The ties that blog: Relationship between social ties and continued participation in blogs. *Workshop on Weblogging* Edinburgh.

Li, C., & Stromberg, C. (2007). *The ROI of blogging*. Cambridge, MA: Forrester Research.

Lin, Y., & Huang, J. (2006). Internet blogs as a tourism marketing medium: A case study. *Journal of Business Research*, *59*, 1201–1205. doi:10.1016/j.jbusres.2005.11.005

Lin, Y., Su, H. Y., & Chien, S. (2006). Knowledge-enabled procedure for customer relationship management. *Industrial Marketing Management*, *35*, 446–456. doi:10.1016/j.indmarman.2005.04.002

Morrison, A. (2002). *Hospitality and tourism marketing* (3rd ed.). Albany, NY: Delmar.

Murphy, L., Moscardo, G., & Benckendorff, P. (2007). Exploring word-of-mouth influences on travel decisions: friends and relatives vs. other travelers. *International Journal of Consumer Studies*, *31*(5), 517–527. doi:10.1111/j.1470-6431.2007.00608.x

Neus, A., & Scherf, P. (2005). Opening minds: Cultural change with the introduction of open-source collaboration methods. *IBM Systems Journal*, *44*(2), 215–225.

Olshavsky, R. W., & Granbois, D. H. (1979). Consumer decision making: Fact or Fiction? *The Journal of Consumer Research*, *6*, 93–100. doi:10.1086/208753

Paolillo, J. C., & Penumarthy, S. (2007, January). *The social structure of tagging internet video on del.icio.u*s. Paper presented at the 40th HICSS, Waikoloa, Big Island, Hawaii.

Park, D. H., Lee, J., & Han, J. (2007). The effect of online consumer reviews on consumer purchasing intention: The moderating role of involvement. *International Journal of Electronic Commerce*, *11*(4), 125–148. doi:10.2753/JEC1086-4415110405

Pereira, C. S., & Soares, A. L. (2007). Improving the quality of collaboration requirements for IM through social networks analysis. *International Journal of Information Management*, *27*, 86–103. doi:10.1016/j.ijinfomgt.2006.10.003

Pitta, D., & Fowler, D. (2005). Online consumer communities and their value to new product developers. *Journal of Product and Brand Management*, *14*(5), 283–291. doi:10.1108/10610420510616313

Rosenbloom, A. (2004). The blogosphere. *Communications of the ACM*, *47*(12), 31–33. doi:10.1145/1035134.1035161

Rowley, J., Teahan, B., & Leeming, E. (2007). Customer community and co-creation: A case study. *Marketing Intelligence & Planning*, *25*(2), 136–146. doi:10.1108/02634500710737924

Sigala, M. (2003). Developing and benchmarking internet marketing strategies in the hotel sector in Greece. *Journal of Hospitality & Tourism Research (Washington, D.C. Print)*, *27*(4), 375–401. doi:10.1177/10963480030274001

Sigala, M. (2005a). In search of online post-modern authenticity: Assessing the quality of learning experiences at eternalegypt.org. In M. SIGALA & D. Leslie (Eds.), *International Cultural Tourism: Management, implications and cases* (pp. 123–136). Oxford, UK: Butterworth Heinemann, Elsevier.

Sigala, M. (2005b). Integrating customer relationship management in hotel operations: Managerial and operational implications. *International Journal of Hospitality Management*, *24*(3), 391–413. doi:10.1016/j.ijhm.2004.08.008

Sigala, M. (2006). e-Customer relationship management in the hotel sector: Guests' perceptions of perceived e-service quality levels. *Tourism: An International Interdisciplinary Journal*, *54*(4), 333–344.

Sigala, M. (2008, January). Developing and implementing an eCRM 2.0 strategy: Usage and readiness of Greek tourism firms. *ENTER 2008 conference*, Innsbruck, Austria.

Smith, D., Menon, S., & Sivakumar, K. (2005). Online peer and editorial recommendations, trust, and choice in virtual markets. *Journal of Interactive Marketing*, *19*(3), 15–37. doi:10.1002/dir.20041

Wang, Y., & Fesenmaier, D. R. (2004). Modeling participation in an online travel community. *Journal of Travel Research*, *42*(3), 261–270. doi:10.1177/0047287503258824

Winer, D. (2005). What is a. River of News. style aggregator? *Really Simple Syndication*. Retrieved June 22, 2005, from http://www.reallysimplesyndication.com/riverOfNews

Ying, Z., & Davis, J. (2007, January). *Web communities in blogspace*. Paper presented at the 40th HICSS, Waikoloa, Big Island, Hawaii.

Yuan, Y., Gretzel, U., & Fesenmaier, D. R. (2006). The role of information technology use in American convention and visitors bureaus. *Tourism Management*, *27*(2), 326–341. doi:10.1016/j.tourman.2004.12.001

This work was previously published in Information Communication Technologies and City Marketing: Digital Opportunities for Cities Around the World, edited by M. Gasco-Hernandez; T. Torres-Coronas, pp. 221-245, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.5
City Brands and their Communication through Web Sites:
Identification of Problems and Proposals for Improvement

José Fernández-Cavia
Universitat Pompeu Fabra, Spain

Assumpció Huertas-Roig
Universitat Rovira i Virgili, Spain

ABSTRACT

City marketing tries to position cities in the mind of the public, although the process of creating and communicating city brands is still at an early stage of its development. One of the main tools for the communication of these brands is now the World Wide Web. This chapter describes the results of two combined studies (qualitative and quantitative) that analyzes a sample of official city Web sites. The results show that official Web sites of cities give much attention to ease of navigation, but interactivity is much less implemented, especially between users. Furthermore, some lack of attention to the communication aspects of city brands can also be found. Finally, the chapter submits a number of improvement proposals.

DOI: 10.4018/978-1-60566-134-6.ch002

INTRODUCTION

In the current world of cities, competition has increased and the centre of interest has moved to include much broader spheres. Already, cities do not try only to be just significant tourist nuclei, but they also compete in aspects such as quality of life, economic development and sustainability. Aside from tourist interest, cities try to position themselves as comfortable areas to live and important centers of economic development that attract all types of investment.

To achieve this, it is necessary to know the opinions and evaluations of the publics (Prebensen, 2007)[1], to find out what image they have of the city and determine the positioning that it would be desirable to achieve. Therefore, a vision of the city must be formulated and, consequently, a program of identity must be created that is transmitted through a brand and a visual logo and, later, an adequate and effective communication program must be run. With this objective, citymarketing tries to position cities in the minds of the public. In spite of this, the process of creating and disseminating city brands is still in a very early stage of its development.

This chapter tries to show, on one hand, that the concept and application of city brands are still very incipient and, on the other hand, that the official websites, in part due to city brands being underdeveloped, do not pay enough attention to the dissemination of the graphic, functional, and emotional aspects of the brand. With this, there is still a long way to go in the dissemination of cities through their brands on the Internet.

THE CONCEPT OF CITY BRAND

The concept of brand applied to destinations, places or cities is relatively new. It started to spread with the Travel and Tourism Research Association's Annual Conference in 1998 (Blain, Levy & Brent Ritchie, 2005). From that moment, the concept has been developed widely and has been studied from diverse perspectives, especially from the point of view of tourism. All in all, studies on city brands and destinations are still under developed and knowledge on the subject is limited. Some authors consider that it's not correct to talk about branding or place branding in relation to territories, cities or countries. They believe that it is incorrect to associate communicational and marketing terms to realities with their own identity like cities. Nevertheless, the majority or authors appreciate that the territories and the cities do not have the same characteristics as commercial products, but agree that they can apply the same marketing strategies to the territories (Olins, 2002).

The first difficulty we come up against in the study of city brands is the confusion of concepts. Therefore, it is fundamental to distinguish between city brand and brand image (Cai, 2002). Many studies confuse the analysis of the brand image with the brand itself and the branding done by the destination.

The city brand is a new concept and is not very well defined. It is currently very much in fashion, and many people have theories, but few have dared to define it. It is a construct composed of a name, a logo, some symbols and some values that we try to associate with a city, representing its identity, with the objective of creating a position and a vision of the city in the minds of the public. Each city must have its own brand, and each city brand must be the result of a citymarketing plan and a competitive city strategy.

A very complete definition of destination brand, fully applicable to the city brand, which is based on the previous definitions of Aaker (1991) and Ritchie and Ritchie (1998) is that of Blain, Levy and Brent Ritchie (2005), which implies:

The creation of a name, symbol, logo, word mark or other graphic that both identify and differentiate a destination; that convey the promise of a memorable travel experience that is uniquely associated with the destination; and that serve to consolidate and reinforce the emotional connection between the visitor and the destination; that reduce consumer search costs and perceived risk; all with the intent purpose of creating a destination image that positively influences consumer destination choice. (p. 337)

However, the brand image is the result of the branding process, which is the perception created in the minds of individuals. It is "*networks of knowledge elements stored in long-term memory, and the core of such a network is the brand name which is linked to a number of other knowledge elements and/or associations*" (Riezebos, 2003).

As confirmed by Bill Baker:[2] "*A destination without a clear and attractive brand image is like a person without a personality. They blend into the crowd, are seen as uninteresting, and don't get the attention they deserve*".

As consequence of this dichotomy between city brand and brand image, there are two types of studies on the topic. On one hand, those that analyze the brand names themselves, their names, symbols, logos, their identification, purposes and meanings. Our research is included within this first type. And on the other, those that are based on the associations and relationships that the brand names create with the public, that is, the brand image that is created among consumers.

We also consider it is important to distinguish another conceptual aspect. Place branding is not the same as destination branding. Place branding is based on the construction of a global image of the territory that promotes the place in its globality: economical, touristic and as a place of residence. Even in some occasions, tourist branding can be contradictory with place branding. This is why we consider it necessary to give a definition of place branding. It is the sum of beliefs and impressions people hold about places. Images represent a simplification of a large number of associations and pieces of information connected with a place. They are a product of the mind trying to process and pick out essential information from huge amounts of data about a place (Kotler, Haider & Rein, 1993).

Having clarified these concepts, we will concentrate on the city brand, which is the purpose of our analysis. This, like all brands, has its raison d'être, which is based on two basic functions (Aaker, 1991). One of these is the identification of the brand with the town and the attribution of a symbology and some values to the destination. The cities must have new signs of identity, an image and a position. Therefore, the first function of the brand is to attribute functional and emotional values to a city that identify the different cities globally and by consensus.

The second function of the brand is based on differentiating the cities from each other. This has always been the principal mission of all brands. According to the American Marketing Association: "the brand is a name, term, design, symbol, or any other feature that identifies one seller's good or service as distinct from those of other sellers".

Blain, Levy and Brent Ritchie (2005) add three concrete purposes of city brands to the two classics stated previously by Aaker (1991). On one hand, to give the visitors the security of a quality experience at the destination. On the other, to reduce the search costs on the part of the visitors and, finally, to offer a single purchase proposal. But we must be conscious of the difficulties of creating city brands, and the limitations that still exist with respect to their functions.

Moreover, Hankinson (2004) uses the concept of brand networks, in which the destination and city brands have four functions: brands as communicators that represent a differentiation between cities, brands as perceptual entities that appeal to the senses and emotions, brands as values, and brands as relationships.

Having reviewed the theoretical framework of destination and city brands, and defined the concept of city brand, from which the study starts, we will now make a simple classification. As we understand, city brands can be classified according to their degree of evolution and development in the following categories:

1. The graphic brand, which only implies the creation of a symbol and a logo. Logos are the basic element for the creation of a brand and the main vehicle for communicating an image.[3]
2. The functional conceptual brand. This type of brand adds the symbolization of some of the territory's characteristics to the logo, which are real and tangible, and which are to be promoted, being adopted as strong points of the city. These attributes may be: good climate, beaches, nightlife, quality of life and, level of innovation.
3. The emotional conceptual brand, created by a body, entity or public institution, also trying to transmit abstract, symbolic and personifiable values to the city, such as innovation,

> multiculturality, modernization, passion, etc. With this combination of values, the aim is to position and distinguish the image of the city from the competition. Previous studies (Ekinci & Hosany, 2006; Hosany, Ekinci & Uysal, 2006)[4] have shown that the emotional and personifiable values of the destination brands have positive influences in the prior choices of purchase and recommendation of these destinations.

Various authors agree that the brand image of a destination has two basic dimensions (Lawson & Band-Bovy, 1977): cognitive and affective, which would correspond with the emotional and functional conceptual brand. The cognitive component would be made up of the beliefs and knowledge of the physical attributes of a city, the functional conceptual brand; meanwhile, the affective component would refer to the feelings about these attributes, the emotional conceptual brand (Baloglu & McCleary, 1999). From the perspective of understanding the brand image as a "cluster of attributes and associations that the consumers connect to a brand", Biel (1997) understands the existence of "hard" associations, which refer to the tangible and functional attributes, and "soft" associations, emotional attributes. Biel acknowledges that the personality of the brand belongs to the emotional aspect of the brand image. Along the same lines, authors such as Etchner and Brent Ritchie (1991), Kapferer (1997) or De Chernatory and Dall' Olmo Riley (1997) confirm that the brand and its image are composed of two attributes: the functional or tangible, and the symbolic or intangible[5].

However, the emotional conceptual brand must be agreed upon by consensus, created jointly by public and private institutions and citizens of the region, which involves both the internal and external public, and is not identified with or property of a single institution, but of the whole region.

The valid significance of a brand is that registered by its public. Certainly, a brand must be adopted by all the public, starting with the city's own residents, companies and institutions; and for this purpose it is fundamental that these are involved in its creation, that they adopt it as their own and thus help with its dissemination. A study by Blain, Levy and Brent Ritchie (2005), based on interviews with the heads of marketing of the destinations, showed that the opinion of the residents and the visitors must be fundamental in the process of creating city brands.

So, in light of the classification above, an existing city brand may be understood as more or less developed according to its degree of preparation. Some city brands are simply logos that do not represent any specific aspect of the destination, while others have elaborate brands agreed by consensus, which come from a prior marketing plan and represent functional and emotional values that can be identified with the city.

QUALITATIVE ANALYSIS OF CITY BRANDS

Having stated the definition, functions and classification of city brands, we performed an initial qualitative study on eight brands of important world cities on their official websites. The study showed the limitations of city brands in their current process of creation, implementation and dissemination.

The objective of the investigation centered on finding out the degree of evolution and development of a sample of city brands, and their dissemination through the websites of their official institutions. The analysis consisted of two stages. The first centered on the prior examination of the city brand itself and its degree of development. The method used in this part of the study was based on the classification of the degree of advance preparation and evolution. In other words, in the analysis of the graphics and logo, the functional and emotional values assigned and the knowledge of the moment of creation, motivation and agents involved in the process of creating the city brand.

The second phase of the investigation centered on the study of how the city brands were treated on the official websites of the destinations. The analytical method used was the BIWAM (Brand Identity Web Analysis Method). This is a technique for qualitative analysis of the establishment of brands on the web, created by Martín Barbero and Sandulli (2005), which includes eight dimensions of analysis, of which we applied six to our study of city brands:

1. Analysis of the Appearance, which corresponds to the strong, real and objective points of the destination (the functional element of the brand) and how it is communicated on the web.
2. Analysis of the Personality. This implies the assessment of the symbolic and emotional elements that are attributed to city brands and their treatment on the websites.
3. Analysis of Humanity. This refers to the interactivity of the page.
4. Analysis of the Style. This analyses the graphic part of the brand, specifically, the relationship of the logo with the colors and the typography of the website.
5. Analysis of the Medium, and the communicative functionality of the websites.
6. Analysis of the Credibility. This refers to errors, slow loading speed, internal coherence, etc.

We only applied six dimensions of analysis created by Martín Barbero and Sandulli (2005) because the other two could be applied to product brands, but no to city brands.

The sample consisted of eight city brands that correspond to international tourism capitals, which have created their city brands, but show different degrees of evolution according to the typology stated above. The city brands selected for the sample were: Barcelona, Madrid, Edinburgh, Amsterdam, Cincinnati, Toronto, Dubai and Hong Kong.

RESULTS OF THE QUALITATIVE STUDY: EVOLUTION AND LIMITATIONS OF CITY BRANDS

The results of the study showed that city brands are a concept of recent creation and still incipient development. Actually, the majority of brands analyzed in the study were created in 2005. Many large world cities still have not created their city brands. And among those that have, the degree of evolution of their brands is still mostly in the initial stages. The vast majority are stuck in the stage of creating logos based on strong points or characteristics of the city that they wish to boost. However, very few cities try to identify themselves with an emotional conceptual brand, based on the appropriation of personifiable values and the creation of a city marketing strategy.

Our results coincide with those of a study by ESADE (2004) on the evolution of the positioning of Spanish tourist destinations and their tourism brands. This stated that the brands, as strategic realities of tourist destinations, are usually fairly general, based on functional values without dealing with emotional aspects. They do not segment their range much and are only transmitted externally, forgetting the internal public, and do not evolve with the passage of time.

Other studies in tourist marketing have shown that, in general, the application of marketing techniques in destinations is still scarcely developed (Gnoth, 1998; Pritchard & Morgan, 2002).

The causes of the lack of evolution of city brands are related to a series of limitations the brands experience in their implementation. The first consists of the complexity of combining a segmentation strategy with the creation of a single brand image. The cities are directed at diverse sectors of the public (citizens, investors, businesspersons, tourists) with whom they wish to communicate. As each sector of the public has certain interests and certain needs, the cities generally create different marketing strategies for each of them. For this reason, it may be complicated, or even contradictory, to create different strategies and integrate them into one single brand positioning.

Another difficulty is found in the existence of more than one brand per city. It often occurs that different institutions create city brands for their websites or for independent use. The result is dispersion, incoherence and the impossibility of creating a single image that is recognizable and adopted throughout the community. For example, the City Council of Madrid has a municipal website (munimadrid) with a brand and a logo that has nothing to do with the Madrid brand on the municipal tourism portal (esmadrid), or with the website of the region (turismomadrid). It is easy to understand that it is absolutely essential to make the effort to coordinate and negotiate when creating a single brand that is not property of one institution in particular, but of all the citizens, and applicable to all the websites related to a destination.

With respect to the second function of the brand, that of differentiating cities from each other, it is possible that different cities try to identify themselves with the same values. The research that we have performed shows that the emotional conceptual brand, that which attributes some personifiable values to a city with the purpose of differentiating it from other cities, is usually created in a very broad and ambiguous way, and this does not fulfill its differentiating function. The majority of cities do not identify with a single value, but with many, some of which are shared by different brands, which, in fact, encourages confusion.

Kotler (1993) already stated that all images of a destination must be simple and distinctive. The main function of a brand must be to differentiate it with respect to the competition. However, in the study it was demonstrated that the majority of cities prefer to be identified with diverse values or characteristics at the same time, perhaps to be attractive to more sectors of the public, maybe because they are values that are currently very attractive in society and which they do not want to renounce although they have been adopted by other destinations. This is shown in the eight city brands analyzed, Amsterdam and Toronto identify themselves as creative; Barcelona and Dubai, adventurous; Barcelona and Edinburgh, friendly; Edinburgh and Toronto, imaginative; Barcelona, Amsterdam and Edinburgh, diverse; Hong Kong, Barcelona and Amsterdam, cosmopolitan; and finally, Barcelona, Amsterdam, Edinburgh and Toronto, innovative. In this sense, the city brand completely loses its distinctive or differentiating function and this limits the creation of a single image for every destination. Similarly, Morgan, Pritchard and Piggott (2002) also showed that the images created of destinations and cities are not different and do not usually contain a single idea or single purchase proposition.

Our research also revealed that the majority of city brands have been created according to a specific event. For example, Madrid's brand was created to promote the destination internationally when the city was selected as a possible candidate for the Olympic games of 2012. Barcelona, however, created its brand to promote the Forum 2004, and Cincinnati, after the results of an economic study, took notice of the need to connect the three States of Ohio, Kentucky and Indiana, and as a result, created its brand with this unifying objective. This fact may also be negative for city brands, which should never be associated with a specific political event, as they must represent a city and not a municipal government, a social or sporting event or any private interest.

The motive that generally moves public bodies to create city brands is mainly based on tourist or economic interests, and thus the brand created usually is identified with these interests. This is an error, a reductionist conception that limits the potential of the city brand. In addition, on occasions, after an election or change of party in the town halls, the city brands and their representation and meanings are changed, precisely with the intention of breaking away from the previous image and promoting a new improved image related to the political party that has formed the municipal government. These changes create dysfunction, as the brands need time to be implemented and require their evolution to be homogenous and coherent. The persistence and durability of a brand is key for its implementation and acceptance by all sectors of the public. Therefore, brand changes only create more confusion in the identities and images of the cities.

Finally, but importantly, difficulty in the creation of city brands lies in coordination, taking into account the umbrella brands of destinations greater than the cities. In tourism, the broader destinations (regions, nations, states or countries) include those within them in their brand for the promotion of tourism. So this presents us with a number of questions: Should the umbrella brands take into account and be coherent with the city brands that they include? Should they all be related? Should they be coherent with each other? Should they have common features? Should the attributes of the umbrella brands be shared by the brands of the respective cities that are represented?

TREATMENT OF CITY BRANDS ON THE WEB

The Internet and new information technologies play a key role in communicating the cities and their brands. They are an important source of information. Destination Management Systems are more than simple websites. In addition to the information, they offer advertising, marketing and sales applications, and have interactive resources that, in an entertaining way, provide services and attract the attention of the users.

Currently, through a city's tourism website, you can get information, make reservations, etc. However, in the promotion of the cities, not only as tourist destinations, but also as business centers and residential areas, portals or broader websites are starting to be created, which offer, in addition to tourist information, business and leisure information for the citizens.

In this sense, and from the field of communication, the brand websites have been marked as the future of marketing communication on the Internet, as they have the potential to provide high levels of information and, in addition, create virtual product experiences (Klein, 2003). Brand websites are capable of combining both of the basic objectives of commercial communication in this channel: to create a brand image and achieve a direct response (Hollis, 2005). As Cho and Cheon (2005) describe, the websites may serve for diverse communication purposes: public relations, sales promotion, advertising or direct marketing.

In the second part of the study, the results of our analysis of city brands on the web showed that the aspects that make up the corporate image, that is, the colors, the lines and the logo of the city brand, in general, are used very little to create graphic coherence and brand image throughout the website. There are some exceptions, such as the sites of Amsterdam or Madrid, that show effective graphic coherence for transmitting the brand, but habitually the typography and the colors of the logo are only used in auxiliary hyperlinked pages, but not the whole official site.

The study showed that what is best transmitted through the web is the functional conceptual brand of the cities, the strong points or the potential that they wish to promote; but in no way the emotional conceptual brand, which ascribes personifiable values to the destination. The exception is a hyperlink that some official websites have, such as Edinburgh or Amsterdam, which links to a page exclusively dedicated to explaining the emotional brand.

Starting with the results of the studies of Hosany, Ekinci and Uysal (2006), where it is demonstrated that the emotional and personifiable values have positive influences on the intention to visit, purchase and recommend destinations on the part of consumers, it is surprising that it is actually the emotional aspect and the personality of the brand that are the least developed aspects on the official websites of the cities analyzed. The heads of marketing of the cities should develop strategies and campaigns that promote the distinctive personality of the destinations, based on the emotional components of these cities, which create better positioning and a more favorable image among users.

Thus, in conclusion, the under use of websites to promote city brands has been verified. The websites centre on the functional conceptual aspects of the brand, that is, all the strong points that are notable in the city, but lack, in general, coherent graphic treatment and the expression of the emotional conceptual brand. The design of the websites tries to be useful and functional to provide the users with the information and services they wish to obtain, but they are not at all creative to disseminate the emotional values attributable to the city through its brand. McMillan (2004) coincides with our statements, arguing that advertising on the Internet and websites must be more creative, that is, better designed, with greater impact, more varied and more entertaining.

Once at this point, and in light of other studies and a bibliography centered on more technical questions of the websites, we decided to analyze other characteristic aspects of the websites that also influence the dissemination and perception of city brands. We considered that dealing with city brands should not be limited to an analysis of the websites' content only, but should take into account aspects such as the interactivity or usability of their pages.

Different research on websites shows that usability is a key aspect in the creation of a good brand image. The sites that seem to be or are easier to open, navigate or use, create a more favorable attitude and image among users (Chen & Wells, 1999; Chen, Gillenson & Sherrell, 2002; Heijden, 2003). Thus, small websites, with very basic iconography and ease of use, such as that of Barcelona, would transmit a good city brand image. However, extensive and complicated sites that are slow to open and that have a confused internal structure, such as that of Toronto, cause less positive or even unfavorable attitudes to be created.

Along the same lines, Jared M. Spool (1996)[6] did a study comparing websites, which demonstrated that the usability of a website considerably and positively affects the brand and the branding process. His results showed that the users that navigate more easily through a website and find the information that they want quickly end up with a better impression of the brand, as it has satisfied their expectations to a greater degree. Contrarily, the obstacles that the users find when navigating negatively and directly affect their perception of the brand. Therefore, usability is essential for effective branding.

Regarding interactivity, Liu (2003) defends the idea that the concept of interactivity unites three correlated but different factors: the active control of information, bidirectional communication and the synchronicity or simultaneity of communication. In a previous article (Liu & Shrum, 2002), the same author classified the brand websites in the maximum range of the three factors stated. This study showed, by the way of bidirectional communication, that the Internet is the only medium that can be used for commercial transactions without the help of other tools, since necessary activities such as showing the product, placing orders, making payments or even, in categories such as music, software or transport titles, distributing the product can happen through the web.

Other authors (Cho & Cheon, 2005) prefer to divide the concept of interactivity into three fields of action or types: consumer-message interactivity, consumer-consumer interactivity and consumer-marketer interactivity. Consumer-message interactivity refers to the capability of the user to personalize his or her relationship with the contents of the page according to his or her interests and motives. Consumer-marketer interactivity centers on the communication between the user of a website and the organizers or those responsible for the content; this relationship may be bidirectional, from user to administrator (questions, suggestions, complaints) or from administrator to user (obtaining personal data, answering questions, etc.). Consumer-consumer interactivity is the relationship that may be created between the people that access a website (virtual communities, chats, forums, etc.).

More recent studies (Sicilia, Ruiz & Munuera, 2005; Ko, Cho & Roberts, 2005) show that interactivity enables the information to be processed better and generates more favorable attitudes towards the website and towards the product and the brand, and greater intention to purchase.

Based on these previous studies on the usability and interactivity of websites we decided to make a broader quantitative analysis about the treatment of city brands on the Internet. In addition to the items related to graphic, functional and emotional aspects of the brand, in the trial we analyzed usability and interactivity characteristics of the websites as elements that also influence the dissemination of a good brand image.

PURPOSE AND METHOD OF QUANTITATIVE STUDY. USABILITY, INTERACTIVITY AND THE CITY BRAND ON THE WEB

The first qualitative study was wider. It analyzed the city brands and their treatment in websites in depth. But later, we decided to carry out a quantitative study analyzing more webs and variables. We had to develop a different questionnaire of analysis that could be measured by quantitative methods.

The main objective of this quantitative study consisted in analyzing the degree of usability, interactivity and treatment of city brands on the official websites of tourist cities.

For this empirical study, we used a quantative method based on a WTO (World Tourism Organization, 1999) analysis model. This was extended with the contribution of recent studies (McMillan, 2003; Liu, 2003; Cho & Cheon, 2005) and adding newly created interactive resources, which appeared as new features on destination websites at the time the study was carried out. At the same time, the model was also extended with aspects to analyze about how city brands are dealt with on the web, the databases that the websites may obtain about their users, and other aspects of website information not considered in the initial model.

The analysis was applied to 40 official websites of important tourist cities of the five continents during 2006. This sample was selected by means of a ranking evaluating the main tourist cities of the world, taking into account the *World's Top Tourism Destinations* of the WTO (World Tourism Organization), the number of visitors and the importance of the city as a tourist destination.

We observed 135 indicators, of which 87 are representative of the three variables analyzed: usability, interactivity and the brand. Interactivity is analyzed in the three factors mentioned previously: consumer-message interactivity, consumer-marketer interactivity and consumer-consumer interactivity. Its analysis, through the SPSS program, was centered on descriptive statistics and the combination of variables, using Gamma as a correlation index.

RESULTS OF THE QUANTATIVE STUDY. USABILITY

The concept of usability is defined as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"[7]. This ease of use, nevertheless, is related to very diverse aspects, that go from the page design, to the content quality, the ease of locating information and the simplicity of navigation, all

Table 1. Items of usability (Source: Authors' compilation)

Indexes of usability	Frequency	Percentage
List of contents on all pages	36	**90.0%**
Link the home page	37	**92.5%**
External links related sites	38	**95.0%**
Up-to-date information	39	**97.5%**
Sitemap	23	**57.5%**
Indication of the navigation path	16	**40.0%**

Table 2. Ranking of websites with the most and least points with respect to usability (Source: Authors' compilation)

Web sites with high scores in structure, design and usability		Web sites with low scores in structure, design and usability	
Score	City Web sites	Score	City Web sites
11	London	5	Cairo
11	Hong Kong	5	Beijing
10	Bangkok	6	Lisbon
10	Budapest		
10	Sydney		
10	Amsterdam		
10	Tokyo		
10	Buenos Aires		
10	Mexico City		

of which are related, in addition, to the subjective perceptions of the user.

The indicators used in the study to measure the usability variable are those that provide the navigability for the website: the access menu for the sections always being visible, the indication of the navigation path, constant links to the home page, the existence of a sitemap and an internal search engine; the possibility of user help by telephone, e-mail, chat or web call; and external links to related sites.

The majority of sites analyzed in the study showed high levels of structure, which means that the websites are, in general, well designed. At a global level, high percentages were seen for the variables that make up the structure, usability and design of the websites.

As seen in Table 1, the majority of the sites analyzed have a list of the site contents that appears on all of the pages, along with a link to the home page from each page. These two resources give the user his or her location on the web, clarify the structure of the site and simplify its use. In addition, almost all the websites have external links with other related sites. Through these links they can provide complementary information and services. In some cases connection to the official sites of the town halls and institutions are provided, in others, the purchase of services, for example the link on Barcelona's website to ServiCaixa, through which you can purchase tickets for entertainment.

With respect to the help that the websites offer their users, 90% of the sites analyzed provided e-mail contact. In addition, 65% provided a telephone helpline. But, however, only one website offered help by chat and none used web calls (calls from marketing staff from the destinations where you request them to call you at a certain time) or human clicks (communication in real time of a visitor to a website with its administrators).

All in all, the results stated up to now show that the majority of websites are well structured and designed. Therefore, they have a high degree of usability.

Table 2 shows the rankings of the websites with best structure, design and usability, along with those that have the least points in these aspects. The structure variable ranges between 5 and 11, with sites with 5 having the worst structure and those with 11 the best.

The initial Hong Kong web page (Figure 1), one of the best in usability of our study, is simple, graphic and structured. It shows only a big image and the list of different languages the user can choose.

When you enter in the initial page of the selected

Figure 1. Hong Kong Web site

language you can observe that it has the content list in all the pages, visual and graphic links and shows a constant simplicity in the entire site.

Interactivity

The interactivity variable, as explained previously, was analyzed using three aspects: consumer-message interactivity, consumer-marketer interactivity and consumer-consumer interactivity. We measured the concept of interactivity using the three aforementioned typologies of Cho and Cheon (2005).

Consumer-message interactivity is measured with indicators such as: the existence of search engines, user help, the option to customize the display, displaying virtual reality, multimedia presentations, directional maps, virtual leaflets, the option of downloads to mobiles, and on-line games, among others. However, consumer-marketer interactivity is based on indicators of relationships with the marketing organizers of the cities. They offer the possibility of questions and complaints from users, opinion surveys, chats with promotional or sales agents, bulletin boards for users, or the possibility of placing orders and making reservations on line. Finally, consumer-consumer interactivity is measured using indicators of the relationship between them, such as the existence of chats or an email service.

Consumer-Message Interactivity

The first result we can see is that, in general, the majority of websites analyzed use many more resources that belong to the consumer-message interaction, than to the consumer-marketer, and consumer-consumer interaction. Therefore, the sites offer the navigator a greater interactivity with the messages that they wish to transmit than with the people that are in charge of marketing the destinations and other consumers. And within the resources of the consumer-message interaction, the most used in all the websites, that is, those that show the highest percentages, are the interactive travel organizers.

So, this shows that the majority of websites analyzed offer services of interactive travel organizers. We refer to organizers without the possibility of purchase. They help users to plan their trips by providing fully personalized information adapted to the needs and interests of the users. The interactive travel organizers allow the users to plan their trips in a fully complete and personalized way, from their transport to the destination to their accommodation and other complementary tourist services.

However, there are other interactive consumer-message resources that are hardly used by the websites analyzed. Certainly, 62.5% have directional maps of the cities to orient the users and 82.5% have geographical markers on the maps with zoom, which offer the possibility of searching for and locating elements on the map in detail. But only three of the sites allow the user to customize the home page, four show virtual excursions or online games for children, five offer virtual flights of the city from the air or allow users to create their own virtual leaflets in folders, which they can save with the personalized information that interests them.

The informative services using optional downloads to mobiles, whether metro maps, information on monuments or audio downloads are also resources that are barely used by the websites analyzed.

All of these interactive resources, in addition to offering information, provide a certain entertainment and distraction to the users of the websites, making them more attractive and interesting.

Consumer-Marketer Interactivity

The resources of the consumer-marketer interaction are still less used by the websites analyzed that those of consumer-message interaction. The possibility of placing orders or online reservations is offered by 35% of the sites, and 32.5% allow the user to track the orders placed. For these two resources, the prior registration is usually required of the user, and with this the websites can obtain personal information about their consumers. In 22.5% of the sites, users can fill in surveys or opinion polls and 12.5% offer the option of complaints by consumers.

The rest of the consumer-marketer interactive resources are barely used by the websites analyzed. For example, only Rome's site offers the users the possibility of asking what they want about the city, the services offered or the entity that organizes the marketing of the destinations. Only the websites of Dublin and Istanbul have an electronic bulletin board available to the users. And finally, the sites of Madrid, Amsterdam and Montreal are the only ones that allow the consumers to sponsor the website. This means that any organization can pay some money to sponsor the website. All the conditions are very well explained on the site. In exchange, the sponsor can put its logo on the website.

Not one of the websites analyzed offered the users the possibility to propose new products or services, chats with the marketing agents, or "call me" buttons with time and language selection for those responsible for promoting the city to call the interested users to provide them with the information they want, personally by telephone.

Consumer-Consumer Interactivity

With respect to the consumer-consumer interaction, there are even fewer resources available. The resource of this type that is most used in the sites analyzed is the on-line postcards service, featuring in 40% of the websites, of which 37.5% do not require registration. This resource, which is offered more and more by the destination websites and is mostly used by young users, enables them to send on-line postcards, without the delay or costs of sending them.

Table 3. Resources used in the websites of interactive travel organizers (Source: Authors' compilation)

		Frequency	Percentage
Interactive travel organizers (without possibility of purchase)	How to get there	33	**82.5**
	What to do	39	**97.5**
	Attractions/events	38	**95.0**
	Leisure activities	38	**95.0**
	Cultural activities	37	**92.5**
	Where to stay	38	**95.0**
	Transport	39	**97.5**
	Excursions	33	**82.5**
	Rentals	**16**	40.0

Moreover, only Krakow's site has a chat for users, and the possibility for consumers to tell their stories, experiences and summaries of trips is only found on the sites of Hong Kong, Rome and Toronto. Finally, none of the websites analyzed offer the services of a cyber club of users with advantages or a cyber community with common interests. Cyber clubs which offer some advantages of information, discounts and special conditions to the loyal users. The cyber club of users would allow the marketing organizers of the destinations to create loyalty programs for clients using the offer of advantages and discounts. Moreover, the creation of a cyber community would enable the users of this group to build relationships with each other, broaden information in their interest, and create a strong position with respect to these common interests.

So, we can see that there is a great inequality in the use of interactive resources by official websites of destinations. The most used are those that belong to the consumer-message interaction, and especially, the interactive travel organizers. All in all, generally, the percentages of use of interactive resources are still underused, in particular those of the consumer-marketer and consumer-consumer interaction. These results coincide with those of the study by Anton (2004) about the Internet presence of the main tourist destinations of the Spanish Mediterranean coast. In it, he highlights the absence of instruments that promote interactivity in the websites, such as on-line forms, pages of visitor comments, etc. (Table 3)

The degree of global interactivity of the websites analyzed, which may range between 4 and 28, can be seen in the ranking in Table 4. These numbers represent the number of interactive tools they use.

Visit Dublin's web page it's the most interactive out of all the analyzed sites. If we only observe the first part of the page we can find a searcher and in the graphic frontal there is the option to click and see a video about the city. The whole page is very interactive.

In the accommodation section the user can search information, but can also reserve and buy the products.

Even the maps of the city and the maps of public transport are interactive. They show the routes that the user requires in personalized way.

Treatment of City Brands on the Web

In the analysis of the communication and dissemination of city brands through websites, it is necessary to differentiate the concepts of city brand and brand image. As previously explained, the city brand is a construct composed of a name,

Table 4. Ranking of websites with the most and least points with respect to interactivity (Source: Authors' compilation)

Most interactive websites		**Least interactive websites**	
Score	City websites	Score	City websites
28	Dublin	4	Moscow
23	Hong Kong	4	Kiev
21	Valencia	7	Sydney
20	Madrid	8	Cairo
18	Berlin	8	Seville
18	Istanbul		
18	Rome		

Table 5. Ranking of the best and worst treatment of city brands in websites (Source: Authors' compilation)

Websites that deal with their city brands best		**Websites that deal with city brands worst**	
Score	**City websites**	**Score**	**City websites**
10	Amsterdam	3	Kuala Lumpur
8	Hong Kong	5	Madrid
8	Athens	5	Zagreb
8	Dublin	5	Cairo
8	Toronto	5	Sydney
		5	Moscow
		5	Lisbon
		5	Montreal
		5	Rome
		5	Mexico City
		5	Kiev

a logo, some symbols and values that we try to associate with a city representing its identity, with the objective of creating a positioning and a vision of the city in the minds of the public, which must be based on a citymarketing plan and projected through a communication program. On the other hand, the brand image is that which is created in the minds of the public as a consequence of the communication of the city brand, in conjunction with the perceptions and subjective values of individuals. (Table 5, Figure 2, Figure 3, Figure 4, Figure 5, Figure 6)

In the study, the degree of communication and dissemination of the city brands through the websites was measured using the presence of the brand and the logo, whether there is a description of the city brand, whether the graphic and photographic images represent the brand, whether the colors and the typography of the page are coordinated with the logo, and the functional and emotional brand is disseminated.

The results show that city brands are treated poorly on the websites analyzed. Their treatment is even less developed than that of interactivity.

All in all, the vast majority of websites (97.5%) have the logo on all of the pages, and this is usually situated in the upper left part. However, it must also be taken into account that three of the websites analyzed show more than one logo for the same city brand, which creates dysfunctionality and incoherence.

Of the websites, 92.5% offer a brief description of the destination, but only 5% (two of the sites analyzed) have a minimum explanation of the city brand, of its creation and symbolism. Similarly, only 35% of the websites state the marketing objectives of the tourist institutions of the cities.

The graphic images, in general, also are an underused resource in the dissemination of the city brands. All the websites transmit the functional brand through photographic images, but only half of them (50%) communicate the emotional brand through the photographs. On the other hand, the rest of the graphic images are not used to disseminate the brand at all.

Finally, the graphic and typographic coherence between the brand and the web has not been taken into account either in the majority of websites

Figure 2. Visit Dublin Web site

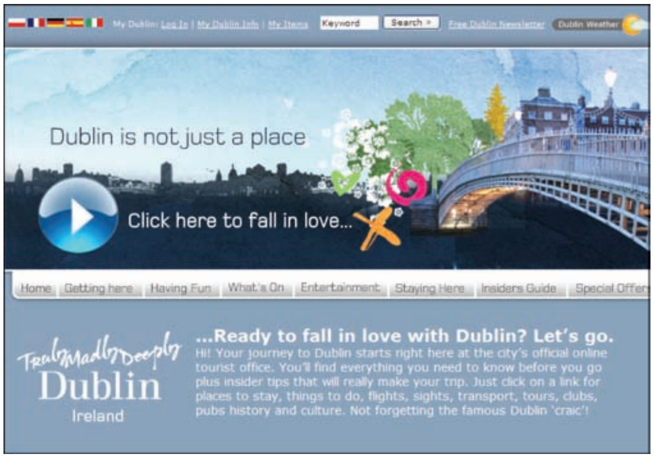
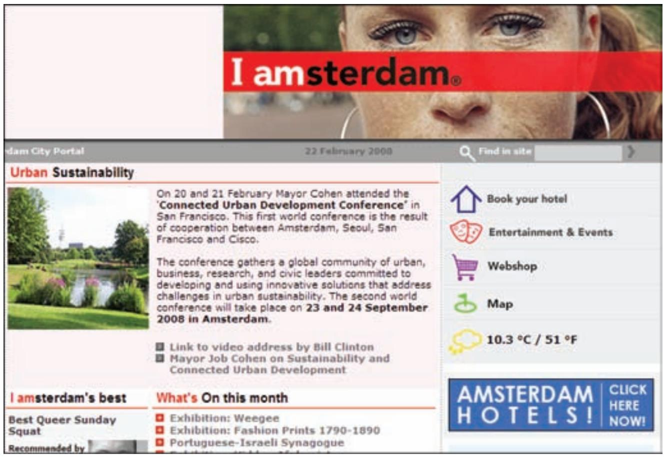


Figure 3. Visit Dublin accomodation section

Figure 4. Visit Dublin maps section

Figure 5. I Amsterdam Web site

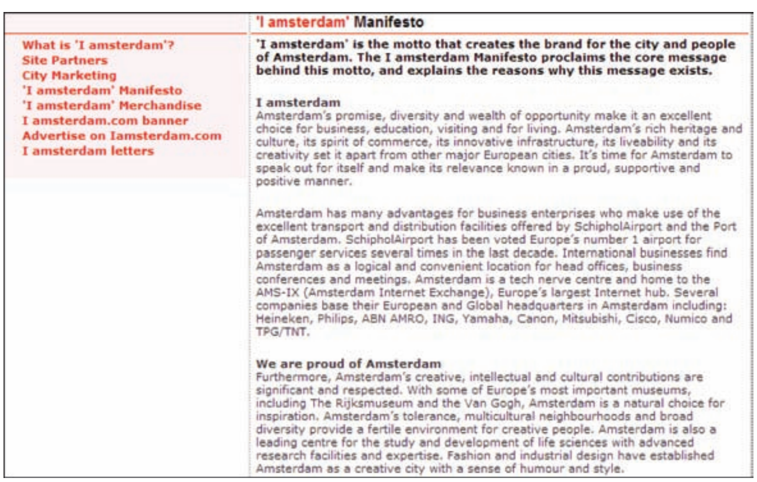


analyzed. Only 27.5% of the sites have the predominant colors of the website coordinated with the logo, and 23.5% use the same typography on the page and the brand. On the other hand, only one of the sites analyzed includes elements of the advertising campaign transmitted through the conventional media.

The degree of global treatment of the city brands on the websites analyzed, which ranges between 3 and 10, can be seen in the ranking in Table 5.

The web of Amsterdam, which best disseminates the city brand, does not only deal with the functional and emotional brand on the web, but in addition it pays attention to many other graphic and visual aspects. For example, the emotional brand, which is disseminated through very few websites, in that of Amsterdam it is dealt with both by changing photographs on the page, which show people of the city, and by an introductory page that explains in depth what the brand symbolizes. In it, it explains what the brand "Iamsterdam" is, who

Figure 6. I Amsterdam manifesto page



created it, when and why; and what it symbolizes and means. As can be seen in the home page, the logotype is present in the entire site. The brand is based on the idea that the value of Amsterdam is their people. It plays with 'I am sterdam' and 'I am Amsterdam'. The photographs of the site try to communicate these emotional values through different types of people of differing ages, race and profession instead of showing monuments or landscapes and scenery.

But the official site of Amsterdam also pays attention to other aspects related to the brand, such as the typography. It uses the same combination of colors and type of letters as the logo in the whole website, creating a graphical coherence that boosts and constantly reminds of the city brand.

Finally, this web page contains information on the city marketing, their objectives and strategies and also an explicit explanation of the brand (which you can read in the I Amsterdam manifesto), advertising of the city and the merchandising of products.

CONCLUSION

The study indicates that the majority of websites analyzed have high usability indexes. Therefore, the official websites of cities pay a lot of attention to ease of navigation, so that the user can easily navigate and find the information he or she wants.

Contrarily, interactivity is much less implemented in the websites analyzed. The interactive resources that are most used are those that correspond to the consumer-message or consumer-marketer relationship, despite being generally underused.

So, it can be confirmed that the websites analyzed in this study have a better structure, design and usability, than interactivity. The websites are more usable and structured than interactive. And, in addition, there is no correlation between these two variables. Therefore, the interactive capacity of the websites may be higher than now without affecting the other variables at all.

As a consequence of the under use of interactive resources, the possibilities of user records, contained in the interactive resources, also decrease. This means that the information obtained on the part of the website users is minimum. As has been previously confirmed, the possibility of obtaining more information for more direct one-to-one marketing that would produce special offers according to the needs of the consumers is being wasted.

On the other hand, there is even less treatment of the brands on the websites analyzed. Only the graphic images, basically the logo, are disseminated and in part the functional brand. However, very few communicate the emotional brand. Similarly, few websites show graphic and typological coherence with the logo, or a relationship with the advertising or the marketing objectives of the destinations.

These results show that, in general, there is a certain lack of attention to the communicative aspects in the construction of the city websites. They seem to be made by information technologists, who pay great attention to usability, and also, to a lesser extent, interactivity, but do not take into account the communication of the brand. Therefore, you can sense the existence of a certain lack of co-ordination between the IT and communication and marketing departments of the entities promoting the cities.

To effectively create websites that disseminate city brands and promote the image of the destinations, it is fundamental to be conscious of what an important tool for image, communication and marketing these websites are for the destinations and cities.

As we go further into the Information Era, the role of websites is changing. They are evolving from being merely sources of information, that is, intermediaries between tourists and destinations, to being involved in tourist transactions. Therefore, tourism and Internet make an ideal combination. The Internet provides the users with a way of obtaining much more varied and detailed information on the destinations and the cities than that which existed before. However, in addition it enables them to, through the same web space, make consultations and reservations quickly and easily.

The websites are considered the future of communication on the Internet, as they offer a large quantity of information and, in addition, create virtual product experiences (Klein 2003). Websites, on one hand, create a brand image and, on the other, can provoke a direct response (Hollis, 2005). As Cho & Cheon (2005) confirm, websites offer different communication possibilities: public relations, sales promotion, advertising, direct marketing and brand image creation. As a consequence, this communicative potential must be promoted and made use of in the field of cities.

With respect to how little city brands are dealt with on the websites analyzed, it is due, firstly, to insufficient conceptualization, creation and development. As we have already stated, city brands are a recently created concept that is still at a very incipient stage. For this reason, the web deal poorly with them. Firstly, it is necessary to create elaborate city brands, based on a strategy of citymarketing, to be communicated later to the public and, if not, disseminated through the websites.

Despite the studies mentioned above (Spool, 1996; Chen & Wells, 1999; Chen & al., 2002; Heijden, 2003) that show the importance of usability in promoting a more positive brand image among users; and the effect of interactivity, which produces more favorable attitudes towards the web, the brand and greater intentions of purchase (Sicilia & al., 2005; Ko & al., 2005), we can confirm that high indexes of usability and moderate indexes of interactivity do little good in the websites analyzed in the promotion of city brands. First it is necessary to correctly create the city brand, to later disseminate it through the websites of the cities. And only after this first step, will usability and interactivity be complementary and effective aspects in the promotion of city brands.

Improvement Proposals

The creation of the city brand must not be an act restricted to the activity of citymarketing. Exactly the opposite, it must be coherent with the whole marketing plan and be derived from a competitive city strategy.

It must start with a prior diagnosis of the current situation of the city image and the competition, to later formulate the vision of the city and the positioning that is desirable to achieve in the world urban system.

Only after the two first stages should an identity program for the city be determined, which must take into account a historical analysis, and the perceptions of the internal and external public. Then the visual identity of the city, its logo, colors, typography should be determined; along with the functional and emotional values of the brand. In this whole process it is essential to identify the internal and external public (citizens, visitors, investors, mass media, public institutions, neighborhood associations), to address them directly and achieve that they take on the identity and the city brand to be promoted as their own and, if possible, this should be by consensus.

Finally, the city brand they must be disseminated by a communication plan to each sector of the public through all types of actions *above* and *below the line,* with the intention of continuity in time, for the values of the brand to catch on coherently in the perceptions of all their sectors of the public and be integrated in the rest of the citymarketing actions.

But city brands must not only comply with a creation process inserted into the citymarketing actions, they must also have certain characteristics:

1. The city brand must be made up of three basic elements: the graphic brand, the creation of a symbol and a logo; the functional brand, based on the real strong and attractive points of the city; and the emotional brand, the symbolic and personalizable values that are associated with the city. Without one of these three elements, the city brand is incomplete, losing its identificative and persuasive power.
2. A single city brand must be created, with a single strategy, that is applicable to the diverse sectors of the public, but that at the same time identifies the brand with unique values, which allow the cities to de identified and differentiated from each other. In addition, this brand must be adopted by all the organizations and entities, avoiding the proliferation of several brands that create dysfunctionality.
3. Finally, and independently of the municipal political changes, city brands must be created to last a long time. If they are agreed upon by consensus when created, and therefore, do not belong to the municipal governments, they will evolve independently of the political channels and have a greater possibility of being consolidated in the minds of the public.

Once the city brand is completely and adequately created, and starting with the importance of websites as tools for communication, promotion and marketing of the cities, attention must be paid to three key aspects in the promotion of city brands through the Internet:

1. The treatment of the city brand through the constant and unique presence of the logo, with an explanation of its symbology and objectives, with the representation of the functional and emotional values that are desirable to associate both textually and as graphically, and by means of the graphic and typographic coherence throughout the website.
2. Maximum development of usability throughout the page, which facilitates user navigation and promotes a positive image increasing the possibilities of recommendation and marketing.
3. Maximum creative use of the interactive resources, as these improve the brand image and the users feel drawn to navigate.

REFERENCES

Aaker, D. (1991). *Managing brand equity*. New York, NY: Free Press.

Anton Claver, S. (2004). La presencia en Internet de los principales destinos turísticos del litoral mediterráneo español. *Actas del Congreso TURITEC 2004*. Universidad de Málaga, 2004. Retrieved June 1, 2008, from http://www.turismo.uma.es/turitec/turitec2004/index.htm

Baker, B. (2007). Places: The new brand frontier. *Total destination management*. Retrieved June 1, 2008, from www.DestinationBranding.com

Baloglu, S., & McCleary, K. (1999). A model of destination image formation. *Annals of Tourism Research*, *26*, 868–897. doi:10.1016/S0160-7383(99)00030-4

Biel, A. (1997). Discovering brand magic: The hardness of the softer side of branding. *International Journal of Advertising*, *16*, 199–210.

Blain, C., Levy, S. E., & Brent Ritchie, J. R. (2005). Destination branding: Insights and practices from destination management organizations. *Journal of Travel Research*, *43*, 328–338. doi:10.1177/0047287505274646

Buhalis, D. (2002). *eTourism. Information technology for strategic tourism management*. UK: Prentice Hall.

Buhalis, D., & Costa, C. (2006). *Tourism, management dynamics. Trends, management and tools.* Oxford: Elsevier Butterworth Heinemann.

Cai, L. (2002). Cooperative branding for rural destination. *Annals of Tourism Research*, *29*, 720–742. doi:10.1016/S0160-7383(01)00080-9

Chaves, N. (2004, April). *La marca destino turístico, cinco estrategias gráficas.* Paper presented at the XIII Simposi Internacional de Turisme i Lleure ESADE-Fira de Barcelona. Retrieved April 30, 2007, from www.esade.es/cedit2004/cat/est_estudios.php

Chen, L. D., Gillenson, M. L., & Sherrell, D. L. (2002). Exciting online consumers: an extended technology acceptance perspective. *Information & Management*, *39*, 705–719. doi:10.1016/S0378-7206(01)00127-6

Chen, L. D., & Wells, W. D. (1999). Attitude toward the site. *Journal of Advertising Research*, *39*(5), 27–38.

Cho, C.-H., & Cheon, H. J. (2005, Summer). Cross-cultural comparisons of interactivity on corporate Web sites. *Journal of Advertising*, *43*(2), 99–115.

De Chernatony, L., & Dall'Olmo Riley, F. (1997). Modelling the components of the brand. *European Journal of Marketing*, *32*(11/12), 1074–1090. doi:10.1108/03090569810243721

Ekinci, Y., & Hosany, S. (2006). Destination personality: An application of brand personality to tourism destinations. *Journal of Travel Research*, *45*, 127–139. doi:10.1177/0047287506291603

ESADE, & BDDO Consulting (2004, April). *Evolución del posicionamiento de los destinos turísticos españoles: De lugares de vacación a marcas de turismo.* Paper presented at the XIII Simposi Internacional de Turisme i Lleure ESADE-Fira de Barcelona. Retrieved April 30, 2007, from www.esade.es/cedit2004/cat/est_estudios.php

Etchner, C. M., & Brent Ritchie, F. (1991). The meaning and measurement of destination image. *Journal of Tourism Studies*, *2*(2), 2–12.

Gnoth, J. (1998). Conference reports: Branding tourism destinations. *Annals of Tourism Research*, *25*, 750–760.

Hankinson, G. (2004). The brand images of tourism destinations: A study of the saliency of organic images. *Journal of Product and Brand Management*, *13*(1), 6–14. doi:10.1108/10610420410523803

Hankinson, G. (2005). Destination brand images: A business tourism perspective. *Journal of Services Marketing*, *19*(1), 24–32. doi:10.1108/08876040510579361

Heijden, H. (2003). Factors influencing the usage of web sites: The case of a generic portal in The Netherlands. *Information & Management*, *40*, 541–549. doi:10.1016/S0378-7206(02)00079-4

Hollis, N. (2005, June). Ten years of learning on how online advertising builds brands. *Journal of Advertising Research*, 255–268. doi:10.1017/S0021849905050270

Hosany, S., Ekinci, Y., & Uysal, M. (2006). Destination image and destination personality: An application of branding theories to tourism places. *Journal of Business Research*, *59*, 638–642. doi:10.1016/j.jbusres.2006.01.001

Kapferer, J. N. (1997). *Strategic brand management.* London: Kogan Page.

Keller, K. (1993). Conceptualizing, measuring, and managing customer-based brand equity. *Journal of Marketing*, *57*, 1–22. doi:10.2307/1252054

Klein, L. R. (2003). Creating virtual product experiences: The role of telepresence. *Journal of Interactive Marketing*, *17*(1), 41–55. doi:10.1002/dir.10046

Ko, H., Cho, C.-H., & Roberts, M. S. (2005). Internet uses and gratifications. A structural equation model of interactive advertising. *Journal of Advertising*, *34*(2), 57–70.

Konecnik, M., & Gartner, W. C. (2006). Customer-based brand equity for a destination. *Annals of Tourism Research, 34*(2), 400–421. doi:10.1016/j.annals.2006.10.005

Kotler, P., Haider, D. H., & Rein, I. (1993). *Marketing places: Attracting investment, industry and tourism to cities, states and nations.* New York, NY: Free Press.

Lawson, F., & Band-Bovy, M. (1977). *Tourism and recreational development.* London: Architectural Press.

Liu, Y. (2003, June). Developing a scale to measure the interactivity of Web sites. *Journal of Advertising Research*, 207–216.

Liu, Y., & Shrum, L. J. (2002). What is interactivity and is it always such a good thing? Implications of definition, person and situation for the influence of interactivity on advertising effectiveness. *Journal of Advertising*, *31*(4), 53–64.

López Lita, R., & Benlloch, M. T. (2005). La marca territorio. El marketing de ciudad, una herramienta al sevicio de las marcas territorio. *99% com*, *2*, 8.

Lynch, P. J., & Horton, S. (2004). *Manual de estilo web. Principios de diseño básico para la creación de sitios Web*. Barcelona: Gustavo Gili, SA.

Martín Barbero, S., & Sandulli, F. (2005). *Marcating en la Web: BIWAM. Identidad desnuda.* Madrid: Dossat 2000.

McMillan, S. (2004). Internet advertising: One face or many? In D. Schumann and E. Thorson (Eds.), *Internet advertising: Theory and research (2nd edition)*. New York, NY: Lawrence Erlbaum.

McMillan, S., Hwang, J.-S., & Lee, G. (2003, December). Effects of structural and perceptual factors on attitudes toward the website. *Journal of Advertising Research*, 400–409.

Morgan, N., Pritchard, A., & Piggott, R. (2002). New Zeland, 100% Pure: The creation of a powerful niche destination brand. *Journal of Brand Management*, *9*(4/5), 335–354. doi:10.1057/palgrave.bm.2540082

Morgan, N., Pritchard, A., & Pride, R. (2004). *Destination branding. Creating the unique destination proposition.* Oxford: Elsevier.

Nielsen, J. (2002). *Usabilidad. Diseño de sitios web.* Madrid: Prentice Hall.

Padín Fabeiro, C. (2004, April). *La formación de la imagen de un nuevo destino.* Paper presented at the XIII Simposi Internacional de Turisme i Lleure ESADE-Fira de Barcelona. Paper presented at the XIII Simposi Internacional de Turisme i Lleure ESADE-Fira de Barcelona. Retrieved April 30, 2007, from www.esade.es/cedit2004/cat/est_estudios.php

Park, C. W., Jaworski, B. J., & MacInnis, D. J. (1986, October). Strategic brand concept management . *Journal of Marketing*, *50*, 135–145. doi:10.2307/1251291

Prebensen, N. K. (2007). Exploring tourists' images of a distant destination. *Tourism Management*, *28*, 747–756. doi:10.1016/j.tourman.2006.05.005

Riezebos, R. (2003). *Brand management. A theoretical and practical approach.* Harlow, UK: Prentice-Hall.

Ritchie, B., & Ritchie, R. (1998). *The branding of tourism destination: Past achievements and future trends*. Paper presented at the 48th Congress, AIEST, St-Gall.

Sicilia, M., Ruiz, S., & Munuera, J. L. (2005). Effects of interactivity in a Web site. *Journal of Advertising*, *34*(3), 31–45.

Simon, S. J. (2001). The Impact of culture and gender on web sites: An empirical study . *The Data Base for Advances in Information Systems*, *3*(1), 18–37.

Spool, J. M. (1996, January 1). Branding and usability. *User Interface Engineering*. Retrieved June 1, 2008, from www.uie.com/articles/branding_usability.

Tsikriktsis, N. (2002). Does culture influence web site quality expectations? An empirical study. *Journal of Service Research*, *5*(2), 101–112. doi:10.1177/109467002237490

World Tourism Organisation. (1999). *Promoción de destinos turísticos en el Ciberespacio*. Madrid: WTO.

World Tourism Organisation. (2001). *Comercio electrónico y turismo*. Madrid: WTO.

ENDNOTES

1 According to Nina K. Prebensen (2007), to build a good city brand, with the most suitable elements, that create a positive image, the opinion and knowledge of the visitors or future visitors is fundamental.

2 Bill Baker is the founder and President of Total Destination Management. He is internationally recognised as an expert in building brand recognition for destinations and communities. TDM is a Portland, Oregon based team of destination branding, tourism planning and marketing specialists. www.DestinationBranding.com consulted in July 2007.

3 Blain, Levy and Brent Ritchie (2005) confirm that the logos of city brands globally represent the experience that the visitors expect of a destination or city. Effectively, logos stimulate the communication of the attributes desired among the visitors and influence the tourists' decision to visit a place.

4 Hosany, Ekinci and Uysal (2006) demonstrated that there is a relationship between the image of a destination and its personality. They also demonstrated that the image of a destination's brand is fundamental to position it effectively.

5 Other authors, such as Keller (1993) and Park (1986) add a third category or element to the city brand: experiential attributes, which refer to the experiences, satisfaction and feelings of the destination's visitors.

6 Spool compared two websites: one more usable and informative and the other very graphically meticulous. He showed that the more usable site created a better brand image, as it satisfied the expectations of the users to a greater degree. Therefore, the graphic aspects of a website, such as logos and photographs, have less effect on the branding than expected.

7 International Organization for Standardization (ISO). In: http://www.iso.org/iso/home.htm

APPENDIX: LIST OF WEBSITES ANALYSED

America

Table 6. America

México D.F.	www.mexicocity.gob.mx
Los Ángeles	www.lacvb.com
Montreal	www.tourisme-montreal.org
New York	www.nycvisit.com/home/index.cfm
Toronto	www.torontotourism.com
Buenos Aires	www.buenosaires.gov.ar/areas/turismo/home
Río de Janeiro	www.riodejaneiro-turismo.com.br/en/home.php

Africa

Table 7. Africa

El Cairo	www.cairotourist.com
Marrakesh	www.ilove-marrakesh.com

Asia

Table 8. Asia

Bei-jing	english.bjta.gov.cn
Tokyo	www.tourism.metro.tokyo.jp/english/index.html
Hong Kong	www.discoverhongkong.com
Kuala Lumpur	www.kualalumpur.gov.my
Bangkok	www.bangkoktourist.com
Macau	www.macautourism.gov.mo
Bali	www.balitourismauthority.net/home.asp
Delhi	delhitourism.nic.in

Oceania

Table 9. Oceania

Sydney	www.cityofsydney.nsw.gov.au

Europe

Table 10. Europe

Amsterdam	www.iamsterdam.nl
Athens	www.cityofathens.gr
Dublin	www.visitdublin.com

Istanbul	english.istanbul.com
Florence	www.firenzeturismo.it/en_default.asp
Helsinki	www.hel2.fi/tourism
Lisbon	www.cm-lisboa.pt/turismo
London	www.visitlondon.com
Paris	www.parisinfo.com
Prague	www.visitprague.cz
Roma	www.romaturismo.com
Vienne	info.wien.at
Berlin	www.berlin-tourist-information.de
Kiev	www.kmv.gov.ua
Krakow	www.krakow.pl
Budapest	www.budapestinfo.hu
Moscow	www.moscowcity.com
Zagreb	www.zagreb-touristinfo.hr

Spain

Table 11. Spain

Barcelona	www.bcn.es/turisme/catala/turisme/welcome.htm
Madrid	www.esmadrid.com
Sevilla	www.turismo.sevilla.org
València	www.turisvalencia.es

This work was previously published in Information Communication Technologies and City Marketing: Digital Opportunities for Cities Around the World, edited by M. Gasco-Hernandez; T. Torres-Coronas, pp. 26-49, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.6
Assessing the Performance of Airline Web Sites:
The ARTFLY Case

Elad Harison
University of Groningen, The Netherlands

Albert Boonstra
University of Groningen, The Netherlands

EXECUTIVE SUMMARY

This case takes place in the increasingly competitive environment of the airline sector. Airline websites and Internet-based booking systems enable transformation of airline operations and become strategic weapons for the majority of airlines worldwide. Established airlines are attempting to stimulate customers to use the Internet, in response to entry of low cost carriers that capture shares of their market and in order to reduce their sales costs. Nonetheless, the development and maintenance of websites and e-commerce platforms requires substantial investments in capital and labor. Therefore, airlines need to assess the performance of their e-commerce channels in terms of profitability, customer appreciation and volume of sales on a continuous basis. However, the design of an assessment model that can serve the managers of ARTFLY, one of the established airlines in the industry that deals with the recent challenges of the intensifying competition, is open for a wide variety of interpretations and should be determined due to the firm's nature of operations and due to its aim to increase the volume of its online sales.

ORGANIZATIONAL BACKGROUND

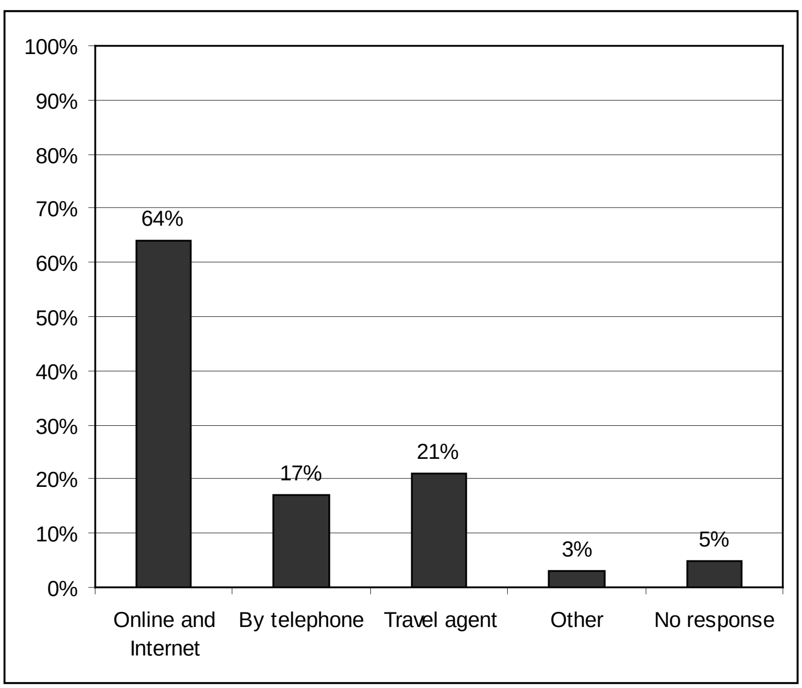
ARTFLY, founded in 1919, is a major airline that operates on a worldwide basis. In 2004 ARTFLY merged with the Air Minoli group and became a division within the joint Air Minoli/ARTFLY group. Air Minoli/ARTFLY has a worldwide coverage, in terms of destinations, flight routes and marketing units. The group uses two main European airports as their main hubs and offers air transport to 128 destinations in 65 countries in 5 continents. To illustrate the volume of the group's activities in 2004-2005 the ARTFLY group transported more than 20 million passengers and more than 600,000 tons of cargo, generating profits of 255 million Euros ($300 million U.S.).

In 1995, ARTFLY was one of the first European airlines that introduced fully integrated e-business into its ticketing process, in response to radical changes in the air transport market (Rubin & Joy, 2005). In particular, the entry of low cost airlines and the price war that followed presented a tangible threat to ARTFLY's market. Low cost carriers succeeded in attracting increasing numbers of passengers who previously purchased their tickets through travel agents and airline branches and have shifted to booking their flights online via the Web sites of those carriers (see Figure 1). For example, the rise of EasyJet and Ryanair to dominance in the British market, flying 55 million passengers from and to the U.K., was achieved in part as a consequence of online ticket sales (ABTA, 2005). ARTFLY replaced its previous procedures with online bookings, electronic tickets, and electronic check-ins and boarding passes. Reducing its costs by direct sales and using the Internet as a prominent marketing channel were immediate actions that were taken by ARTFLY's management and assisted in surviving the intense competition in the market and maintaining its major position in this rapidly changing business environment.

In March 1996, the company launched its first Web site, which was mainly an electronic brochure with information on flights. A year later, new functionality was added including real-time information on departure and arrival times, a reservations module and electronic ordering of tickets. From 2001, a complete electronic booking system was included in ARTFLY's homepage and, since then, the Web site is continuously maintained and often face-lifted due to changes in the market and new technological possibilities.

ARTFLY bases its marketing and sales activities on four main distribution channels that include its own Web site, online travel agencies (e.g., Expedia.com and Kayak.com), ARTFLY branches and "physical" travel agencies (see Table 1). The firm has defined its Web site, ARTFLY.

Figure 1. Channels of low cost flight bookings (Source: ABTA, 2005)



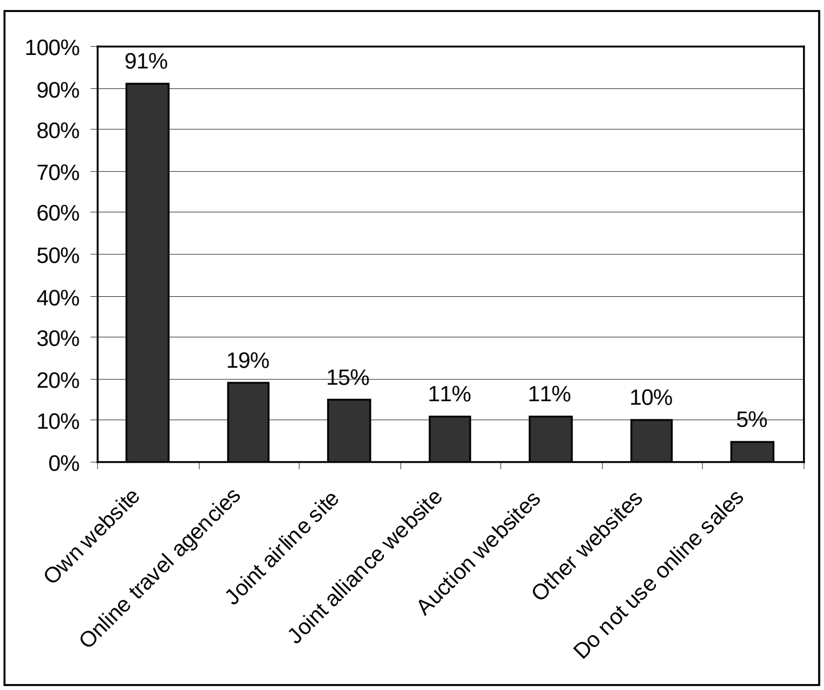
com, as the most important channel in its home market. The company is keen to increase the share of online bookings via the Web site (together with other online travel agencies) from 25% up to 40%. Outside its home market, ARTFLY branches are its main sell points. The branches are located at major airports and in the centers of large cities worldwide, hence enabling accessibility of customers to its booking services. In addition, ARTFLY operates through local travel agencies. However, the firm aims at replacing part of its offline sales by increasing the share of online bookings, as the physical distribution channels involve either higher operating expenses (e.g., personnel and location costs) or commissions paid to independent travel agencies.

The targets of ARTFLY can be compared to the industry's benchmarks. In 2006, the Airline IT Trends Survey (SITA, 2006) measured the relative share of bookings in each distribution channel in 98 major airlines worldwide. It concludes that, on average, the airlines' own Web sites (the direct online channel) captured 21.5% of the bookings, 8.2% of the tickets were purchased through online travel agencies (the indirect online channel), 17.7% of the bookings were made in the branches

Table 1. Four distribution channels of ARTFLY

	Direct	**Indirect**
Online channel Now (2005/06): 12% Target (2008/09): 40%	1 www.ARTFLY.com	2 Online travel agencies (e.g., Expedia.com)
Offline channel Now (2005/06): 88% Target (2008/09): 60%	3 ARTFLY branches	4 Travel agencies with branches (e.g., Thomas Cook, but mostly operate on local, regional or national scales)

Figure 2. Use of online channels for flight bookings (Source: SITA, 2006)



of flight companies (the direct offline channel), while the majority of bookings – 52.6% of the flight tickets sold in 2006 – were made through "physical" travel agencies (the indirect traditional channel). On the supply side, 91% of the firms in the survey offer their customers the possibility to book their flights via their own Web sites, and only a small proportion of the firms (5%) do not have any direct or indirect online channels (see Figure 2).

The firm's policy, that is, setting a 40% share of sales from online sources, stems from three main objectives, as follows:

- *Cost-reductions* that result from simplifying the booking and payment processes and reducing the scale of human interventions in the ticketing process by emailing e-tickets to customers.
- *Better services* are provided by a faster check-in procedure and by enabling online customers to select their seats while booking their flight. Doing so, customers do not have to wait for their boarding pass and can shorten their stay at the airport.
- *Increased market share* is achieved by gaining better performance on the Internet than its competitors. This aim involves both high degrees of Internet presence and use of intelligent strategies and customer relationship management (CRM) tools to identify the preferences and patterns of purchasing decisions among online customers.

Over time, ARTFLY's Web site has become increasingly important for the firm in terms of its growing volume of sales and as an attractor of potential customers. To illustrate the growth of its online channel of distribution, in 1997 ARTFLY's Web site received approximately 3,000 visitors per week, while during 2007 it attracted more than one million weekly visitors. In 2001, the site reached average monthly revenues of 0.5 million Euros, while the monthly online revenues in 2007 were approximately 156 million Euros (source: Reference Document ARTFLY/Air Minoli, 2006). However, the exponential increase in the use of ARTFLY's Web site led to larger volumes of online traffic and to demands from customers for new and advanced features, availability and reliability of the Web site.[1] Consequently, ARTFLY's online operations required a new assessment model that captures the multiple dimensions of its online operations and their effectiveness in terms of additional profits and growth of its customer base.

Currently, the evaluation of ARTFLY's online activities and performance is coordinated by the CEO through a series of monthly meetings that involve managers of different divisions and units. The managers report the changes in the indicators for which they are responsible and recommend on adapting existing online tools and campaigns or highlight the needs to develop new systems. Then, the feasibility of developing or modifying new tools, their expected contribution to the organization and their costs are discussed within this forum. However, this process suffers from two major shortcomings: First, the configuration of the current performance evaluation process requires significant coordination efforts due to the involvement of different divisions and organizational units and due to the broad distribution of responsibilities among them (see Figure 3). Second, the CEO can assess the performance of ARTFLY's Web sites only in a monthly resolution and cannot receive continuous measurements of performance to assess it within shorter periods (for example, on a daily or on a weekly timeline).

SETTING THE STAGE

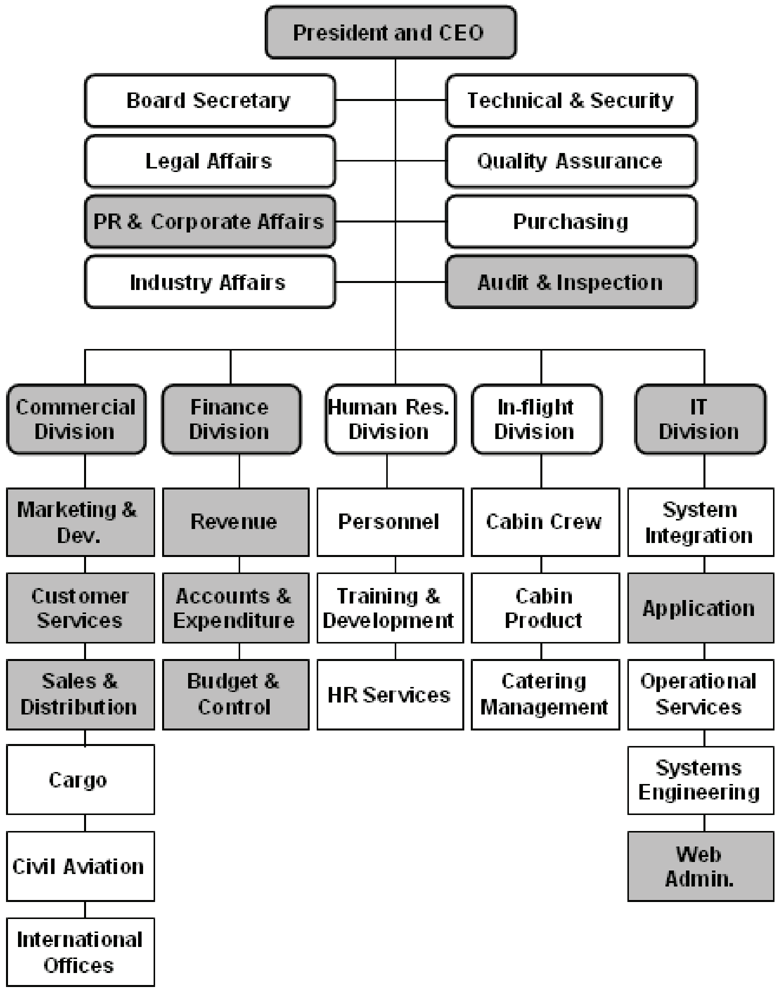
Recent studies in marketing and in ICT emphasized the potential of Web sites and e-commerce tools as new distribution channels that can create new opportunities for cost savings and profits for new and existing companies (Buhalis, 2004; Xing & Grandt, 2006). In particular, early air transport studies identified how Internet technologies can be implemented by airlines to increase the volumes of their online ticket sales and to expand their marketing channels (Alamdari, 2002; Jarach, 2002; Lubbe, 2007; Shon, Chen & Chang, 2003; Yoon, Yoon & Yang, 2006). Recent studies focussed on assessing the impact of commercial Web sites on broader dimensions of activities and performance of firms (see, for example, Agarwal & Venkatesh, 2002; Gianforte, 2004; Huizingh, 2002; Otim & Grover, 2006; Park & Gretzel, 2007).

Those studies seem to be of a major importance particularly to airline companies, as their market is characterized on the one hand by strong competition and on the other hand by opportunities to sustain customer relations and loyalty (e.g., via frequent flyer programs, Gasson, 2003). In addition, the entry of low-cost airline companies in recent years intensified price competition and put higher pressure on the profit margins of many firms (Klein & Loebecke, 2003; Klein, Klein, Kohne & Oorni, 2004). In this respect, the Internet assists airlines in developing alternative marketing and distribution channels, reducing direct costs, reaching new customers and maintaining long-term relations with existing customers. However, the strong competition in the market and the openness of the Internet as a communications medium enabled firms to learn and to apply similar online business practices, Web site designs and e-commerce tools. In other cases, the rapid development of Internet technologies has made within a short period Web sites less user friendly or less attractive in comparison to Web sites of competitors, and sometimes even obsolete (Law & Leung, 2000). Consequently, marketing managers in airline firms, as well as in other sorts of organizations that base a large part of their activities on electronic commerce, are in need of an evaluation tool to assess the effectiveness and the reach of their online operations and to indicate what improvements in it should be made.

Internet marketing and e-commerce are largely based on site building, promoting it online (e.g., via search engine) and via traditional marketing campaigns and attracting potential customers that enter it to purchase goods and services (Chu, 2001). The accessibility to the Web site and the knowledge of consumers of its existence and the range of information and services that it provides are major factors in its success, as well as the choice of advertising channels with high potential to attract customers (Monsuwe, Dellaert & Ruyter, 2004). It is also important to enhance their presence and the volume of Web traffic to them by other means, such as association with related search terms and Web-based campaigns. Complementary off-line campaigns can also reinforce the exposure of Web sites to potential Web-surfers and customers. Drèze and Zufryden (2004) describe nine categories of traffic building strategies that should be taken into account while constructing Web-traffic plans: off-line advertising of a Web site, offline news reports, Internet advertising/banner ads, links from other Web sites, links from search sites, links from online directories, online news reports, reference in an e-mail and discussion/newsgroups. The exposure of Web sites increases not only but including it in search engines and search-word based advertising, but also by more "traditional" means, such as banners, TV campaigns and press releases. Attracting large numbers of visitors to the firm's Web site is an important step to the successful of Web-based exposure and merchandizing.

Although financial indicators of the operations of the firm (including its online marketing, distribution and sales)[2] are essential for evaluating the performance of Web sites, a different approach that is based on a broader framework of the firm's operations and competitive position is often needed (Toh & Raven, 2003). Financial measurements of the performance of Web sites and online tools mostly reflect short-term measurements, such as costs and revenues and overlook other performance measures of online activities and long-term effects on the firm.

Figure 3. ARTFLY's organizational structure (Divisions and units participating in the evaluation of the performance of ARTFLY's Web sites are marked in grey)



Analyses of Internet ventures, including those of *non-Internet* product and service providers, should be done at the industry level. When an individual tool fails to meet the industry's standards, the firm is advised not to engage in developing online marketing and distribution channels. For example, if an airline company invests an online booking system and the volume of bookings are inferior to other players in the market, it should either modify the system or abolish it and turn to its traditional distribution channels (e.g., agents and call centers, Smyth & Wagner, 2006).

CASE DESCRIPTION

By the end of 2006 the Finance Director, Mrs. Eva van Duinen, started a discussion during a weekly board meeting, in which she expressed her doubts about the performance of the Web site of ARTFLY. She felt that the usual growth of online sales were stagnating. The CEO, Mr. Jim Manzetti, agreed with her that it would be wise to give more attention to this theme. The first thing to do for a more thorough discussion in the management team would be to collect all relevant data on Web site performance. The members of

Table 2. Comparative measures of ARTFLY's Web site use in subsequent periods

Measures	2005 Q4	2006 Q4
1 Financial overview:		
Revenues from online bookings (in USD)	88,359	84,563
Number of tickets sold via Web site	209,523	206,251
Costs of Web site maintenance (in thousands of USD)	12,450	14,429
2 Customer appreciation of Web site according panel		
Ease of finding the Web site (on scale of 1 – 10)	7.3	7.3
Overall appreciation of Web site (on scale of 1 – 10)	6.9	6.5
3 Contact		
Number of visitors	5.133.935	4.988.238

the management team agreed that they would look for this information, especially the information that compares Web site performance between the 4th quarter of 2005 and 2006.

One week later, the Web site performance theme was again on the agenda. The Finance director reported that the revenues from online bookings had gone down with nearly 5%, compared with the last quarter of 2005. She also mentioned that the costs of Web site development and maintenance had gone up by 12%. John Simons, IT director, explained this rise of IT costs by growing costs of maintenance as a result of the increasing number of Web pages and the hiring of external experts. He said that more Web pages and more online traffic automatically lead to higher costs in terms of maintenance. External experts were needed in order to make Web sites more aesthetically pleasing. Developing such expertise in-house would be much more expensive.

Mr. Otto Caroll, director of the sales department delivered additional figures that showed that the number of tickets sold is also slightly decreasing. In addition to this, other indicators revealed that the volume of use of the Web sites by new and returning ("loyal") customers is lower than in the previous years and its popularity among them is decreasing. During the same meeting, the public relations manager, Mrs. Anna Jacobsson, reported from a panel discussion with customers that the overall appreciation of the Web site by the panelists was also gone down a little bit. These panel discussions were held with a group of twenty customers on a quarterly basis and covered many issues related to customer services and customer satisfaction of ARTFLY.

The data on Web site performance were grouped by a staff member in Table 2. Based on these data, a discussion started among the various directors. Eva van Duinen, the Finance Director, felt strongly that a program of cost cutting would be necessary to align investments in online channels with the performance of those channels.

Director Otto Caroll from Sales argued that the figures reflect a growing and intense competition, especially with low-cost airlines, which directly compete with ARTFLY in its existing routes, and that a substantial share of ARTFLY's customers have preferred to book their flights online via their Web sites. He argued that ARTFLY should know what they are doing wrong with their Web sites, before cutting costs.

The IT director contributed to the discussion by saying that a strategy of promoting the online channel should be supported by high investments in quality and attractiveness of the Web sites. He thinks that cost saving programs will jeopardize such a policy.

All directors agreed that the available information as reflected in table 1 was not appropriate for making far reaching decisions on investment programs or savings on ARTFLY Web sites. The participants were missing comparisons with the performance of other airlines as well as figures about the performance of off-line channels. Others were asking what the determinants of customer satisfaction of airline Web sites are. Speed? Attractiveness? Completeness? It would not be wise to invest or to save costs without having a thorough knowledge on important relevant issues that determine the performance of online channels.

When the discussion continued, it became clear the directors did not completely agree on the most relevant issues that determine Web site performance. The financial director emphasized the financial justification of investments in Web sites. Others agreed that this is important, but the financial performance will be determined by actual sales. These are determined by issues like the easiness to find a Web site, the speed and loading time of a Web site, attractiveness and so on. The IT director, John Simons added to this by saying that ARTFLY also pays considerable amounts of money to search engines like Google and Yahoo. Costs of clicks from those sites varied from 0.20 – 0.30 cents. To his opinion, it would be important to assess whether these costs can be justified by increased online sales. Marketing manager, George Veldman emphasized that Web sites also realize important cross sales in terms of hotel bookings, car rentals and so on. He felt that the profits from cross selling should be captured in the whole decision making process.

The CEO concluded this discussion by saying that the company needed a Web site performance monitoring system that includes the issues raised by the various directors. Such a system should have to lead to frequent reports that show how the company's Web sites are performing. This would help to make rational decisions with regard to investments and cost saving programs much easier. He emphasized that the system should also be related to the strategy of ARTFLY, which includes a careful move to online channels, while maintaining off-line channels and keeping customers satisfied who prefer off-line channels.

IN NEED OF AN ASSESSMENT MODEL: DEFINING KEY INDICATORS TO ASSESS THE PERFORMANCE OF ARTFLY'S WEB SITE

Recent studies in marketing and in ICT have addressed the potential of Web sites and e-commerce tools to provide new distribution channels that can increase the profitability of flight companies. Online bookings can significantly increase the volume of sales by reaching new customers via Internet presence and campaigns and can reduce the operation costs of firms by eliminating the need to pay commissions to travel agents (Xing & Grandt, 2006). Major airlines were aware of the need for an assessment model to evaluate the performance of their online activities (notably, their Web sites and online ticket sales) and have adopted various models for this purpose (Alamdari, 2002; Jarach, 2002; Lubbe, 2007; Shon et al., 2003; Yoon et al., 2006). However, most of these models address only limited aspects of the online activities and Web site operation, such as the number of entries per period or the user-friendliness of the booking system, and do not provide a coherent overview of the indicators that define in detail the role of the Web sites. For example, most of these assessment methods overlook the additional added value of the Web sites in terms of their impact on sales, contribution to the airlines' profitability, reputation and presence in the market. A second type of indicators that is often omitted from the assessment models that are described in the literature evaluates the Web sites in terms of the quality of their technical operation. For example, the proportion of down time and the number of malfunctions can be included in the model as indicators for technical performance. The assessment of the technical performance of Web sites is particularly important as the availability of their online booking systems strongly affects the volume of online sales and the level of customer satisfaction. On the other hand, the period in which a Web site cannot be accessed can be directly translated to loss of sales and potential bookings. Even when some of the customers decide to complete their order via another marketing channel, such as call centers or travel agencies, it would be less profitable for the firm in comparison to online bookings.

The impact of commercial Web sites on the performance of the firm can be assessed through a broad prism that includes both financial key performance indicators as well as via qualitative and quantitative measures. Those measures should reflect the satisfaction of clients from the airline's Web site and the accessibility to its contents and to the booking system.

The financial appraisal of the firm's operations is the direct and most important indicator to evaluate its e-commerce channel. Therefore, online operations, and particularly the construction and the maintenance of a Web site, can be assessed from the financial standpoint by means used to evaluate other IT investment decisions: Profitability is the most important objective and it reflects the proportion between profits and investments. In addition, the risks associated with the development and operation of a Web site and an online booking system should be taken into the account of the expected financial gains (Renkema & Berghout, 1997). However, the financial measurements present the costs and revenues and do not include other measures that can affect the performance of the firm, such as service quality, reputation and brand loyalty. Therefore, the financial measures should be complemented by qualitative and quantitative indicators.

In addition to those indicators, nonfinancial factors that affect the performance of airline Web sites can be assessed via the following categories:

- **Customer satisfaction:** the ability of customers to locate the Web site of the company, to access it with ease and to use its contents in ways that meet their needs. This indicator is measured through customer surveys and focus groups.
- **Accessibility:** the ease of finding a Web site and the exposure of potential customers to it. This criterion includes also the technical aspects of accessing the Web sites with different Web browsers, presenting the contents properly and downloading the Web site in a reasonable time when various Internet connections are used.
- **Traceability:** measures the traffic to the Web site, access to the different contents and Web pages within it, its presence in online portals and its position (rating) in search engines. The criterion is measured by online traffic measurement tools.
- **Contact:** a sales-based criterion that reflects the interest of customers in the online services of the firm. It measures the volume of online visitors that access the Web site in a given time and the distribution between new and returning visitors. The number of visitors is especially important as it is used also to assess the traceability of the Web site and the accessibility of consumers to it. Contact data are obtained from customers that visit Web sites after viewing advertisements in the media (newspapers, radio and TV), follow Internet advertisements (banners flash videos and pop-up windows) and *adsense* advertisements that link terms used in search engines to commercial Web sites[3]. Therefore, it is possible to provide a measure for the effectiveness of each advertising channel by tracking the volume of traffic that originates from it.
- **Sales:** indicators measure the revenues and the profits from online sales in a given period. The firm can compare between sales from "off-line" marketing channels, such as

travel agents and ARTFLY's offices, and the volume of orders that are completed via its Web site. Changes in the relative share of online bookings reflect changes in consumers' behavior and trust in purchasing services via ARTFLY's Web site and can signal the need for a proactive online marketing, such as Internet-based campaigns and additional Web advertising. Sales data by channel can be obtained from ARTFLY's financial and accounting systems.

- **After-sales:** records data on service quality, the volumes and costs of online transactions (in comparison to off-line sales) and the satisfaction of customers from their online bookings.
- **Customer support:** measures the total number of visitors that apply for support services, such as technical assistance and follow-up on online bookings, in a given period, as well as assessing the costs of customer support services regarding the operation of the Web site and its various features.

Table 3. Criteria and measures for assessing ARTFLY's online products and services

Criteria	Indicators
Financial overview	Costs Revenues Direct revenues from bookings
Customer satisfaction	Ease of finding the Web site Ease of booking a flight online Speed of downloading the Web site and contents Speed of operating the Web site Overall appreciation of Web site Appreciation of services
Traceability	Number of domain names in use
Accessibility	Loading time Number of languages supported by Web site
Contact	Total: Number of visitors Number of new visitors Per channel: Number of visitors Costs per visitor Sales per visitor Average bookings per visitor
Sales	Total revenues Total sales Sales per product Profits per product Total profits Volume of bookings Total conversion Conversion per product Ratio of online to offline revenues
After-sales	Costs per booking Number of bookings Use in relation to sales Customer appreciation
Customer support	Use of the product Use of the product per visitor Costs of use Customer appreciation

Since the beginning of 2007, the model, presented in Table 3, was applied in various ways to measure the short-term and long-term performance of ARTFLY's Web site. The various indicators have formed a baseline that can provide information on changes in consumers' behaviour (such as growing demand for particular lines or seasonal trends). Further, by following the indicators of the model over time ARTFLY can monitor its online operations and assess the effectiveness of new contents and features, Internet-based promotions, different advertising channels and online marketing campaigns. Table 4 demonstrates the measures of the assessment model after implementing it for two years. However, the application of ARTFLY's assessment model is only in its initial stages and the firm plans to expand its use (such as assessing the use of the booking system by regular passengers vs. members of the frequent flier program) and to broaden the set of measurement indicators.

The metrics collected for the assessment model provide useful insights that can assist in managing the firms' online strategy and activities. After the implementation of the assessment model, ARTFLY's managers can identify major changes and market trends with ease and to respond to them. The model is often used to assess current practices and strategies of the firm, such as its advertising strategy and the distribution of its advertising budget between online and traditional media.

CURRENT CHALLENGES/PROBLEMS FACING THE ORGANIZATION

As mentioned previously, the airline market is dynamic and it is dominated by intensifying competition, particularly from low-cost carriers. In this business environment, ARTFLY aims at increasing the share of the online bookings ac-

Table 4. Comparative measures of the assessment model in subsequent periods.

Measures	2006 Q4	2007 Q1
	105,489	108,549
1 Financial overview:	14,759	15,563
Revenues (in thousands of Euros)	89,467	91,263
Costs (in thousands of Euros)		
Direct revenues from online bookings		
2 Customer satisfaction	7.1	7.2
Ease of finding the Web site	6.6	7.1
Speed	6.3	6.7
Overall appreciation of Web site	365	389
Appreciation of services		
3 Traceability		
Number of domain names		
	6.123669	6.23657
4 Accessibility		
Loading time in seconds		
Languages supported by Web site		
	8.36	7.92
5 Contact	11	14
Number of visitors		
Visitors from Google		
Sales to visitors from Google (millions of Euros)	4.563.236	5.239.246
Costs per click from Google (Euros)	562.346	599.756
Visitors from Yahoo	6.235	6.539
Sales to visitors from Yahoo (millions of Euros)	0.21	0.23
Costs per click from Yahoo (Euros)	123.756	142.153
	953	843
	0.30	0.29

complished by its customers, as those sustain its market share. Further, online bookings accommodate a tangible potential to lower ARTFLY's operational costs.

ARTFLY's management assesses the performance of its Web sites and online tools particularly on the basis of financial indicators, that is, the additional revenues and profits gained from those distribution channels in comparison to the financial investments necessary to construct and to maintain those electronic commerce platforms and online campaigns. However, the impact of ARTFLY's presence over the Internet goes beyond the direct and immediate financial prospects and has significant positive effects in terms of customer satisfaction, reputation and the quality of the information and services that it provides. Further, ARTFLY's management concluded the way the performance assessment of the Web sites and online booking facilities is carried out could be considerably improved. Hence, ARTFLY's actual challenges and questions are as follows and are followed by elaboration of the relevant aspects:

- What are the possible dimensions through which the performance of Web sites can be determined and assessed?

The current assessment model can be seen as a starting point for the evaluation of the performance of Web sites. Firms should address the following questions: Which dimensions are used in the current model, which aspects are emphasized and which are ignored? What are the choices that are made and how do they reflect the strategy of the company?

- How can ARTFLY's assessment model can be extended and improved?

Since the model as reflected in Table 3 and 4 are an early version of the model management should think about further refinements of it, as well as about removing less relevant elements of the model.

- How frequent should a performance measurement system of Web sites be operated to report to firm managers?

Information should be provided to support decision making that promote the use, the improvement and the maintenance of Web sites. These decisions also support the outphasing or abandonment of Web sites when needed. Management teams should determine how often they have to consider these issues.

- How should the evaluation process of the performance of Web sites be organized?

To realize actual use, data have to be collected, data have to entered into the system, information has to produced by the system and the information has to be distributed to the relevant people. Management should specify these generic activities and the responsibilities of carrying them out on a continuous basis.

- Which organizational units should be responsible for the performance evaluation and take part in it?

This issue is related to the question of who is primary responsible for the use and for the success of Web sites. Depending on the organizational structure and the roles and responsibilities of workers, Web site responsibilities can be divided in different ways. Alternative options are: Web site managers, business process managers, business unit managers, financial managers, division managers or CEO's.

- How can a performance measurement system be related to the strategies and to the strategic priorities of firms?

Firms have various strategic priorities, for example, they can follow a low cost strategy, a differention strategy, a global strategy or a niche strategy that can be monitored by the measurement system. However, those different strategies may be evaluated and monitored by different criteria that assess online processes and functions, as well as the Web sites themselves.

REFERENCES

Agarwal, R., & Venkatesh, V. (2002). Assessing a firm's web presence: A heuristic evaluation procedure for the measurement of usability. *Information Systems Research, 13*(2), 169.

Alamdari, F. (2002). Regional development in airlines and travel agents relationship. *Journal of Air Transport Management, 8*(5), 339-348.

Association of British Travel Agents (2005). *Travel statistics and trends.* London: ABTA Press.

Buhalis, D. (2004). eAirlines: Strategic and tactical use of ICTs in the airline industry. *Information & Management, 41*(7), 805-825.

Chu, R. (2001). What online Hong Kong travelers look for on airline/travel websites? *International Journal of Hospitality Management, 20*(1), 95-100.

Drèze, X., & Zufryden, F. (2004). Measurement of online visibility and its impact on Internet traffic. *Journal of Interactive Marketing, 18*(1), 20-37.

Gasson, S. (2003). *The impact of e-commerce technology on the air travel industry.* Hershey, PA: Idea Group Inc.

Gianforte, G. (2004). The world at our fingertips—How online travel companies can turn clicks into bookings. *Journal of Vacation Marketing, 10*(1), 79-87.

Huizingh, K. R. E. (2002). Towards successful e-business strategies: A hierarchy of three management models. *Journal of Marketing Management, 18*(7-8), 721-747.

Klein, S., & Loebbecke, C. (2003). Emerging pricing strategies on the web: Lessons from the airline industry. *Electronic Markets, 13*(1), 46-58.

Klein, S., Kohne, F., & Oorni, A. (2004). Barriers to online booking of scheduled airline tickets. *Journal of Travel & Tourism Marketing, 17*(2/3), 27-39.

Jarach, D. (2002). The digitalisation of market relationships in the airline business: The impact and prospects of e-business. *Journal of Air Transport Management, 8*(2), 115-120.

Law, R., & Leung, R. (2000). A study of airlines' online reservation services on the internet. *Journal of Travel Research, 39*(2), 202-213.

Monsuwe, T. P., Dellaert, B. G. C., & Ruyter, K. (2004). What drives consumers to shop online? A literature review. *International Journal of Service Industry Management, 15*(1), 102-121.

Lubbe, B. (2007). The effect of Internet apprehension and website satisfaction on air travellers' adoption of an airline's website. *Journal of Air Transport Management, 13*(2), 75-80.

Otim, S., & Grover, V. (2006). An empirical study on Web-based services and customer loyalty. *European Journal of Information Systems, 15*, 527–541.

Park, Y. A., & Gretzel, U. (2007). Success factors for destination marketing web sites: A qualitative meta-analysis. *Journal of Travel Research, 46*(1), 46-63.

Renkema, T. J. W., & Berghout, E. W. (1997). Methodologies for information system investment evaluation at the proposal stage: A comparative view. *Information and Software Technology, 39*(1), 1-13.

Rubin, R. M., & Joy, J. N. (2005). Where are the airlines headed? Implications of airline industry structure and change for consumers. *Journal of Consumer Affairs, 39*(1), 215-228.

Shon, Z. Y., Chen, F. Y., & Chang, Y. H. (2003). Airline e-commerce: The revolution in ticketing channels. *Journal of Air Transport Management, 9*(5), 325-331.

SITA (2006). *Airline IT trends survey.* Geneva: SITA Press.

Smyth, A., & Wagner, C. M. (2006). Business-to-business electronic marketplaces in the airline industry: Tool for enhanced efficiency in a volatile business. Transportation Research Record. 1951, 60-68.

Toh, R. S., & Raven, P. (2003). Perishable asset revenue management: Integrated internet marketing strategies for the airlines. *Transportation Journal, 42*(4), 30-43.

Xing, Y., & Grandt, D. B. (2006). Developing a framework for measuring physical distribution service quality of multi-channel and "pure player" internet retailers. *International Journal of Retail and Distribution Management, 34*(4), 278-289.

Yoon, M. G., Yoon, D. Y., & Yang, T. W. (2003). Impact of e-business on air travel markets: Distribution of airline tickets in Korea. *Journal of Air Transport Management, 12*(5), 253-260.

ENDNOTES

[1] Since the beginning of 2005, the latest version of the Web site is available in all the 65 countries and supports 20 different languages. ARTFLY's management demanded that the Web site would be available online in 99.99% of the time.

[2] On average, airlines expect to save 13% of their administrative costs by moving to online bookings (SITA, 2006).

[3] Google bases its activities and business model on linking the search terms of users to textual *adsense* advertisements. When customers click on featured advertisements, they open the promoted Web site. Then Google charges the advertiser per click.

This work was previously published in Journal of Cases on Information Technology, Vol. 11, Issue 1, edited by M. Khosrow-Pour, pp. 47-64, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 5.7
Aviation-Related Expertise and Usability:
Implications for the Design of an FAA E-Government Web Site

Ferne Friedman-Berg
FAA Human Factors Team - Atlantic City, USA

Kenneth Allendoerfer
FAA Human Factors Team - Atlantic City, USA

Shantanu Pai
Engility Corporation, USA

ABSTRACT

The Federal Aviation Administration (FAA) Human Factors Team – Atlantic City conducted a usability assessment of the www.fly.faa.gov Web site to examine user satisfaction and identify site usability issues. The FAA Air Traffic Control System Command Center uses this Web site to provide information about airport conditions, such as arrival and departure delays, to the public and the aviation industry. The most important aspect of this assessment was its use of quantitative metrics to evaluate how successfully users with different levels of aviation-related expertise could complete common tasks, such as determining the amount of delay at an airport. The researchers used the findings from this assessment to make design recommendations for future system enhancements that would benefit all users. They discuss why usability assessments are an important part of the process of evaluating e-government Web sites and why their usability evaluation process should be applied to the development of other e-government Web sites.

INTRODUCTION

On November 15, 2007, President Bush announced actions to address aviation delays during the Thanksgiving holidays. As part of this announcement, he directed people to visit the Web site fly.

faa.gov, which is a Federal Aviation Administration (FAA) e-government Web site that provides real time information about airport delays.

Fourth, the federal government is using the Internet to provide real-time updates on flight delays. People in America have got to know there's a Web site called Fly.FAA.Gov; that's where the FAA transmits information on airport backups directly to passengers and their families. If you're interested in making sure that your plans can -- aren't going to be disrupted, you can get on the Web site of Fly.FAA.Gov. As well, if you want to, you can sign up to receive delay notices on your mobile phones. In other words, part of making sure people are not inconvenienced is there to be -- get transmission of sound, real-time information. (Bush, 2007)

There has also been a concerted effort by the FAA to publicize its Web site by placing advertisements in airports across the United States. Many news outlets now provide airport delay information as part of their weather forecasts, and this delay information comes, most often, directly from the fly.faa.gov Web site.

Because this Web site is the public face of a large federal agency, it is important that it presents the agency in the best light possible. An agency Web site should be a positive public relations vehicle and should not, in itself, create any public relations problems. Although use of e-government Web sites is increasing annually, low user acceptance of e-government Web sites is a recognized problem (Hung, Chang, & Yu, 2006). Many factors affect whether or not someone will use or accept an e-government Web site, including past positive experience with e-government Web sites (Carter & Bélanger, 2005; Reddick, 2005); the ease of use of the Web site (Carter & Bélanger, 2005; Horst, Kuttschreutter, & Gutteling, 2007); the perceived trustworthiness of the information presented on the Web site (Carter & Bélanger; Horst, et al., 2007); the perceived usefulness of the Web site (Hung et al., 2006); and personal factors such as education level, race, level of current internet use, and income level (Reddick, 2005). If a Web site has many functional barriers, such as having a poor layout or producing incomplete search results, customers of the site may not use it (Bertot & Jaeger, 2006).

Early work in e-government has consistently ignored studying the needs of end users, and there has been little research focusing on the demand side of e-government (Reddick, 2005). That is, *what are customers looking for when coming to an e-government Web site?* Although there have been many benchmarking surveys conducted on e-government Web sites, benchmarking surveys often do not describe the benefits provided by a Web site and only enumerate the number of services offered by that site (Foley, 2005; Yildiz, 2007). Benchmarks do not evaluate the user's perception of sites and do not measure real progress in the government's delivery of e-services. However, governments often chase these benchmarks to the exclusion of all other forms of evaluation (Bannister, 2007).

E-government academics emphasize the importance of usability testing and highlight the need to focus on Web site functionality, usability, and accessibility testing (Barnes & Vigden, 2006; Bertot & Jaeger, 2006). However, despite its importance, many organizations still are not performing usability testing on e-government Web sites. Current work often does not address the needs of different user communities, employ user-centered design, or use rigorous methods to test the services being delivered (Bertot & Jaeger; Heeks & Bailur, 2007).

Governments around the world are working to review best practices for e-government evaluation methods (Foley, 2005). Because of the social and economic benefits of providing information online, it is important that e-government Web site designs meet the needs of its targeted users. In addition, it is important to document the benefits provided by the Web site to increase public support

(Foley). Carter and Bélanger (2005) point out that e-government Web sites should be easy to navigate. They note that the organization of information on the site should be congruent with citizens' needs. When consumers visit an e-government Web site, they are most frequently looking for information (Thomas & Streib, 2003), which they need to be able to find quickly and easily. If users encounter problems while using a Web site, they may become frustrated and be less likely to adopt or utilize e-government services in the future. A positive experience with an e-government Web site will be communicated to others (Carter & Bélanger), and a usable Web site can play a significant role in engendering trust in the agency itself.

Most Web usability research focuses on e-commerce sites and privately run Web sites (Hung et al., 2006), and people expect e-government Web sites to be as good or as usable as private sector sites (Irani, Love, & Montazemi, 2007). People are more likely to use an e-government Web site if the transactions with that site are compatible with previously conducted transactions on similar, non-government Web sites (Carter & Bélanger, 2005).

However, there are clear differences between e-government and e-commerce Web sites. For instance, e-government sites must provide universal accessibility so that all citizens have access to information. Additionally, e-government Web sites are accountable to the public, whereas commercial Web sites are only accountable to people who have a financial stake in the Web site. It is not always clear, however, where the boundary between these two types of sites lies (Salem, 2003). Additionally, there are often challenges faced in producing e-government Web sites that are not faced by commercial sites (Gil-Garcia & Pardo, 2005). For example, when creating e-government Web sites, designers need to consider whether the project goals align with the goals or mission of the government agency (Yildiz, 2007). They also must make sure that all project stakeholders are involved, determine whether they are in compliance with all relevant government regulations, and work within government budget cycles and changing government contractors.

The FAA and fly.faa.gov

The FAA Air Traffic Control System Command Center provides information about airport conditions, such as arrival and departure delays, to the public and the aviation community via their Web site, www.fly.faa.gov. This Web site allows users to view airport conditions for specific airports.

The Web site has many different functions that help the user to search for delay information (see Figure 1). Using the **Search by Region** function, users are able to look up airports in different geographic regions, such as the Northeastern states and the Southeastern states. When using the **Search by Airport** function, users are able to search for airport delay information by typing in the name of a city, airport, or a three-letter airport code. The **View by Major Airport** function allows users to search for delay information using a drop down list of 40 major airports.

The site is also a repository of information for use by airlines, pilots, passengers, government personnel, academics, individual aircraft operators, and other stakeholders in the aviation community. It provides access to real-time and historical advisory information, real-time airport arrival demand information, current reroutes, and reroute restrictions. It also provides access to information related to air traffic management tools, a glossary of aviation terms, a national routes database, pilot tools for making arrival and departure reservations, a collection of National Airspace System documents, and many other air traffic tools.

The focus of this assessment was on the evaluation of site elements that the general public would access the most, such as the airport delay information and the glossary of aviation terms. From the user's point of view, the Web site needs to provide accurate information quickly, with minimal effort,

Figure 1. The www.fly.faa.gov home page, illustrating the View by Region, Search by Airport, View by Major Airport, and Site Map search methods

while minimizing potential mistakes. The site should be easy for users to learn and provide an appealing and satisfying experience.

We faced some unique issues and challenges when evaluating the fly.faa.gov Web site. First, the fly.faa.gov Web site presents real-time, up-to-the-minute data, whereas most e-government Web sites often present static information or information that changes infrequently. It was also clear that the expectations of site users were likely to be influenced by the information found on more commercial aviation sites. Because people have preconceived notions about the airlines and the reliability of information provided by airlines, it was possible that this perception could transfer to their perception of this Web site.

The Web site was also originally designed for use by people associated with the aviation industry, such as pilots and local airport authorities, who have at least a working knowledge of various aviation concepts. Because it is accessible on the internet and other travel sites have links to it, members of the traveling public (who may have little, if any, understanding of aviation or its associated jargon) also frequently use the site. The Web site is also being touted (Bush, 2007) as the first place the public should visit on the Web when looking for travel-related delays in the aviation system. Therefore, it was important to evaluate whether this site is usable by people who do not have a background in aviation. In this usability assessment, we examined how effectively people with different levels of domain knowledge were able to use the site.

It was difficult to identify a single typology that described the Web site. Although the site often looks like a Government to Consumer (G2C) site (Hiller & Bélanger, 2001), its original purpose was to function as a Government to Business (G2B) site or a Government to Employee (G2E) site. The site allows people to perform basic transactions (Hiller & Bélanger, Stage 3), but it also attempts to be a full-service, one-stop site for many types of aviation related information (Hiller & Bélanger, Stage 4). For instance, although this evaluation did not focus on the G2B information, airlines often use the site to find delay information, and general aviation pilots use the site to make route reservations. Although the site tries to organize its content to meet the different needs of these different categories of users (Ho, 2002; Schelin, 2003), it is not clear how the organizational structure was determined or whether it is the most optimal organization for all types of users.

We conducted this formal usability assessment to determine how successfully the Web site meets these usability goals and the needs of its users, including both expert and novice users. The assessment employed techniques commonly used in usability evaluations (Ahlstrom & Longo, 2003; Nielsen, 2003). The participants completed a set of representative tasks using the Web site, while researchers observed and recorded their actions and comments. Users also answered a series of questions rating the usability of the site. The data collected through these activities helped us identify a number of problems. After identifying the final list of usability issues, we used a part of the heuristic evaluation technique (Nielsen) to determine the most critical issues. This article discusses the technique used in this evaluation, highlights some of the most critical issues, and provides suggestions to designers on how to fix them. We also discuss the benefits of applying this formal process to the development of other e-government Web sites.

METHODOLOGY

Participants

We recruited 32 adult volunteers from the FAA William J. Hughes Technical Center to serve as participants. Because the participants were FAA employees, many had greater aviation-related knowledge than the general public. However, many FAA employees, such as administrative assistants and facility support workers, do not have significant knowledge of aviation or air traffic control. We included participants of both categories.

Equipment

The laptops used in the experiment contained fully interactive offline versions of the fly.faa.gov Web site. A User Script asked the participants to use the Web site to find information to answer 17 questions: 12 asked users to search for delay information, 3 asked users to find the definitions for aviation-related terms, and 2 asked users to identify the authority to be contacted when trying to obtain specific information. The script also asked users to use the **Search by Region**, **Search by Airport**, and **View by Major Airport** methods for specific questions. This allowed us to evaluate the usability of each function.

Procedure

Each session lasted 30 to 45 minutes. After signing an informed consent form, the participants completed a Background Questionnaire that collected information about the participants' knowledge of computers, Web sites, and aviation terminology.

After completing the Background Questionnaire, the participants next completed the User Script. We observed each participant during the experiment and recorded pertinent actions or comments. At the end of the experiment, the participants completed a Post-Session Questionnaire, where they rated their experience and identified usability issues.

Because using participants who all had a high level of aviation-related knowledge could have biased the results, we used the data to categorize the participants into three groups (novices, moderate knowledge users, and experts), based on their aviation-related knowledge. We analyzed the data by level of expertise to determine whether aviation-related knowledge had an impact on user performance. By analyzing the results in this way, we could make recommendations targeted toward making the site usable for the different user populations. When even individuals with a high level of aviation-related expertise had trouble using certain features, this provided strong evidence that those features needed to be redesigned. Even if novices were the only ones who had a problem with a feature, we rated that problem as severe if the impact for those users was severe.

RESULTS

Background Questionnaires

The Background Questionnaire asked the participants questions regarding their familiarity with aviation-related terms and acronyms. For example, participants were asked to list three-letter abbreviations for airports (e.g., Philadelphia International Airport = PHL), or were given the three-letter abbreviations and asked to list the airports associated with those abbreviations (e.g., MIA = Miami International Airport). Using the correct responses to these and other aviation-related questions, we categorized the participants as novices (n = 8), moderate knowledge users (n = 15), and experts (n = 9). The novices were slightly younger than both the experts and those with moderate-knowledge (M_{novice} = 41.6 years, $M_{moderate}$ = 49.9 years, M_{expert} = 49.1 years). More than 70% of novices and those with moderate knowledge reported never using the fly.faa.gov Web site. In contrast, 75% of the experts reported using the Web site a few times a year.

All the participants had extensive experience using computers and the Web. Because we found no discernable differences in reported Web and computer use among the participants, we were unable to stratify the participants based on these factors.

User Script Data: Overall Analysis

Of the 12 questions that asked users to find specific delay information, the participants answered 79.4% correctly. For the subset of five delay questions that allowed the participants to use their preferred search method, the participants answered 71.2% correctly. For the subset of four **Search by Airport** questions, 84.5% of the participants answered the questions correctly. For the **View by Major Airport** question, 90.6% of participants found the correct answer; for the **View by Region** question, 81.3% found the correct answer; and for the **Site Map** question, 87.5% found the correct answer.

Three questions asked the participants to use the site to provide the definition of three aviation related terms and abbreviations. Although 84.4 % of participants answered all three questions correctly, 6.3% answered one incorrectly, 3.1% answered two incorrectly, and 6.3% were not able to answer any of the questions. By comparing the percentage of participants who answered a question correctly, we determined that all three questions were equally difficult.

Two questions asked the participants to find whom to contact to obtain information about the status of an individual flight or why an airport was closed. For these questions, only 28.1% of the participants answered both questions correctly, 56.2% answered one incorrectly, and 15.6% answered both incorrectly.

User Script Data: Analysis by Level of Expertise

We analyzed the data by level of expertise to determine whether aviation-related knowledge had an impact on user performance. Analyzing all 17 questions, we found an effect of expertise on overall task performance, F (2, 29) = 3.54, p =

.04. Post hoc pairwise contrasts indicated expert participants were able to answer significantly more questions than novices (85.6% vs. 69.1%, $p = .01$), and there was a trend suggesting moderate-level users answered more questions than novices (79.6% vs. 69.1%, $p = .07$).

We performed ordinal (linear) chi-square tests on individual questions to determine whether the percentage correct increased or decreased across the user categories (Howell, 2007). Although only three of the questions were significant, 7 of the 12 delay questions showed the expected pattern of results (see Table 1). Therefore, we also tested the binomial probability that 7 of the 12 delay questions would show the expected ordering of expert > moderate > novice. We found that it was unlikely that this pattern would occur by chance 7 out of 12 times, $p < .001$. This suggests that experts were better able to find information on the fly.faa.gov Web site than moderate users,

Table 1. Percentage correct by level of aviation-related expertise

Questions	**% Correct**		
	Novices	**Moderate Users**	**Experts**
1. Los Angeles to Salt Lake City.**	75.0	100.0	100.0
2. Portland to Memphis.	25.0	53.3	33.3
3. Denver to Philadelphia. **Search by Airport**.	87.5	93.3	88.9
4. Houston to Chicago. **Search by Airport**.**	62.5	73.3	100.0
5. Newark to Burlington.	50.0	73.3	88.9
6. Las Vegas to New York. **View by Major Airport**.	75.0	93.3	100
7. Phoenix to Dallas.*	12.5	73.3	77.8
8. Cincinnati/Northern Kentucky to Detroit. **View by Region**.	75.0	80.0	88.9
9. Pittsburgh to Washington DC. **Site Map**.	75.0	86.7	100.0
10. New York to San Jose. **Search by Airport**.	75.0	80.0	100.0
11. Orlando to St. Louis. **Search by Airport**.	87.5	86.7	77.8
12. Houston to Tulsa.	87.5	86.7	100.0
Using information available on the site, provide the definitions of the following aviation-related terms or abbreviations:			
13. CIGS	87.5	93.3	88.9
14. MULTI-TAXI	87.5	86.7	88.9
15. VOL	75.0	93.3	100.0
Using information available on the site, who should a visitor contact to obtain information about the following:			
16. Status of an individual flight	100.0	78.6	87.5
17. Why an individual airport was closed	50.0	26.7	44.4

** $p < .10$, two-tailed. ** $p < .05$, two-tailed.*

who in turn were better than the novices. We did not find the same pattern for the aviation term or contact information questions.

We grouped the questions to analyze performance on the different subsets of questions. For the 12 questions that asked users to find specific delay information, novices, moderate-level users, and experts answered 65.6%, 81.7%, and 88% of the questions correctly, F (2, 29) = 5.04, p = .01. Post hoc pairwise contrasts indicated experts and moderate-level users were better able to find delay information than novices (p = .005 and p = .021, respectively).

We further divided the 12 delay questions into subcategories based on search method. For the subset of questions that allowed people to find information using their preferred search method, we found an effect of expertise on user performance, F (2, 29) = 9.93, p = .001. Experts and moderate users performed better than novices when searching for delay information using their preferred search method, answering an average of 80% and 77.3% of the questions correctly, while novices only answered an average of 50% correctly (p < .001 for both post hoc pairwise comparisons).

For the four delay questions that asked users to specifically use the **Search by Airport** method, novices, moderate users, and experts answered 78.1%, 83.3%, and 91.7% of them correctly. Although these results were not statistically significant, they demonstrated the same trend as the other sets.

Post-Session Questionnaire: Overall Analysis

The Post-Session Questionnaire asked the participants to rate their subjective experience with the fly.faa.gov Web site using 6-point scales. Except for the question asking about the level of detail, higher ratings indicated positive responses and lower ratings indicated negative responses. For the question that asked the users how detailed the information on the site was, a rating of 1 indicated too little detail and a 6 indicated too much detail. For these summaries, we omitted responses from the participants who chose more than one number on the rating scale. The ratings indicated that the participants thought it was fairly easy to find information on the site (M = 4.4, SD = .8) and that they understood information once they found it (M = 4.8, SD = 1.0). The participants also found it fairly easy to navigate between pages on the site (M = 4.9, SD = 1.2) and found the design of the site to be consistent (M = 4.9, SD = 1.0). They indicated that there was somewhat too much detail (M = 3.9, SD = 0.8), but that information on the site was fairly readable (M = 4.8, SD = 1.1). Finally, they indicated that, overall, they were mostly satisfied with the site (M = 4.7, SD = 0.8). When we compared satisfaction ratings to actual performance, it was apparent that participants were not able to accurately estimate performance, given that they answered an average of 20.1% questions incorrectly. However, despite their performance, the participants still reported high satisfaction with the site. Given this dissociation between performance and satisfaction, it is important that usability experts evaluate not just user satisfaction, but actual user performance, when evaluating a Web site.

Post-Session Questionnaire: Analysis by Level of Expertise

We found no significant differences in the ratings between experts, moderate-level users, and novices. There were, however, some interesting trends in the data. The ratings on information comprehensibility indicated that experts found the information to be somewhat more comprehensible than moderate-level users, who, in turn, found the information to be more comprehensible than novices. In evaluating design and layout consistency, the experts were the least satisfied with the design consistency, with novices being the most satisfied, and moderate users falling somewhere in the middle. For the ratings on the level of detail, experts gave the highest ratings (i.e., slightly too much detail), with novices giving the lowest ratings (i.e., slightly too little detail), and moderate users falling in the middle (i.e., an appropriate level of detail).

Rating of Usability Issues

Using comments and questionnaire ratings made by the participants, along with our observations of the participants while they completed the User Script, we compiled a consolidated list of usability issues and rated the severity of each issue (for a comprehensive list, see Friedman-Berg, Allendoerfer, & Pai, 2007). When rating the severity of each problem, we considered the following factors (Nielsen, 2003).

1. **Frequency:** Is the problem very common or very rare?
2. **Impact:** How easy is it for the users to overcome the problem when navigating through the Web site?
3. **Persistence:** Can users overcome the problem once they know about it, or will the problem bother users repeatedly?

The researchers rated each issue as having high, medium, or low frequency, impact, and persistence, and then used these three ratings to determine a severity rating from 0 to 5. The severity rating scale was adapted from Nielsen (2003).

0 = I don't agree that this is a usability problem at all
1 = minor/ cosmetic problem only: not necessary to fix, should be given lowest priority
2 = usability problem: small benefit from fixing, should be given low priority
3 = moderate usability problem: moderate benefit from fixing, should be given medium priority
4 = major usability problem: important to fix, should be given high priority
5 = usability catastrophe: extremely important to fix, should be given highest priority

After each researcher independently assigned a severity rating for each issue, we averaged them to compute a consolidated severity rating (Nielsen, 2003). These consolidated severity ratings provide a good estimate of additional usability efforts needed when developers establish priorities for future enhancements. We rank ordered the usability issues from those having the highest severity rating to those having the lowest.

The following section discusses the eight usability issues that had the highest severity rating and provides suggestions and design recommendations regarding how these issues could be resolved. User interface design standards and best practices drive these suggestions (Ahlstrom & Longo, 2003). In some cases, we developed simple prototypes to demonstrate potential design concepts that designers could use to remediate some of these issues.

Issue 1: User Confusion Regarding Delay Types

The primary purpose of fly.faa.gov is to provide travelers with airport delay information. For example, a traveler going from Philadelphia to Miami might want to find out about departure delays at PHL and arrival delays at MIA. The traveler also might have some interest in the causes of delays, which can include factors like weather, airport construction, and traffic flow programs. However, the difference between delay types was not readily apparent to many participants. For example, one question asked users to find information about delays at their arrival destination. The arrival airport had no arrival delays, but did have general departure delays. Because the instructions indicated that they were arriving at that airport, the participants should have focused on the lack of an arrival delay, but only 40.6% of the participants answered this question correctly. Those who answered incorrectly seemed to be looking at the departure delay, which indicated that they did not understand which delays were relevant for them. This issue received a mean severity rating of 4.3, *SD* = 0.5.

It is important that the site provide users with the information they want without requiring them to understand difficult air traffic concepts. We also found that novices had greater difficulty in finding delay information than both moderate level users and experts. This was likely due to novice users not understanding more technical concepts. We recommend that the site not try to present difficult concepts to the lay public, but instead present information in a less technical manner. For instance, instead of referencing ground delay programs as the cause of a delay, the site could indicate that a delay was due to congestion. For users seeking more detailed information, the Web site could provide additional information about ground delay programs using links to additional pages.

Because the participants were not always able to identify relevant delays, we recommend that the site provide users with a capability that gives them easy access to pertinent delay information. For example, the site might provide an interactive tool that allows users to input departure and arrival airports or click on city pairs to generate a single report on relevant delays for air traffic traveling between a pair of airports.

Issue 2: Information Presentation: Clutter and Redundant Information

The participants' comments and researchers' observations suggested that there was too much information on the typical search results page (see Figure 2). This issue received a mean severity rating of 4.3, *SD* = 0.5. The site sometimes presented information for a single airport in multiple places on the same page. The information was dense, used too much text, and was not well organized. In many instances, the participants had difficulty finding the delays that were relevant for them. Displaying so much information can be especially problematic when users are in a hurry to find information. Users may scan too quickly and get lost. They may read the wrong line, overlook information they are looking for, or see a big block of text and give up.

We recommend simplifying and reorganizing these pages to make it easier for users to find and understand information on the page. The page could use a tabular layout arranged in columns and organized by arrivals and departures (see Figure 3). Much of the text information is not useful, creates clutter, and should therefore be removed. Because the distinction between general departure delays and destination-specific delays is not clear to users, it should be deemphasized or eliminated. Finally, all delay information related to an individual airport should be consolidated.

Presenting two sets of delay information for one airport, especially if the data are inconsistent, is confusing. The Web site should avoid going into too much technical detail regarding the causes of delays. It might instead use icons or graphics (e.g., clouds with snow, clouds with rain) to depict weather or other causes of delays. The Web site could offer links to additional information for advanced users.

Issue 3: Overuse of Aviation-Related Acronyms and Jargon

The site uses too many aviation-specific acronyms and jargon when providing specific information about the causes of delays. This issue received a mean severity rating of 4.0, *SD* = 0.0. Aviation-specific acronyms, abbreviations, and jargon are difficult for the general public to understand, and the glossary is difficult to find. The average user of the Web site may never be aware that it exists.

Figure 2. Crowded Airport Status Information page

AIRPORT STATUS INFORMATION
provided by the FAA's Air Traffic Control System Command Center

Waterloo Muni Airport (ALO) Real-time Status

The status information provided on this site indicates general airport conditions; it is not flight-specific. Check with your airline to determine if your flight is affected.

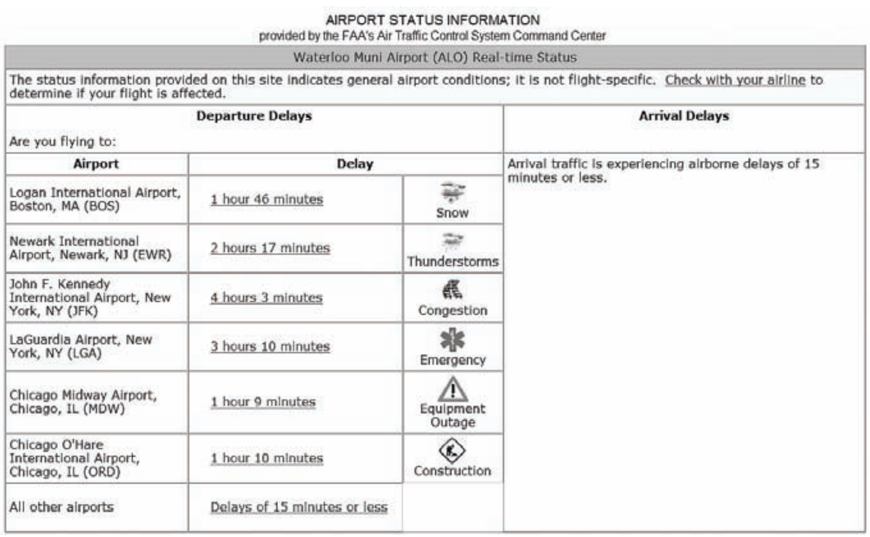
Delays by Destination:

- Due to WEATHER/WIND, departure traffic destined to **General Edward Lawrence Logan International Airport, Boston, MA (BOS)** is currently experiencing delays averaging **1 hour and 46 minutes.**
- Due to WEATHER/WIND, departure traffic destined to **Newark International Airport, Newark, NJ (EWR)** is currently experiencing delays averaging **2 hours and 17 minutes.**
- Due to WEATHER/LOW CIGS, departure traffic destined to **John F Kennedy International Airport, New York, NY (JFK)** is currently experiencing delays averaging **4 hours and 3 minutes.**
- Due to WEATHER/WIND, departure traffic destined to **La Guardia Airport, New York, NY (LGA)** is currently experiencing delays averaging **3 hours and 10 minutes.**
- Due to WEATHER/SNOW, departure traffic destined to **Chicago Midway Airport, Chicago, IL (MDW)** is currently experiencing delays averaging **1 hour and 9 minutes.**
- Due to WEATHER/SNOW, departure traffic destined to **Chicago OHare International Airport, Chicago, IL (ORD)** is currently experiencing delays averaging **1 hour and 10 minutes.**
- Due to WEATHER/WIND, departure traffic destined to **Philadelphia International Airport, Philadelphia, PA (PHL)** is currently experiencing delays averaging **3 hours and 38 minutes.**
- Due to WEATHER/WIND, departure traffic destined to **Teterboro Airport, Teterboro, NJ (TEB)** is currently experiencing delays averaging **1 hour and 3 minutes.**
- Due to WEATHER/TSTMS, departure traffic destined to **Newark International Airport, Newark, NJ (EWR)** will not be allowed to depart until at or after 4:15 pm CST.
- Due to WEATHER/TSTMS, departure traffic destined to **La Guardia Airport, New York, NY (LGA)** will not be allowed to depart until at or after 4:15 pm CST.
- Due to WEATHER/ENROUTE WX, departure traffic destined to **Philadelphia International Airport, Philadelphia, PA (PHL)** will not be allowed to depart until at or after 4:15 pm CST.

General Departure Delays: Traffic is experiencing gate hold and taxi delays lasting 15 minutes or less.

General Arrival Delays: Arrival traffic is experiencing airborne delays of 15 minutes or less.

Figure 3. Airport Status Information in a redesigned format



AIRPORT STATUS INFORMATION
provided by the FAA's Air Traffic Control System Command Center

Waterloo Muni Airport (ALO) Real-time Status

The status information provided on this site indicates general airport conditions; it is not flight-specific. Check with your airline to determine if your flight is affected.

Departure Delays Are you flying to:			Arrival Delays
Airport	**Delay**		Arrival traffic is experiencing airborne delays of 15 minutes or less.
Logan International Airport, Boston, MA (BOS)	1 hour 46 minutes	Snow	
Newark International Airport, Newark, NJ (EWR)	2 hours 17 minutes	Thunderstorms	
John F. Kennedy International Airport, New York, NY (JFK)	4 hours 3 minutes	Congestion	
LaGuardia Airport, New York, NY (LGA)	3 hours 10 minutes	Emergency	
Chicago Midway Airport, Chicago, IL (MDW)	1 hour 9 minutes	Equipment Outage	
Chicago O'Hare International Airport, Chicago, IL (ORD)	1 hour 10 minutes	Construction	
All other airports	Delays of 15 minutes or less		

When the participants had to find the definition of three aviation-related terms, 16% were unable to find the definition for at least one of them. Therefore, we recommend eliminating the use of these terms when they are not essential. This would eliminate unnecessary detail, simplify the site, and make it easier to use and understand.

Issue 4: User Confusion with Using the View by Region Maps

The fly.faa.gov Web site provides users with a **View by Region** search function that allows users to look up airports by searching in different geographic regions. These regions include the Northeast, North Central, Northwest, Southeast, South Central, and Southwest regions, along with Alaska and Hawaii. When a user uses the **View by Region** function, they are taken to a map that contains only states that are part of a region. However, it is not easy for someone with little knowledge of geography to determine the region for a particular state. The participants got lost when looking for airports that were not on the main U.S. map because they were unable to determine the relationship of regional maps to the main U.S. map. This was especially difficult for states such as Ohio that lie at the edge of a region. These issues make the **View by Region** method difficult for the general public to use and the participants found the **View by Region** maps to be confusing. This issue received a mean severity rating of 4.0, *SD* = 0.0.

One question asked the participants to find delay information for an airport that was not available on the main map or on the **View by Major Airport** menu. Only 71.9% of the participants found the correct answer for this question, indicating that the participants had some difficulty finding information when they needed to drill down on the maps.

There are several recommendations that could alleviate some of the issues related to the use of the **View by Region** method. First, the site could place an outline around the different regions or use color coding to highlight the different regions on the U.S. map. This would help users identify which states belong in which region. The site could display split portions of the main U.S. map on the same page to better orient users to the different regions. To familiarize people with relevant geographic information, the site could label states, both on the main U.S. map and on the smaller regional maps. The site could also offer users a drop-down menu that listed the various airports by state.

Issue 5: Lack of User Knowledge Regarding Three-Letter Airport Identifiers

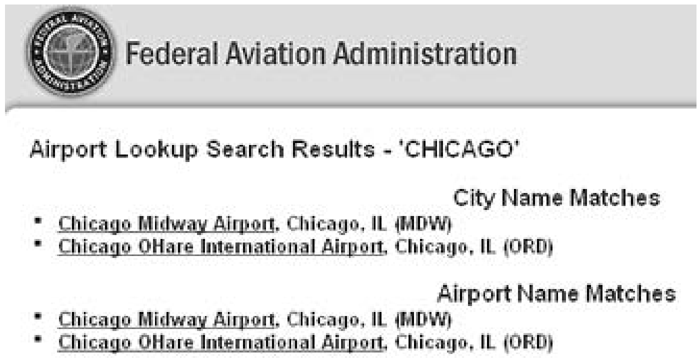
All commercial airports have three-letter identifiers, and using them is an efficient way to obtain delay information about an airport. The site provides a function that allows users to type a three-letter identifier directly into the **Search by Airport** text box, which will take the user to the details page for that airport. It also provides cues to site users by labeling airports on the main U.S. map with their three-letter identifiers (see Figure 1). However, many participants did not know the correct three-letter identifiers for airports and did not use the cues on the main map to determine the correct identifier. This issue received a mean severity rating of 3.3, *SD* = 0.6.

The site should emphasize that the **Search by Airport** text box accepts regular airport names and city names in addition to three-letter identifiers. Although the **Search by Airport** text box does have a label indicating that users can enter city, airport code, or airport name information in this field, we recommend that the Web site provide the user with specific examples to highlight and better explain the different search options.

Issue 6: The Search by Airport Function Returns Redundant and Irrelevant Results

City name searches using the **Search by Airport** function generate an intermediate results page that lists multiple airports. These listings often contain redundant and irrelevant results. This issue received a mean severity rating of 3.3, *SD* = 0.6. For example, a search for Chicago generates a search results page listing two airports: Midway and O'Hare International. The site lists each result twice, once under **City Name Matches** and once under **Airport Name Matches** (see Figure 4). This format is confusing and users may not

Figure 4. The www.fly.faa.gov results page for a Search by Airport search for Chicago



realize that both links take them to the same information. Some participants questioned why the site listed an airport twice. We recommend that the **Airport Lookup Search Results** page consolidate search results and list airports only once in any search results list.

Issue 7: User Spelling and Misspellings and Their Impact on the Search by Airport Function

User spellings and misspellings can have a serious impact on the **Search by Airport** function. In some instances, the correct spelling does not work, but a misspelling does. For example, typing *O'Hare* does not return any results, but *Ohare* does. Typing *LaGuardia* returns no results, but *La Guardia* does. In addition, common misspellings do not produce any results at all, even when the system could provide reasonable guesses about what the user intended. For example, *Newyork* does not produce any search results at all. This issue received a mean severity rating of 3.3, *SD* = 0.6. The participants quickly became frustrated and confused when the site did not return any search results for correct spellings or reasonable misspellings. The search function should always result in a hit when the correct spelling is used, should provide "best guess" search result even when users make spelling mistakes, and should ignore spacing errors.

Issue 8: Inconsistent Use of Pop-up Windows

The fly.faa.gov Web site is inconsistent in its use of pop-up windows. When users access information using the **Search by Airport** method or when they click on the color-coded dots on the main site map, the Web site displays the search results in a pop-up window. However, when users access information using the **View by Major Airport** method, the site displays the same information in the current browser window rather than in a pop-up window. This issue received a mean severity rating of 3.0, *SD* = 0.0.

During the assessment, some participants accidentally closed the browser by clicking the **Close** button when search results appeared in the main browser window. These participants had become accustomed to results appearing in a pop-up window. When search results appeared in the main browser window, they still reacted as if they were in a pop-up window and accidentally closed down the site, along with the browser.

We recommend that the site be more consistent in how it returns search results and **Airport Status Information** pages. Users become confused when the site responds differently to similar actions. If the standard convention of the site is to bring up search results in pop-up windows, then the site should bring up all search results in pop-up windows.

DISCUSSION

The level of aviation-related expertise had an impact on many aspects of user performance. Experts were more likely than novices and moderate-level users to have had some prior interaction with the fly.faa.gov Web site. They were also better at finding delay information on the Web site. Experts appeared to have a better conceptual understanding of the different types of airport delays than both novices and moderate users. Finally, experts indicated that they found the information on the Web site to be slightly more comprehensible than both novices and moderate level users. Although we realize that there may be some performance decrement for people who have no affiliation with the FAA, we expect that their performance and their issues should be most similar to our novice users.

On the basis of performance differences, we recommend that the primary goal of site designers should be to make the site more usable for people who do not have an aviation background. If people in the general public visit this site without an aviation-related background, we would expect them to have substantial difficulty (a) understanding which delays were relevant for them, (b) understanding how airport delays differ from airline delays, and (c) interpreting much of the jargon used by aviation experts. Although both experts and novices use the site, simplifying the Web site should help all users, not just novices. Links to additional information can be provided for expert users.

Subjective reports indicated that the participants were generally satisfied with the fly.faa.gov Web site, and objective data revealed that they could successfully complete most tasks using the site. By evaluating user performance data in conjunction with user comments and researcher observations, we were able to identify a number of human factors issues with the Web site that we would not have identified by relying solely on subjective data.

After identifying issues, we rated each one in terms of its impact on site usability, discussed each issue in detail, identified supporting data when appropriate, and provided recommendations for improving the usability of the Web site. Many of the suggested improvements should be easy to implement and should further increase user satisfaction and site usability.

CONCLUSION

One of the primary lessons that we learned from this usability evaluation is that developers should not simply rely on subjective reports of usability when evaluating e-government Web sites. It is just as important to observe users interacting with a Web site and collect objective performance data to better identify usability issues. By having people use the Web site to find different types of information, we were better able to identify those areas of the site that caused problems for users. To encourage organizations to perform usability evaluations on e-government sites, we should ensure that they provide value by identifying important usability issues that can be remedied through redesign. As we saw in this evaluation, subjective reports often fail to identify these issues. If research on Web site usability fails to identify significant usability issues, it is likely that such evaluations will not be used.

We also found that having researchers rate the severity of usability issues improved our evaluation. Future e-government usability assessments could reap benefits by using this technique. Many times, when a usability assessment is performed, the output of the assessment is a laundry list of issues that usability experts present to site designers. If guidance is given on issue severity or criticality, it is usually ad hoc and is not derived using any formal methodology. By requiring evaluators to explicitly rate each item on frequency, impact, and severity, they are required to think about how and in what ways the problem will affect the user. This user-centric focus is the key element of this methodology. It allows site evaluators to provide designers with a roadmap of how they can best focus their effort to provide a more optimal user experience. Additionally, we recommend that usability assessments use more than one evaluator to make severity ratings. We found that different evaluators might have different priorities, but by using combined severity ratings from three or more evaluators, you can increase the reliability of the ratings (Nielsen, 2003).

By employing an evaluation processes like the one used in this study to evaluate e-government sites, whether they are G2B sites, G2C sites, or G2E sites, designers and system developers can better allocate limited resources during the design process. In general, it is important that e-government Web site designers take into consideration the demographics of those who will use their Web site or application. If an e-government Web site or application, initially targeted to users with a specific area of expertise, is going to be redesigned for use by the general public, the site must be evaluated for usability. Based on the results of such an evaluation, changes need to be made to ensure that the site is usable by the broadest possible audience.

REFERENCES

Ahlstrom, V., & Longo, K. (Eds.). (2003). *Human factors design standard for acquisition of commercial-off-the-shelf subsystems, non-developmental items, and developmental systems* (DOT/FAA/CT-03/05/HF-STD-001). Atlantic City International Airport, NJ: FAA William J. Hughes Technical Center.

Bannister, F. (2007). The curse of the benchmark: An assessment of the validity and value of e-government comparisons, *International Review of Administrative Services, 73*, 171-188.

Barne, S. J., & Vidgen, R. T. (2006). Data triangulation and Web quality metrics: A case study in e-government source. *Information and Management, 4*, 767-777.

Bertot, J. C., & Jaeger, P. T. (2006). User-centered e-government: Challenges and benefits for government Web sites. *Government Information Quarterly, 23*, 163-168.

Bush, G.W. (2007, November 15). President Bush discusses aviation congestion. *Office of the Press Secretary* [Press release]. Retrieved February 22, 2008, from http://www.whitehouse.gov/news/releases/2007/11/20071115-6.html

Carter, L., & Bélanger, F. (2005). The utilization of e-government services: Citizen trust, innovation and acceptance factors. *Information Systems Journal, 15*, 5-25.

Foley, P. (2005). The real benefits, beneficiaries and value of e-government. *Public Money & Management, 25*, 4-6.

Allendoerfer, K., Friedman-Berg, F., & Pai, S. (2007). *Usability assessment of the fly.faa.gov Web site* (DOT/FAA/TC-07/10). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

Gil-García, J. R., & Pardo, T. A. (2005). E-government success factors: Mapping practical tools to theoretical foundations. *Government Information Quarterly, 22*, 187-216.

Heeks, R., & Bailur, S. (2007). Analysing eGovernment research. *Government Information Quarterly, 22*, 243-265.

Hiller, J., & Belanger, F. (2001). *Privacy strategies for electronic government*. E-Government series. Arlington, VA: Pricewaterhouse Coopers Endowment for the Business of Government.

Ho, A. (2002). Reinventing local government and the e-government initiative. *Public Administration Review, 62*, 434-444.

Horst, M., Kuttschreutter, M., & Gutteling, J. M. (2007). Perceived usefulness, personal experiences, risk perception, and trust as determinants of adoption of e-government services in the Netherlands. *Computers in Human Behavior, 23*, 1838-1852.

Howell, D. C. (2007). *Chi-square with ordinal data*. Retrieved January 22, 2007, from http://www.uvm.edu/~dhowell/StatPages/More_Stuff/OrdinalChisq/OrdinalChiSq.html

Hung, S. Y., Chang, C. M., & Yu, T. J. (2006). Determinants of user acceptance of the e-government services: The case of online tax filing and payment system, *Government Information Quarterly, 23*, 97-122.

Irani, Z., Love, P. E. D, & Montazemi, A. (2007). E-Government: Past, present, and future. *European Journal of Information Systems, 16*, 103-105.

Nielsen, J. (2003). *Severity ratings for usability problems*. Retrieved April 15, 2006, from http://www.useit.com/papers/heuristic/severityrating.html

Reddick, C.G. (2005). Citizen interaction with e-government: From the streets to servers? *Government Information Quarterly, 22*, 38-57.

Salem J. A. (2003). Public and private sector interests in e-government: A look at the DOE's PubSCIENCE. *Government Information Quarterly, 20*, 13-27.

Schelin, S. H. (2003). E-government: An overview. In G. D. Garson (Ed.), *Public information technology: Policy and management issues* (pp. 120-137). Hershey, PA: Idea Group Publishing.

Thomas, J. C., & Streib, G. (2003). The new face of government: Citizen-initiated contacts in the era of e-government. *Journal of Public Administration Research and Theory, 13*, 83-102.

Yildiz, M. (2007). E-government research: Reviewing the literature, limitations, and ways forward. *Government Information Quarterly, 24*, 646-665.

This work was previously published in International Journal of Electronic Government Research, Vol. 5, Issue 1, edited by V. Weerakkody, pp. 64-79, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 5.8
Quality Enhancing the Continued Use of E-Government Web Sites:
Evidence from E-Citizens of Thailand

Sivaporn Wangpipatwong
Bangkok University, Thailand

Wichian Chutimaskul
King Mongkut's University of Technology Thonburi, Thailand

Borworn Papasratorn
King Mongkut's University of Technology Thonburi, Thailand

ABSTRACT

This study empirically examines Web site quality toward the enhancement of the continued use of e-government Web sites by citizens. The web site quality under examination includes three main aspects, which are information quality, system quality, and service quality. The participants were 614 country-wide e-citizens of Thailand. The data were collected by means of a web-based survey and analyzed by using multiple regression analysis. The findings revealed that the three quality aspects enhanced the continued use of e-government Web sites, with system quality providing the greatest enhancement, followed by service quality and information quality.

INTRODUCTION

Electronic government, so called e-government, has been broadly defined as the use of information and communication technology (ICT) to transform government by making it more accessible, effective, and accountable (infoDev & CDT, 2002). The Internet is indeed the most powerful and popular means of delivering e-government. Hence, Web sites have been employed as a platform for delivering a wide range of government services electronically.

By using e-government Web sites, citizens can conveniently access government information and services and gain greater opportunities to participate in the democratic process (Fang, 2002).

Citizens can access government information and services anywhere and anytime. Thus, the time spent in traveling and waiting is reduced. From the government's point of view, the more citizens that use e-government Web sites, the more operation and management costs are reduced.

To obtain these benefits, the initial adoption and subsequent continued use of e-government Web sites by citizens are required. In general, an information system indicated that its eventual success depends on its continued use rather than first-time use (Bhattacherjee, 2001; Limayem, Hirt, & Cheung, 2003). Likewise, initial use of e-government Web sites is an important indicator of e-government success. However, it does not necessarily lead to the desired outcome unless a significant number of citizens move beyond the initial adoption and use e-government Web sites on a continual basis. To enhance the continued use, this study proposes that quality of e-government Web sites is one significant factor.

According to DeLone and McLean (2002), the three quality aspects, information quality, system quality, and service quality, are the determinants that effect user's intention to use an information system. In practice, these three aspects have been employed to study the initial intention to use the information system and to evaluate the quality of information system (e.g., Lee & Kozar, 2006; Negash, Ryan, & Igbaria, 2003; Wilkin & Castleman, 2002). However, there is a lack of prior research that uses information quality, system quality, and service quality to examine the continued use in the context of e-government Web sites.

This study therefore aims to examine the Web site quality toward enhancement of the continued use of e-government Web sites by citizens. The population of interest for this study is e-citizens of Thailand, a group of citizens who has experienced Thailand's e-government Web sites. The reason that makes Thailand an ideal place to study is that e-government is considered a new innovation to Thai citizens and is conceived as a fundamental element to encourage the country development.

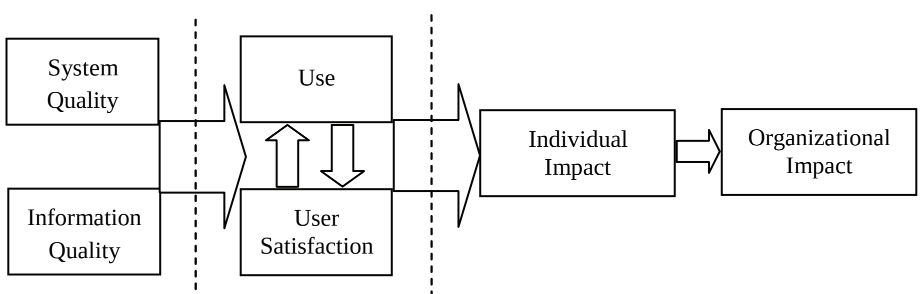
In the next section, the background of this study is briefly reviewed. Thereafter, the research model and hypotheses development, research methodology, and data analysis are presented. Finally, the discussion, limitations, and suggestions for future research are given.

BACKGROUD OF STUDY

DeLone and McLean's Information System Success Model

In order to ascertain the success of an information system, DeLone and McLean (1992) proposed the Information System Success Model (referred hereafter as the 'D&M IS Success Model') as shown in Figure 1. The model asserts that system quality and information quality are the determinants of system

Figure 1. Original DeLone and McLean's Information System Success Model (DeLone & McLean, 1992, p.87)



use and user satisfaction which effect individual and organizational impact respectively.

However, Pitt, Watson, and Kavan (1995) noticed that commonly used measures of information system effectiveness focus on the products, rather than the services. They then proposed that service quality needs to be considered as an additional measure of the D&M IS Success Model. DeLone and McLean (2002) therefore reformulated the D&M IS Success Model by including service quality as an additional determinant that effects the use and user satisfaction as shown in Figure 2.

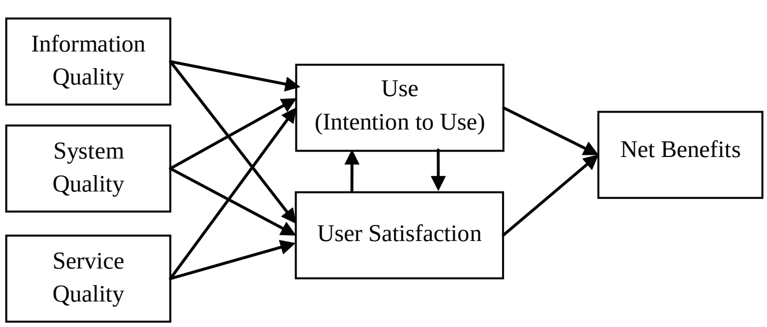
D&M IS Success Model has become popular for the specification and justification of the measurement of the dependent variable in an information system research. In the summer of 2002, a citation search yielded 285 refereed papers in journals and proceedings that referenced D&M IS Success Model during the period 1993 to mid-2002 (DeLone & McLean, 2003). In practice, a number of empirical studies (e.g., Iivari, 2005; Molla & Licker, 2001; Seddon & Kiew, 1994) gave support for the associations among the measures identified in the D&M IS Success Model.

E-Government in Thailand

Like many countries, Thailand has been fully cognizant of both the potentials and the benefits of e-government. Since 1996, the e-government development has been driven by the National Information Technology Committee (NITC). Several programs, such as computer training for mid-level officers, specifying minimum requirements of information technology equipment for government agencies, and the establishment of a Chief Information Officer (CIO) in the public sector, have been imposed to support and promote this initiative (Chamlertwat, 2001). The SchoolNet Thailand (a national school informatization program to empower all schools to access a large pool of information resources using the Internet), Government Information Network or GINet (a government backbone network to facilitate intra- and inter- agencies communication and information exchanges), and the development of legal infrastructure to support the application of information technology have also been initiated under the first National IT Policy Framework for the year 1996–2000 (IT 2000) (Ateetanan, 2001).

In March 2001, the two-year period e-government project was established by NITC to establish a framework for building up e-government and to implement some pilot projects. In March 2002, the National IT Policy Framework for the year 2001–2010 (IT 2010) was approved by the cabinet and e-government was a manifest flagship, in addition to e-industry, e-commerce, e-society, and e-education. Subsequently, in September 2002, the cabinet further endorsed the first National ICT Master Plan for the year 2002–2006. The master

Figure 2. Reformulated DeLone and McLean's Information System Success Model (DeLone & McLean, 2002, p.2974)



plan devises seven key strategies. One of which is e-government (NECTEC, 2003).

As a result, e-government has been developing rapidly in Thailand. In early 2004, NECTEC initiated the first on-line survey of government e-services. The survey revealed that all 267 government agencies have Web sites to provide information to the public (NECTEC, 2005a). With accordance to the global ranking, the E-government Readiness Survey of United Nations (United Nations, 2003; United Nations, 2004; United Nations, 2005) reported that Thailand owned an E-government Readiness Ranking at 56 from 191 global countries in 2003, moved up to the 50th rank in 2004, and finally at the 46th rank in 2005.

In terms of overall usage, there is not much usage of e-government Web sites compared to other Web site categories. According to the Truehits 2005 Award (Truehits, 2006) government Web sites occupied the 15th rank from 19 Web site categories and had proportion of usage only 1.64%, which went down 0.13% from the previous year.

At this stage it is not clear if Thai citizens will continue using Thailand's e-government Web sites. Hence, better understanding of the factors that enhance citizens to continue using e-government Web sites can create greater value for Thailand's government and also other governments all over the world.

RESEARCH MODEL AND HYPOTHESES DEVELOPMENT

Based on the review of the aforementioned literature, the conceptual research model used to guide this study is proposed as shown in Figure 3. The model is based on the three quality aspects of the D&M IS Success Model, adapting to the e-government Web site context. In the following, the meaning of all constructs and the theories supporting the relationships are presented.

Information Quality

According to DeLone and McLean (1992), information quality is concerned with the measure of the information that the system produces and delivers. When applied to this study, the information quality focuses on characteristics of information produced by e-government Web sites.

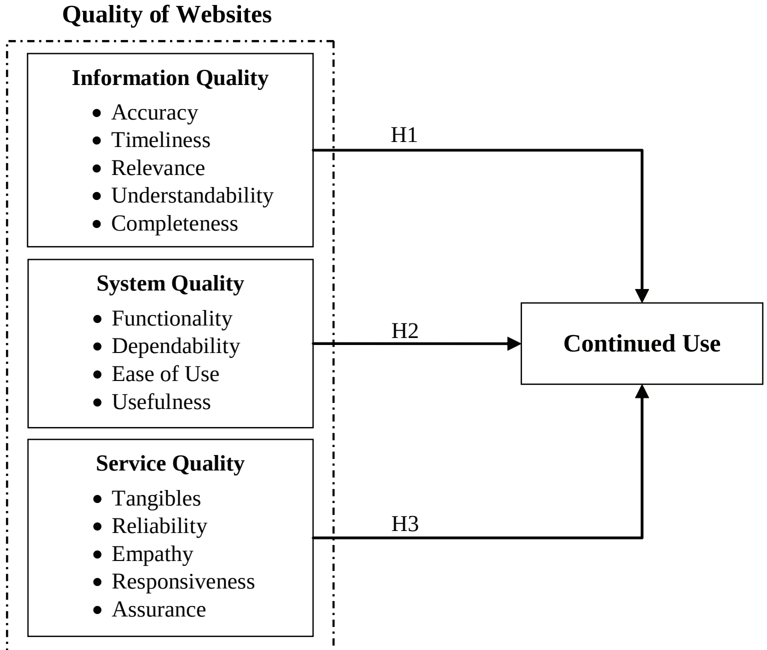
Quality of information is believed to be the most salient factor for predicting customer decision-making behavior (Jeong & Lambert, 2001) and user intention to use a particular system (DeLone & McLean, 1992, 2002; Molla & Licker, 2001). Furthermore, information quality has long been found associated with customer or user satisfaction in previous empirical studies (Seddon & Kiew, 1994; Spreng, MacKenzie, & Olshavsky, 1996; Szymanski & Hise, 2000; Negash, Ryan, & Igbaria, 2003; Iivari, 2005).

Concerning to the case of e-government, the quality of information on the Web sites is very significant since most citizens use e-government Web sites for informational purposes (Accenture, 2004), and the first phase of e-government implementation is to publish government information (infoDev & CDT, 2002). If e-government Web sites contain low information quality, they are useless. Furthermore, high quality information encourages citizens to use the Web sites (Cullen & Hernon, 2004). Hence, the following hypothesis is proposed:

H1: *Information quality of e-government Web sites enhances the continued use of e-government Web sites.*

According to a review of related literature (Bailey & Pearson, 1983; Doll & Torkzadeh, 1988; Wang & Strong, 1996), the fundamental

Figure 3. Conceptual research model



dimensions of information quality is composed of five dimensions: accuracy, timeliness, relevance, understandability, and completeness. This study thereby uses these five dimensions to measure citizens' perceptions toward information quality of e-government Web sites. Table 1 shows a brief definition of each dimension.

System Quality

According to DeLone and McLean (1992), system quality is concerned with the measure of the actual system which produces the output. The system quality in this study therefore focuses on features and performance characteristics of e-government Web sites regarding the quality in use or the citizen's view of quality.

System quality, in the sense of quality in use, has been found as a significant determinant of overall user satisfaction (DeLone & McLean, 1992, 2002; Seddon & Kiew, 1994; Negash, Ryan, & Igbaria, 2003; Iivari, 2005), user acceptance (Bevan, 1999), and system use (DeLone & McLean, 1992, 2002). The more satisfied the user is with the system the more he or she will be inclined to use it. Conversely, if system use does not meet the user's needs, satisfaction will not increase and further use will be avoided (Baroudi, Olson, & Ives, 1986). Therefore, this study postulates that:

H2: *System quality of e-government Web sites enhances the continued use of e-government Web sites.*

Based on a review of related literature (Bailey & Pearson, 1983; Doll & Torkzadeh, 1988), this study identifies and categorizes the characteristics related to the quality in use and user satisfaction into four core dimensions: functionality, dependability, ease of use, and usefulness. Ease of use and usefulness are also excerpted from Davis's (1989) Technology Acceptance Model (TAM). This study therefore uses theses four dimensions to measure citizens' perception toward system

Table 1. Information quality dimensions

Dimension	Definition	Contributing Authors
Accuracy	The information is correct and reliable	Bailey and Pearson (1983) Doll and Torkzadeh (1988) Wang and Strong (1996)
Timeliness	The information is current and timely	
Relevance	The information corresponds to the need and is applicable for the task at hand	
Understandability	The information is clear and easy to comprehend	
Completeness	The information has sufficient breadth and depth for the task at hand	

Table 2. System quality dimensions

Dimension	Definition	Contributing Authors
Functionality	The required functions are available in the system	Bailey and Pearson (1983) Doll and Torkzadeh (1988) Davis (1989)
Dependability	The system is accurate and dependable over time	
Ease of Use	The system can be accessed or used with relatively low effort	
Usefulness	The benefits that the user believes to derive from the system, including convenience, saving time, and saving cost	

quality of e-government Web sites. The definition of each dimension is summarized in Table 2.

Service Quality

Service quality refers to the quality of personal support services provided to citizens through e-government Web sites, such as answering questions, taking requests, and providing sophisticated solutions to citizen's problems. This definition is consistent to service quality of DeLone and McLean (2002, 2004) that concerns the measure of the user support services delivered by the service provider.

Table 3. Service quality dimensions

Dimension	Definition	Contributing Authors
Tangibles	Physical facilities, equipment, and appearance of personnel	Parasuraman, Zeithaml, and Berry (1988)
Reliability	Ability to perform the promised service dependably and accurately	
Empathy	Caring, individualized attention the service provider gives its customers	
Responsiveness	Willingness to help customers and provide prompt service	
Assurance	Knowledge and courtesy of employees and their ability to inspire trust and confidence	

Prior literature on marketing has indicated that service quality is an important determinant of customer satisfaction (Cronin & Taylor, 1992; Bitner, Booms, & Mohr, 1994; DeLone & McLean, 2002, 2004) and repeat patronage (Zeithaml, Berry, & Parasuraman, 1996), especially in pure service situations where no tangible object is exchanged (Parasuraman, Zeithaml, & Berry, 1985; Solomon, Surprenant, Czepiel, & Gutman, 1985). With regards to e-government, service quality is needed since citizens differ in education, knowledge, and experience. The service quality therefore acts as an enabler of the citizen's capability to use e-government Web sites. Hence, this leads to the hypothesis:

H3: *Service quality of e-government Web sites enhances the continued use of e-government Web sites.*

Based on the SERQUAL developed by Parasuraman, Zeithaml, and Berry (1988), the quality of service is composed of five dimensions: tangibles, empathy, reliability, responsiveness, and assurance as defined in Table 3. The SERVQUAL is both a reliable and a valid measure of service quality and is also applicable to a wide variety of service contexts (Parasuraman, Zeithaml, & Berry, 1988). Thus, the SERVQUAL dimensions are used to measure citizens' perception toward system quality of e-government Web sites.

RESEARCH METHODOLOGY

Participants

The participants were 614 e-citizens from five regions of Thailand. The majority of the participants are living in the capital of Thailand (Bangkok) and vicinity (77.36%), followed by the central region (9.77%), the northern region (5.70%), the northeast region (3.58%), and the southern region (3.58%). The dispersion of participants in this study was comparable to the Internet user profile of Thailand, wherein Internet users are concentrated in Bangkok and vicinity, and the rest are distributed in other regions with nearly equivalent proportion (NECTEC, 2005b). Figure 4 shows the participant dispersion in this study compared to the Internet user profile of Thailand.

Demographic characteristics of the overall participants are summarized in Table 4. The proportion of the gender of participants is equal. Most of them are between 21–30 years of age (64.01 %), have a bachelor's degree (54.89%), work in private sectors (58.79%), and have monthly income between 10,001–20,000 Baht (36.64%). About half of participants (62.21%) have experienced Internet for 6-10 years.

In terms of experience with e-government Web sites, the most frequently mentioned experience is searching, inquiry, or complaint (79.32%), followed by online transactions (68.24%) and downloading forms (60.75%). The five most frequently mentioned topics are tax (75.90%), tourism (56.84%), education (55.37%), citizen registration (35.67%) and communication (31.60%). The participants' experience with e-government Web sites is illustrated in Table 5.

Instrument Development

The questionnaire was used as an instrument to gather data from participants. The measurement items for information quality, system quality, and service quality were rated on a 5-point Likert scale (1=Strongly Disagree; 2=Disagree; 3=Neutral; 4=Agree; 5=Strongly Agree). Table 6 lists the measurement items.

The validity of the questionnaire was strengthened through an extensive review of the literature and an agreement among professionals. In addition, the pretest through 25 convenience samples was employed to determine if the intended audiences had any difficulty understanding the questionnaire and whether there were any ambiguous

Figure 4. Dispersion of participants compared to Internet user profile of Thailand

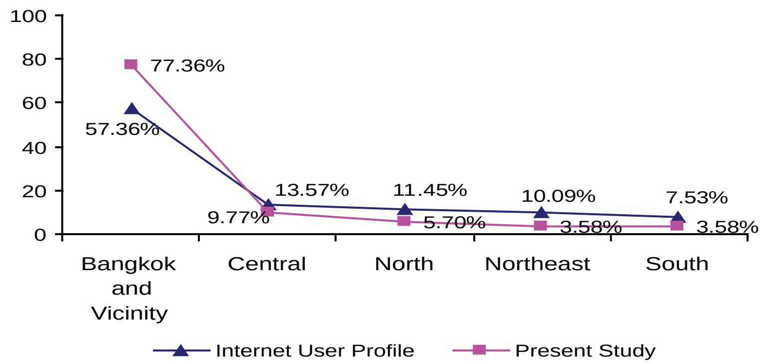


Table 4. Participants' demographic characteristics

Characteristics		**Frequency**	**Percent** [a]
Gender	Female	307	50.00
	Male	307	50.00
Age	21–30 years	393	64.01
	31–40 years	179	29.15
	41–50 years	35	5.70
	Older than 50 years	7	1.14
Level of Education	Bachelor's Degree	337	54.89
	Master's Degree	268	43.65
	Doctorate Degree	9	1.47
Occupation	Student	51	8.31
	Government Employee	113	18.40
	State Enterprises Employee	30	4.89
	Private Sector Employee	361	58.79
	Self Employment	54	8.79
	Unemployed	3	0.49
	Retiree	2	0.33
Monthly Income	Less than or equal to 5,000 Baht	12	1.95
	5,001–10,000 Baht	75	12.21
	10,001–20,000 Baht	225	36.64
	20,001–30,000 Baht	130	21.17
	More than 30,000 Baht	172	28.01
Years on Internet	Less than 1 year	2	0.33
	1–5 years	92	14.98
	6–10 years	382	62.21
	More than 10 years	138	22.48

[a] *Percentages subject to rounding*

Table 5. Participants' experience with e-government Web sites

Experience [a]		**Frequency**	**Percent** [b]
Type of Use	Search / Inquiry / Complaint	487	79.32
	Conduct Online Transaction	419	68.24
	Download Form	373	60.75
Topic of Use	Tax	466	75.90
	Tourism	349	56.84
	Education	340	55.37
	Citizen Registration	219	35.67
	Communication	194	31.60
	Research	155	25.24
	Transportation	140	22.80
	Employment	138	22.48
	Recreation	137	22.31
	Health	130	21.17
	Commerce	110	17.92
	Foreign Affairs	97	15.80
	Housing	92	14.98
	Safety and Regulation	88	14.33
	Government Welfare	82	13.36
	Agricultural	71	11.56
	Public Utility	66	10.75
	Industry	65	10.59
	Politics	59	9.61
	Family and Community	40	6.51

[a] *Participants could tick all that apply*

[b] *Percentages subject to rounding*

or biased questions. Based on the feedback of the pretest, one reverse word item was dropped since it caused confusion.

To ensure the measurement items are measuring the same construct, the most widely used measure named Cronbach's alpha was employed for each construct to measure the internal consistency among all items. As observed from Table 7, the reliability analysis gave alpha coefficients exceeding .70 which are typically regarded as an acceptable reliability coefficient (Nunnaly, 1978); one exception was for the scale of completeness which is a little bit lower than .70. However, the lower limit of acceptability may decrease to .60 in exploratory research (Hair, Anderson, Tatham, & Black, 1998). Therefore, the items measuring the constructs were acceptable and reliable.

Data Collection

In order to eliminate costs, data coding time, and human-error, and to easily reach citizens in different geographic areas across the country, a web-based survey with a probability list-based method, which samples participants based on a list, was employed to collect data. The bundled script program was used to check and advise participants, thereby ensuring that all items in the questionnaire were filled in completely and appropriately. To take steps toward ensuring the integrity of the data, the IP address of each participant and the time used for completing the survey were recorded.

After the survey was uploaded to the server, the 3,600 invitations for participation, including a link to the Web site, were randomly emailed to an alumni mailing list in a variety of faculties (e.g., science, agriculture, engineering, pharmacology, liberal arts, business administration, information technology, and humanities) of five universities across five regions of Thailand. The selection of these five universities resulted from a three-stage sampling. First, a stratified sampling was performed to cluster Thailand into five regions. Second, a simple random sampling was done to select a university corresponding to each of the five regions. Third, a simple random sampling was employed to select some email addresses corre-

Table 6. List of measurement items

Construct	Dimension	Description
Information Quality	Accuracy	Using e-government Web sites enables me to have accurate information.
		I can trust the information on e-government Web sites.
	Timeliness	Using e-government Web sites enables me to access the newest information.
		Using e-government Web sites enables me to access up-to-date information when compare to deal with other sources.
	Relevance	Using e-government Web site enables me to have information that is relevant to the site.
		Using e-government Web site enables me to have the information that I need.
	Understandability	Information on e-government Web sites is easy for me to comprehend.
		Information on e-government Web sites is clear for me.
	Completeness	Using e-government Web sites enables me to access adequate information.
		I find information on e-government Web sites is sufficient for the task at hand.
System Quality	Functionality	E-government Web sites provide necessary information and forms for downloading.
		E-government Web sites provide necessary online transactions.
		E-government Web sites provide service functions that I need.
		E-government Web sites present a variety of services
	Dependability	E-government Web sites perform right at the first time.
		Every time I request e-government Web sites, the Web sites are available.
		The government will not misuse my personal information.
		I feel safe in my online transaction with e-government Web sites.
	Ease of Use	I can easily login to e-government Web sites.
		Getting the information that I want from e-government Web sites is easy.
		It is easy for me to complete transactions through e-government Web sites.
		The organization and structure of e-government Web sites is easy to follow.
	Usefulness	Using e-government Web sites enable me to accomplish tasks more quickly.
		The results of using e-government Web sites are apparent to me.
		Using e-government Web sites can cut traveling expense.
		Using e-government Web sites can lower traveling and queuing time.
		Using e-government Web sites enable me to do business with the government anytime, not limited to regular business hours.

continued on following page

Table 6. continued

Service Quality	Tangibles	If I need help, I can find a way to reach a government staff such as email or webboard on e-government Web sites.
		There is staff who will respond to my request indicated on e-government Web sites.
	Reliability	If I send a request via email or webboard to the government, I will receive the right solution from the government staff.
		If I send a request via email or webboard to the government, I will receive the solution that matches to my needs from the government staff.
	Empathy	If I send a request via email or webboard to the government, I will receive the response that shows the willingness to help from the government staff.
		If I send a request via email or webboard to the government, I will receive the response that shows the friendliness of the government staff.
	Responsiveness	If I send a request via email or webboard to the government, I will receive prompt response from the government staff.
		If I have a problem with e-government Web sites, the government staff will quickly resolve my problem.
	Assurance	The government staff seem to have sufficient knowledge to answer my questions.
		The government staff seem to have an ability to solve my problem.
Continued Use		In the future, I would not hesitate to use e-government Web sites.
		In the future, I will consider e-government Web sites to be my first choice to do business with the government.
		In the future, I intend to increase my use of e-government Web sites.

Table 7. Reliability analysis results

Construct	No. of Items	Mean	SD	Cronbach's Alpha
Information Quality	10	3.067	.604	.899
Accuracy	2	3.394	.748	.839
Timeliness	2	2.715	.795	.783
Relevance	2	3.148	.725	.737
Understandability	2	2.997	.766	.840
Completeness	2	3.079	.741	.688
System Quality	17	3.100	.617	.925
Functionality	4	3.068	.722	.837
Dependability	4	3.019	.699	.749
Ease of Use	4	2.875	.716	.872
Usefulness	5	3.372	.811	.878
Service Quality	10	2.587	.778	.951
Tangibles	2	2.649	.961	.819
Reliability	2	2.582	.899	.937
Empathy	2	2.625	.868	.925
Responsiveness	2	2.406	.870	.905
Assurance	2	2.674	.843	.907
Continued Use	3	3.232	.865	.873

sponding to each of the five selected universities. Sending the invitation emails to the alumni mailing list can guarantee that the participants have experience with the Internet and hence probably enabled us to reach e-citizens. In addition, previous studies indicated that e-government Web sites are particularly popular among those who have at least a college education (Larsen & Rainie, 2002; Wangpipatwong, Chutimaskul, & Papasratorn, 2005). Finally, 1,159 e-mail addresses turned out to be invalid and the invitation emails could not be delivered to the recipients. However, there were 2,441 valid e-mails that did reach the recipients.

Responses to the survey were collected for a two-month period (February 1, 2006 to March 31, 2006). Respondents were screened according to whether they had experience with e-government Web sites. Only those who had previous experience continued with the survey. Out of 799 responses, 614 responses indicated experience with e-government Web sites. All these 614 responses were then used in the analysis after they were verified to be valid and complete without any unusual data or multiple responses. The number of valid responses conforms to finite population sampling formula (Yamane, 1973), along with a 95% confidence level and a 5% precision level.

Data Analysis

A multiple regression was chosen as the appropriate method to examine whether information quality, system quality, and service quality of e-government Web sites will enhance the continued use of e-government Web sites. Together with the analysis, assumptions of multivariate normal distribution, linearity, and homogeneity of variance were tested. There were no violations of these

Table 8. Regression analysis results of information quality, system quality, and service quality on continued use

Construct	Unstandardized Coefficients		Standardized Coefficients	t	p	Collinearity Statistics
	B	Std. Error	Beta			VIF
(Constant)	.293	.144		2.031	.043	
Information Quality	.218	.067	.153	3.251	.001	2.371
System Quality	.547	.073	.390	7.515	.000	2.901
Service Quality	.221	.044	.199	5.072	.000	1.660

R^2 = .434; F = 155.793; p = .000

Table 9. Regression analysis results with the finest model

Dimension	Unstandardized Coefficients		Standardized Coefficients	t	p	Collinearity Statistics
	B	Std. Error	Beta			VIF
(Constant)	.283	.137		2.067	.039	
Usefulness	.441	.041	.413	10.826	.000	1.713
Empathy	.179	.042	.179	4.298	.000	2.043
Accuracy	.123	.043	.106	2.846	.005	1.635
Assurance	.105	.043	.102	2.406	.016	2.111
Relevance	.094	.046	.079	2.066	.039	1.715

R^2 = .483; F = 113.514; p = .000

assumptions. The Variance Inflation Factor (VIF) less than 5 confirms the lack of multicollinearity (Studenmund, 1992). Finally, the number of cases is very well above the minimum requirement of 50+8k for testing the multiple correlation and 104+k for testing individual predictors, where k is the number of independent variables (Green, 1991).

ANALYSIS AND RESULTS

To examine the Web site quality toward enhancement of the continued use of e-government Web sites, information quality, system quality, and service quality were simultaneously regressed on the continued use of e-government Web sites. The results revealed that these three quality aspects significantly accounted for 43.4% of the variance in the continued use of e-government Web sites ($R^2 = .434$, $F = 155.793$, $p < .001$). As shown in Table 8, system quality ($\beta = .390$, $p < .001$) yielded the greatest enhancement on the continued use, followed by service quality ($\beta = .199$, $p < .001$), and information quality ($\beta = .153$, $p < .01$). Therefore, all proposed hypotheses were supported.

Afterward, all fourteen dimensions corresponding to information quality, system quality, and service quality were regressed, using stepwise method to investigate the finest model of enhancement. As shown in Table 7, there were five dimensions, usefulness ($\beta = .413$, $p < .001$), empathy ($\beta = .179$, $p < .001$), accuracy ($\beta = .106$, $p < .01$), assurance ($\beta = .102$, $p < .05$), and relevance ($\beta = .079$, $p < .05$) which formed the finest model of enhancement.

CONCLUSION AND DISCUSSION

The aim of this study was to examine the Web site quality toward enhancement of the continued use of e-government Web sites by citizens. The study was motivated by the lack of empirical studies that uses information quality, system quality, and service quality to examine the continued use in the context of e-government Web sites.

As predicted, the results revealed that Web site quality corresponding to information quality, system quality, and service quality enhanced the continued use of e-government Web sites. The higher the level of information quality, system quality, and service quality, the higher the citizens' intention to continue using e-government Web sites. The results thereby corroborate that information quality, system quality, and service quality enhance not only initial intention as DeLone and McLean (2002) asserted, but also the continued use in the context of e-government Web sites.

Further, the results also revealed that system quality provided the greatest enhancement on the continued use of e-government Web sites, followed by service quality and information quality. This outcome resembles e-business study (Lee & Kozar, 2006) that found online customers considered system quality as the greatest significant factor in selecting the most preferred e-business Web sites.

When considering the dimensional perspective, the results showed that there were five dimensions which formed the finest model of enhancement. These five dimensions were ordered in significance as usefulness (of system), empathy (of service), accuracy (of information), assurance (of service), and relevance (of information). This outcome thereby suggests that government should ensure that these five dimensions are well integrated in the e-government Web sites. The following are some of suggestions.

- **Usefulness:** E-government Web site should provide useful services compared to the traditional way, such as convenience, saving time, and saving cost.

- **Empathy:** Responsible staff should give caring and individualized attention to citizen, such as providing individualized attention to individual concerns and requests, through email communication rather than a generic auto-reply message.
- **Accuracy:** Information on e-government Web site should be correct and reliable.
- **Assurance:** Responsible staff should have the knowledge and ability to inspire trust and confidence. The staff should provide impeccable response to convey trust and confidence to citizens.
- **Relevance:** Information on e-government Web site should be relevant to the site and corresponds to the need.

Furthermore, government should obviously disclosure the usefulness of e-government Web sites. The government may highlight the unique features of the Web sites compared to dealing with government staff for the same services and promote the idea that the Web sites facilitate the access to services anywhere and anytime with saving time and cost.

To conclude, it is a necessity for the government to recognize the quality of e-government Web sites, since it enhances the continued use of the Web sites. Government should ensure that the significant dimensions corresponding to information quality, system quality, and service quality are well established. Finally, the next challenge for government involves changing the citizens' perception and the means in which the information, system, and service are presented and delivered to the citizens corresponding to their needs.

LIMITATIONS AND RESEARCH DIRECTIONS

Although the study provides meaningful implications, it has two limitations. First, the dimensions used to measure information quality, system quality, and service quality are equally weighted. Future research may try using dimensions that are unequally weighted. Second, this study intends to elicit data from e-citizens who are ready for e-government. To regard the digital divide, future research should elicit the data from citizens who have lower level of education, lower income, and also citizens who lack access to the Internet.

REFERENCES

Accenture (2004). *e-government Leadership: High Performance, Maximum Value*. Retrieved May 7, 2005 from: http://www.accenture.com/xdoc/en/industries/government/gove_egov_value.pdf

Ateetanan, P. (2001). *Country Report Thailand*. Retrieved November 14, 2006 from: http://unpan1.un.org/intradoc/groups/public/documents/APCITY/UNPAN012806.pdf

Bailey, J.E., & Pearson, S.W. (1983). Developing a tool for measuring and analyzing computer user satisfaction. *Management Science*, *29*(5), 530−545.

Baroudi, J.J., Olson, M.H., & Ives, B. (1986). An empirical study of the impact of user involvement on system usage and information satisfaction. *Communications of the ACM*, *29*(3), 232−238.

Bevan, N. (1999). Quality in use: Meeting user needs for quality. *Journal of Systems and Software*, *49*(1), 89−96.

Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. *MIS Quarterly*, *25*(3), 351−370.

Bitner, M.J., Booms, B.H., & Mohr, L.A. (1994). Critical service encounters: The employee's viewpoint. *Journal of Marketing*, *58*(4), 95−106.

Chamlertwat, K. (2001). Current status and issues of e-government in Thailand. *15th Asian Forum for the Standardization of Information Technology*, Kathmandu, Nepal.

Cronin, J.J. Jr., & Taylor, S.A. (1992). Measuring service quality: A reexamination and extension. *Journal of Marketing, 56*(3), 55−68.

Cullen, R., & Hernon, P. (2004). *Wired for Well-being Citizens' Response to E-government*. Retrieved March 15, 2005 from: http://www.e-government.govt.nz/docs/vuw-report-200406/

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319−340.

DeLone, W.H., & McLean, E.R. (1992). Information systems success: The quest for the dependent variable. *Information System Research*, *3*(1), 60−95.

DeLone, W.H., & McLean, E.R. (2002). Information systems success revisited. *Proceedings of the 35th Hawaii International Conference on System Science*, *3*(1), 2966−2976.

DeLone, W.H., & McLean, E.R. (2003). The DeLone and McLean model of information systems success: A ten year update. *Journal of Management Information Systems*, *19*, (4), 9−30.

DeLone, W.H., & McLean, E.R. (2004). Measuring eCommerce success: Applying the DeLone and McLean information system success model. *International Journal of Electronic Commerce*, *9*(1), 31−47.

Doll, W.J., & Torkzadeh, G. (1988). The measurement of end-user computing satisfaction. *MIS Quarterly*, *12*(2), 259−274.

Fang, Z. (2002). E-government in digital era: Concept, practice, and development. *International Journal of the Computer, The Internet and Management*, *10*(2), 1−22.

Green, S.B. (1991). How many subjects does it take to do a regression analysis?. *Multivariate Behavioral Research*, *26*(3), 499−510.

Hair, J.F. Jr., Anderson, R.E., Tatham, R.L., & Black, W.C. (1998). *Multivariate Data Analysis* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

Iivari, J. (2005). An empirical test of the DeLone-McLean model of information system success. *The DATA BASE for Advances in Information Systems*, *36*(2), 8−27.

infoDev and CDT (Center for Democracy and Technology) (2002). *The E-government Handbook for Developing Countries*. Retrieved November 14, 2006 from: http://www.cdt.org/egov/handbook/2002-11-14egovhandbook.pdf

Jeong, M., & Lambert, C.U. (2001). Adaptation of an information quality framework to measure customers' behavioral intentions to use lodging web sites. *Hospitality Management*, *20*(2), 129−146.

Larsen, E., & Rainie, L. (2002). *The Rise of the E-Citizen: How People Use Government Agencies' Web Site*. Retrieved April 12, 2006 from: Available: http://www.pewinternet.org/pdfs/PIP_Govt_Website_Rpt.pdf

Lee, Y., & Kozar, K.A. (2006). Investigating the effect of Web site quality on E-Business success: An analytic hierarchy process (AHP) approach. *Decision Support Systems*, *42*(3), 1383−1401.

Limayem, M., Hirt, S.G., & Cheung, C.M.K. (2003). Habit in the context of IS continuance: Theory extension and scale development. *Proceedings of the 11th European Conference on Information Systems (ECIS 2003)*. Retrieved April 12, 2006 from: http://is2.lse.ac.uk/asp/aspecis/20030087.pdf

Molla, A., & Licker, P.S. (2001). E-Commerce systems success: An attempt to extend and respecify the Delone and Maclean model of IS success. *Journal of Electronic Commerce Research*, *2*(4), 131–141.

NECTEC (National Electronics and Computer Technology Center) (2003). *Thailand Information and Communications Technology (ICT) Master Plan (2002-2006).* Retrieved November 6, 2006 from: http://www.nectec.or.th/pld/masterplan/document/ICT_Masterplan_Eng.pdf

NECTEC (National Electronics and Computer Technology Center) (2005a). *Thailand ICT Indicators 2005*. Retrieved November 6, 2006 from: http://www.nectec.or.th/pub/book/ICTIndicators.pdf

NECTEC (National Electronics and Computer Technology Center) (2005b). *Internet User Profile of Thailand 2005*. Retrieved November 6, 2006 from: http://www.nectec.or.th/pld/internetuser/Internet%20User%20Profile%202005.pdf

Negash, S., Ryan, T., & Igbaria, M. (2003). Quality and effectiveness in web-based customer support systems. *Information & Management*, *40*(8), 757–768.

Nunnaly, J.C. (1978). *Psychometric Theory* (2nd ed.). New York, NY: McGraw-Hill.

Parasuraman, A., Zeithaml, V.A., & Berry, L.L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing*, *49*(4), 41–50.

Parasuraman, A., Zeithaml, V.A., & Berry, L.L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, *64*(1), 12–40.

Pitt, L.F, Watson, R.T., & Kavan, C.B. (1995). Service quality: A measure of information systems effectiveness. *MIS Quarterly*, *19*(2), 173–185.

Seddon, P.B., & Kiew, M.-Y. (1994). A partial test and development of the DeLone and McLean's model of IS success. *Proceedings of the International Conference on Information Systems (ICIS 94*), 99–110.

Solomon, M.R., Surprenant, C.F., Czepiel, J.A., & Gutman, E.G. (1985). A role theory perspective on dyadic interactions: the service encounter. *Journal of Marketing*, *49*(1), 99–111.

Spreng, R.A., MacKenzie, S.B., & Olshavsky, R.W. (1996). A reexamination of the determinants of consumer satisfaction. *Journal of Marketing*, *60*(3), 15–32.

Studenmund, A.H. (1992). *Using Econometrics: A Practical Guide* (2nd ed.). New York, NY: HarperCollins.

Szymanski, D.M., & Hise, R.T. (2000). E-Satisfaction: An initial examination. *Journal of Retailing*, *76*(3), 309–322.

Truehits (2006). *Truehits 2005 Awards*. Retrieved November 16, 2006 from: http://truehits.net/awards2005

United Nations (2003). *UN Global E-government Survey 2003*. Retrieved November 16, 2006 from: http://unpan1.un.org/intradoc/groups/public/documents/un/unpan019207.pdf

United Nations (2004). *UN Global E-government Readiness Report 2004: Towards Access for Opportunity*. Retrieved November 16, 2006 from: http://unpan1.un.org/intradoc/groups/public/documents/un/unpan019207.pdf

United Nations (2005). *Global E-government Readiness Report 2005: From E-government to E-Inclusion*. Retrieved November 16, 2006 from: http://unpan1.un.org/intradoc/groups/public/documents/un/unpan021888.pdf

Wang, R.Y., & Strong, D.M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–34.

Wangpipatwong, S., Chutimaskul, W., & Papasratorn, B. (2005). Factors influencing the use of eGovernment Web sites: Information quality and system quality approach. *International Journal of the Computer, the Internet and Management*, *13*(SP3), 14.1–14.7.

Wilkin, C., & Castleman, T. (2003). Development of an instrument to evaluate the quality of delivered information systems. *Proceedings of the 36th Hawaii International Conference on System Sciences.* Retrieved December 11, 2007 from: http://csdl2.computer.org/comp/proceedings/hicss/2003/1874/08/187480244b.pdf

Yamane, T. (1973). *Statistics: An Introductory Analysis* (3rd ed.). New York, NY: Harper & Row.

Zeithaml, V.A., Berry, L.L., & Parasuraman, A. (1996). The behavioral consequences of service quality. *Journal of Marketing, 60*(2), 31–36.

This work was previously published in International Journal of Electronic Government Research, Vol. 5, Issue 1 , edited by V. Weerakkody, pp. 1-28, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 5.9
Social Aspects of Mobile Technologies on Web Tourism Trend

Fernando Ferri
IRPPS-CNR, Rome, Italy

Patrizia Grifoni
IRPPS-CNR, Rome, Italy

Tiziana Guzzo
IRPPS-CNR, Rome, Italy

ABSTRACT

This chapter analyzes how the development and use of mobile and Web technologies are changing the way to search information, to plan, to buy, and to travel. The new technologies are changing several aspects of our life, such as the way in which people work, buy, learn, travel, and how they relate to each other, and so on. The tourist sector certainly represents one of the most dynamic markets, able to capture innovations and opportunities provided by the Web, in such a way that gets to be an out-and-out model of e-business. Internet access now is not restricted to personal computer. In fact the use of mobile devices is becoming increasingly important. The chapter's goal is to analyze social implications of Web applications and mobile devices and how they are improving the attitude of the customers both the fruition of tourism services and to development of sustainable tourism.

INTRODUCTION

The widespread use of Internet and Web technology in every aspect of our daily life has brought great change in the consumers habits in any field but mainly in the tourism sector. Every year, million of tourists approach to Internet in order to find tourist information: vacations, flights, guides, last minute, cruises, destinations and routes. This situation is changing the concept of tourism. In particular, tourism was defined by the World Tourism Organisation as "the activities of persons travelling to and staying in places outside their usual environment for not more than one consecutive year for leisure, holidays, business, health treatment, religion and other purposes".

DOI: 10.4018/978-1-60566-156-8.ch027

A new concept of "intelligent tourism" is spreading, based on all those innovative technological solutions offered by Web that allows achieving information on cultural, artistic and other kinds of natural interest.

The Web imposes itself more and more as relevant reference and indispensable resource in the tourism sector both for customers and tourism companies, thanks to undisputed advantages such as:

- Speed for information exchange,
- Improvement of interaction among people located in different places,
- Improvement of information sharing, knowledge and services availability for all the users.

These above three aspects are very important because they are producing the markets globalisation and the spatial and of temporal boundaries break down. Moreover the information and knowledge production and sharing improves both quality development of social inclusion. Moreover, the success of the Web and mobile technologies in the tourism sector is given by very competitive prices, but also by the ability to differentiate the offer, by the improvement of possibility to reach the market niches and by proposing new services with a good usability degree. The user is not a passive subject and when s/he visits a business site s/he knows that s/he will not be able to find negative aspects of a choice. This has brought to the creation of virtual communities where users share their own travel experiences with other tourists and where the potential tourist consumer finds relevant information. The tourist seems in fact to prefer descriptions and testimonies of other people that have already visited that place.

In the next sections we introduce the passage from the old economy to the new economy in tourism sector, in particular we describe how the advent of e-commerce has marked the passage from traditional travel agencies to Internet. In succession we describe the different tools tourist used for Web and mobile and their social aspects. Finally we describe the new scenarios of tourism using the new technologies and how the mobile devices can develop the sustainable tourism, increasing both tourism demand and tourism supply.

FROM TRAVEL AGENCIES TO ONLINE TOURISM E-COMMERCE

The tourism initially involved an elite activity. In the last years it has beenbecoming a mass phenomenon shared by million of people all over the world, and it has been becoming one of most relevant economic sector of most countries. Data provided by World Tourism Council (1997), in fact talk of seven hundred million of arrivals of tourists all over the world. The tourism represents the 7% of the total occupied people and the 2% of global gross domestic product. It is a dynamic phenomenon, mutable and complex, it can be defined as a "social fact", it changes with transition of tendencies, of orientations, of necessities and needs of society.

In the past tourists had to go to the travel agencies, i.e. in the physicals commercial places where to plan and buy a travel. These agencies execute activities of reservation and selling of single tourism services or services packages confectioned by Tour Operators. This means that people, have to go to a physical place to use such services. The choice of the agency can be limited by the physical distance. Moreover services obtained depended by operators, by their personal skills and by their limited information.

When tourists visit different locations guidebook can be very useful. The paper guidebooks more frequency used by tourist in the old economy. Even if they are still now the principal tool used by travellers, because it is easy to consult and information is well structured, they have a lot of limits. Information in tourist paper guidebooks can be outdated because items written many years before could not be updated, hotels and other tourist activities could be ceased (Schwabe, 2005). Actually the new Web and mobile technologies can provide more timely and complete information than paper guidebooks. In fact user can obtain an updated answer to her/his question, more than a paper guidebook and improve the information quality and consequently the travel quality.

Thanks to technology innovation of the last thirty years, not only users have obtained several advantages but also the tourist companies. The most innovative ones in fact, can actually be able to redefine their own organization structure and relationship with partners, optimising the operating costs and improving the quality of services.

Information and Communication Technologies (ICT) have allowed tourism companies to increase their efficiency and their market value. In fact ICT offer the chance to share data-bases with other organizations and other customers' information resources and services. Besides ICT allow to optimize other internal functions, either lowering costs and by expanding services to offer (Poon, 1993).

In latest three decades the tourism sector has been characterized by three technological phases: Computer Reservation System in the 70s, Global Distribution System in the 80s and Internet Revolution since the second half of 90s. (Buhalis, 1998). The first two have allowed to create, to develop and to globalize availability of services by travel agencies, who have exclusive access to automatic booking systems. The last phase has allowed the customer to perform bookings by themselves, redefining the entire business tourism system, modifying the same tourism fruition and improving the tourist experience. (Stipanuk, 1993).

The first change registered in tourism market according to these technological evolutions is a great increase of e-commerce. On-line tourism is one of the most meaningful achievement cases of the e-commerce in the world. The tourism products in fact have ideal characteristics for e-commerce, they can be represented in Web site utilizing potentiality of multimedia and hypertextual communication. Some studious, such as Werthner and Klein, assert in fact, that tourism is considered as one of more important field of application in the World Wide Web.

Some statistic data can help to understand the Internet impact on tourism sector. According to research of Eyefor Travel Research (2007), the tourism e-commerce field represents about the 30% of Web purchases. In 2005 the on-line booking represented the 33,6% of the worldwide tourism market, in 2006 the market quote of travel agencies is of 36,6%, it goes over with the 37,5% by on-line market. However the United States remains on the top of e-commerce for tourism services creating a gap with other countries.

In Figure 1 it is possible to see how the online market is very important and relevant respect to offline market. In Europe the off-line canal is traditionally the one preferred by consumers but it is loosing more and more market quote to vantage of on-line tourism, the online selling since 2002 to 2006 are increased constantly of 43% per annum. Data in the next table is showing the Internet relevance to improve relation between demand and supply operating on promotion and vending of tourism services to consumers. The trend of on-line tourism market (Figure 2), in fact, shows an yearly increase on-going ascent of 30% in 2006 respect to 2005. Among reasons of this increase there are different factors; first of all the raise of Internet utilise, then the coming of low cost companies and last the expansion of large band-width and of electronic credit card.

Analysing the European countries scenario of on-line travel market, United Kingdom is the best one with 34% thanks mainly to large presence of low-cost flight. In the second position there is Germany with 20% of on-line travel market. On the bottom of it there is the Southern Europe, the reason of this gap is also due to socio-cultural factors. The UK consumers in fact, are used to buy everything on the Web, thanks to wide presence of large band-width and frequent usage of

Figure 1. Value of tourism European market: Comparison online offline (Reference: Mele, 2007)

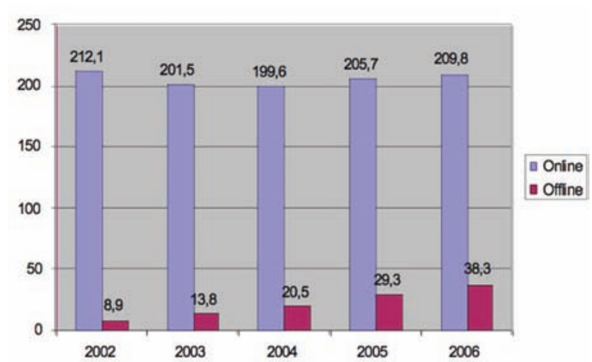


Figure 2. Trend in European online travel market (Reference: Mele, 2007)

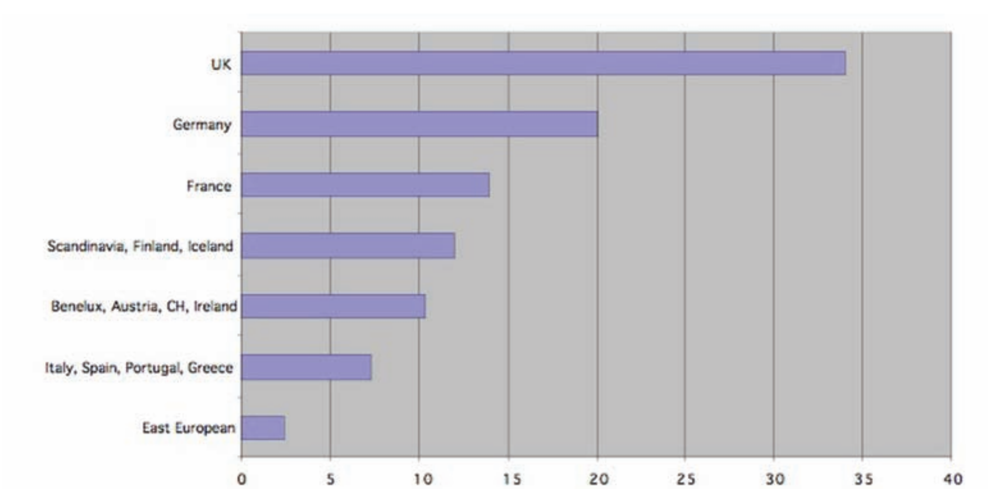
YEAR	TOURISM MARKET (SELLING)	ONLINE SELLING %	ONLINE % SELLING INCREASE
2002	221.036	4,04	77
2003	215.303	6,4	54
2004	220.094	9,3	48
2005	235.017	12,4	43
2006	247.16	15,5	30
2007	253.506	18,4	22
2008	259.723	21,1	17

electronic credit card. In Southern Europe there is distrust to buy in Internet especially by people more traditionalists. There is a cultural distaste to use credit card because they fear to be deceived.

Moreover, there is a short access to the large band-width. It's noticeable that the Web diffusion depends on structural and cultural factors linked to countries' policies and economies. In Figure 3 is represented the European scenario of on-line travel market 2006.

Regarding the service typologies the purchases more frequent on the Web in the European online market are occupied by: Air travel 56%, following booking hotel and travel packages 16%, train 8%, car renting 3% and finally car ferry 1%. Increasing also others segments as holiday houses and cruises. Another data about European scenario is about selling channel: in 2006 the direct sellers

Figure 3. Geographic status of the European online travel market 2006 (Reference: Mele, 2007)



accounted for 69% of online sale and intermediaries the 31%. This is way low-coast companies sells flight tickets directly to consumers by their sites, avoiding intermediaries actions.

The new trend of tourism is also influenced by duration and kind of holiday. In particular short trips, especially in the weekend, are growing. The Online Travel Agencies (OTA) have understood (exploited) this tendency and created the Dynamic Packaging, that allow travellers to organize the travel about their needs.

As reported by travel weekly (www.travelweekly.co.uk) "Dynamic Packaging is the practice of selling holiday components separately rather than in a single package." OTA have to create the Dynamic Packaging to compete with the prices and flexibility offered by online retailers. In general, we can affirm that tourist behaviour shows the wish to adapt vacation to his/her own needs, adopting consequently travel solutions.

The business travel are the most purchased on Internet due to several advantages: convenience, best prices, speed and availability of tools to search travel and services, possibility to change default packages for personal needs, possibility to choice the hotel franchising, the renting society, possibility to read on one page all options about flight and so on.

After to have analysed the transition from offline to online market tourism, in the next section we will describe the new online tools for Web and mobile that improve the quality of tourism services fruition.

NEW TOOLS FOR WEB AND MOBILE INTERACTION

New technologies in hand (at disposal) of tourists have deeply enhanced the tourism information quality. We know that Internet usage is still increasing dramatically, mainly through personal computers and mobile devices. So far, much of this growth has come from new Internet users. Another aspect to consider is that in the short to medium future, users will be accessing the Internet more frequently using a variety of other mobile devices for different purposes and some of these will be quite surprising.

Users can use advanced tools such as wikis, blog, personal spaces, mush-up and their own passage to mobile devices to organise their travels and relate to each others.

Wiki is a Web-based application that allows users to add content and also to edit content supporting collaborative writing, opening discussions, interaction and Web-authoring (Desitlets 2005). At the same time it also allows to delete every content of a new page or another already written by others, to restore an old version, to avoid risks of damages by hackers or non-collaborative members. Thanks to wiki everybody can write everything about a place, a city or a country, and can read everything, or improve what is already written.

Wiki is the most important example of collaborative online community, and applied to a tourism site give to every traveller the chance to share their own experience and to collaborate with other members, activeness and loyalty to the site is guaranteed. The Web site www.wikitravel.org represents an example of a collaborative and updated online global travel guide.

Another kind of Web application important for tourism is represented by personal pages (that can be part of a social network). Technically they are more simple than wikis, in fact they only represent a page created and modified by only its owner. But their technical ease allows users to concentrate on their own contents, and besides they can also build their own buddy network by adding other user's personal pages to their "friend list". Every page becomes in this way an host joint to a bigger network, and every user can discover new paths by starting from the page of a friend. A personal space usually includes blogs: diaries of text or pictures in chronological order (such as Blogspot), personal pages where the contents

Figure 4. Different types of usage of wikis and wikipedia (Reference: White, 2007)

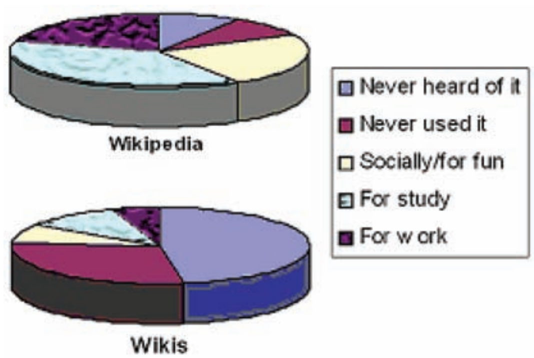
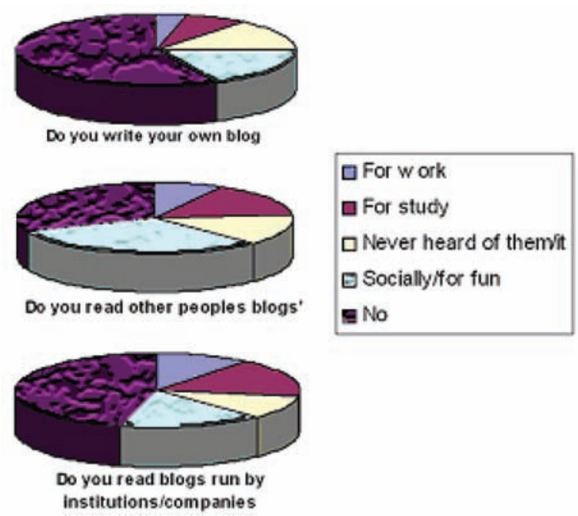


Figure 5. Different types of usage of blog (Reference: White, 2007)



are more static and includes every kind of multimedia items (such as MySpace) or every other thematic social content (list of Web links, list of news, musical tastes, and so on).

It is easy to understand that Internet promotes the making of thematic sub-network such as travel blogs. In the blog every author talks about her/his own favourite topics, and it's easy to quote other's news or opinions in this way; for example, a list of travel blogs links each other to compose a thematic travel network. In fact, new Internet users do not utilize only one tool at time. They use a wide set of Web tools: they write their daily thoughts on a blog, post their best pictures on a photo blog, collect their favourite Web sites on social bookmarks, and so on.. Every kind of new Web application has one special feature that makes the difference among the infinite offer of Internet; however there is the necessity to do not neglect the role of the "pass the word" in the choices of potential tourists. After the vacation, contacts are kept by email and messenger, different opinions are compared on forum, suggestions before leaving are read online, the best offers and discounts are accessible everywhere and every time. In fact, the advanced tools do not swap tools such as emails or forums.

It is important to have an idea about how Internet users relate to wikis and blogs. (Fig. 4-5). About wikis, David White (2007) reports in his research, that most users not know what they are, but when talking about Wikipedia (the famous free encyclopaedia created and continually updated by its users) he reports that nearly the 80% have used it for study, work or simply for fun (see fig. 3). Instead about blogs, users have been asked about their own blogs, other blogs and institutional blogs. What emerged is that only the 20% write their own blog and nearly the 75% read other's blog to work, to study but the most of them is for fun; but if we talk about blogs run by institutions or company the percentage decrease to 45%.

We can easily see that blogs are more used than wikis for fun because they are easier to use and similar to personal diaries then more appropriated to socialise, they are more used by young target under 18 then from 18 to 34 years, less by adult public. Instead Wiky are more complex and reference target is heterogeneous, they are in fact more used for study and for work.

The question is: what does make new Web tools really powerful? This question find its answer in the fact that one of the main goals of Web operator is actually of giving to every content a real meaning; this is the goal of the semantic Web (also called Web 2.0). The simple text of any content over the Internet can always be misconceived by search engine and other Web applications; the goal of semantic Web is giving every object (a picture, a paper, a clip and also papers) a real meaning, and this is possible by linking one or more keywords to that particular object. In this way it is possible for every user to find e.g. pictures of a "puma" without obtaining sport wear brand "Puma" images. The same text, can now have two or more meanings! That's the reason of users' vitality for social networks and new collaborative Web tools, which represent the first step to give texts, pictures, video clip or every other object a real meaning. Moreover, tagging is a new emerging way to categorise, share and search information based on meaning of keywords that facilitate tourists in their choices.

Once contents of every kind are meaningful, becomes important accessing these information and the Web tool that helps user to do that is: RSS (Really Simple Syndication). RSS is a new file format used to communicate information by following a syntax. It is used to spread news, new blog posts and new multimedia object, and it is read by software (Web based or not and also called "aggregator"). Thanks to RSS provided by almost every kind of site, users can read news about their favourite site without visiting every single site. Users can also rearrange their own subscribed feeds to obtain thematic news, by adding labels or tags, the same kind of keywords we talked about Web 2.0. It is easy to understand that the use of RSS allows everybody to improve their own information needs, besides it represents a tool available by every kind of technological device, from a desktop computer or even a small cell phone or a palmtop.

This technology has several positive impacts also for tourism supply. Many companies have adopted RSS feeds in their Web sites to keep a communication with their customers and enhance their search engines optimisation. It is difficult keep update offers and purposes because they are short life. This technology is characterized as a demand-pull rather than a supply push model. (Sigala, 2007). Furthermore, information flows about travels, special offers or more interesting places are real time provided on computer's monitor by a special software (Web-based or not) called "aggregator".

We have already talked about sites that offer the chance to upload pictures, tagging them with keywords (giving these pictures a real meaning), organizing them by creating albums or slideshows; but now the new scenario is about people, or travellers, who can take pictures with digital camera and/or smart phones and immediately send to these to online services. For travellers this means they can prove their skills and their travel experience; for all the other Internet users means to have the chance to find pictures from all over the world, especially when some special happening occurs. This new trend allows old media such as newspaper and television news to report flash news by showing images or video clip, taken from (respectively from Flickr or YouTube).

Another very important field on mobile devices applications for social activities (and consequently for tourists) are organizers and planning sites. They allow to set a calendar of events (to be shared with other users or customers) that can be subscribed, edited and shared. For example Google Calendar allows a hypothetic travel or service agency, to set a special calendar of event about new destinations, prices updates or happy-hour promotion. The users and customers can subscribe this particular calendar, and they can even receive on their own cellular phones SMS reminders that advice about expiring offers or new item added in that particular moment. These mobile technologies make available tools to plan the complete travel and the localization of the places of interest. Among these technologies we find the Internet mapping: the digital interactive maps that supply information on the hotel proximity, restaurants, stores, services, monuments, situated historians, archaeological sites. It is also possible to know the territory classified in different topics (sport, well-being, wine and food, information on traffic, weather forecast). The customer can generate personalized itineraries, search useful services and visualize more specific information on places of interest.

Newer technologies also allow mixing different applications (usually Web-based applications) to each other, to obtain a completely new one with new different functions. These are called "mash-ups", and they represent the newest scenarios in new Web technologies. The most famous mashups are those regarding Google; the Mountain View company in fact, has built a global system of satellite maps covering the whole world, and thanks to GPS devices, users can add new information such as pictures, short videos or information of any kind, exactly relating to a place on the map. In this way tourists can share information of any kind, in a very simple and funny way. Tourists can create their own path by describing with words, by drawing the route on a digital map, by adding digital pictures relating to a special point on the map. So they can create thematic itinerary and share for other tourist, just as every other digital tool we previously described.

Almost every Web 2.0 site involves the chance to interact with the site itself and other users too. The whole Web based application (and so, also online tourism market) is going towards Web 2.0 and so called Mobile2.0, this means that collaboration among users and participation is fundamental for the content site. The future of the e-commerce plays a very important role in the field of the collaboration and sharing: friendships, fellow traveller, socialization are carrying elements of each kind of travel. The main use of mobile devices by tourists is photo and video sharing whit others by personal blog or site.

Portable mp3 devices also give another chance for the tourists. In fact, they allow tourists to bring with them, in a very low weight, a large amount of video or audio files, such as thematic guides. Some operators have also started to provide their customers, with mp3 devices already filled by guides about the subject of the travel (a museum tour, a walk in the historical centre of a town, and so on).

The main reason that allows to tourism to benefit from the use of mobile technologies is the new services to travellers on the move. An example of this technology is the location – based Services.

The term *location-based services* (LBS) refers to information services that are accessible through a mobile handset and based on the current geographic location of the mobile device (Antikainen, H., 2006). The most commonly used is the satellite-based Global Positioning System (GPS). The conventional application areas of LBS include mapping, tracking, routing and logistic, electronic yellow pages, data collection and public safety (Beaulieu & Cooper 2001, Maguire 2001, Veijalainen et al. 2001, Zipf & Malaka 2001).

The primary functions of LBS for tourism are usually regarded as being the localization of persons, objects, and places, search of restaurants, shops, hotels, or points of interest in proximity and information about traveling conditions. Currently, mobile services facilitate the reservation of last-minute trips, rental cars, and hotels; and they provide information about changes and delays of flights and trains, offer guides on restaurants, events, and sightseeing opportunities at the destination (Berger *et al.* 2003, Eriksson 2002).

In the last years mobile devices such as mobile phones with embedded camera, palmtop, notebook and last but not the least GPS systems, have enhanced the use and the production of personal sites and blogs. In fact they allow everybody to post, not only reviews of a new bed & breakfast but also the pictures of the rooms or a short movie showing the landscape.

This brings to birth of the mobile virtual communities. In the next section will be analyse the social aspect of virtual communities.

SOCIAL ASPECTS OF VIRTUAL TRAVEL COMMUNITY

All these new tools that we have described are changing the way of people to interact and to communicate among them. In fact, the users can use these tools and meet new people in a virtual community that is a virtual place where people can speak (textual chat), can meet (video chat), can discuss about different matters (newsgroup and forum), can play and can exhibit themselves (personal home page and free Web).

The first sociological definition of "virtual community" was given from Rheingold in 1993 where he defined the virtual communities as: "… social aggregation that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling, to form Webs of personal relationship in cyberspace. A virtual community is a group of people who may or may not meet one another face to face, and who exchange words and ideas through the mediation of computer bulletin boards and networks".

The first element of a virtual community is the absence of territorial boundaries due to its missing physical dimension; it changes for anyone the perspective to interact with other people according to their own needs and interests enhancing the real interaction and communication possibility. Whatever reason motivates a user to join network, sooner or later, s/he will need or curiosity to interact with other.

The advantage of Web is that it encourages humans to establish "weak" relationship with unknown people; this enables the communication also among persons that have different social characteristics. Moreover the on-line communication usually is uninhibited favouring sincerity in the discussion. The impact in the real social life is not decried but reinforced. The networks represent an aggregation form similar to society in which we live: weak and strong relationship, need comparison and exchange.

This form of social aggregation is not ground on politic, ethnic, linguistic, religious affiliation but volunteer cooperation between individuals that share the same interests, hobbies and goals.

It is easy to understand that Web community has become a term who involves any group of people who communicate online. These people can share different goals, interests or hedonistic pleasures. The term "online community" is also used to mean community network. One of the fields where all these concepts are successfully applied is the online travel community. This is a virtual community where backpackers, globetrotters, and other adventurers from all over the world to join together at different online platforms to exchange information, experiences, and plans in their favorite pursuit travel. In fact, in the travel and tourist industry, Internet encouraged more and more people to join into virtual communities to satisfy their needs, to fulfil their asking tips and suggestion before having a "real" travel.

Recently also travel organizations have realized the power of the new technologies for the core of their activity and the importance of virtual travel communities for their own marketing actions, by broadening their borders.

The travel and tourism virtual communities represent an ideal place without space and time, where people can meet experiences and different worlds. The travel is in its own nature delocalised in respect to point in which oneself is; for this reason it needs a strong communications and information exchange.

The travel experience is rich of emotional and relational contents that for this nature can be shared in the community. A person accesses a virtual community for different reasons: to search information and services, to contact different kind persons, to find partners to share experiences, amuse (oneself). All that is very compatible with tourism that is an "experience reality" and that needs of aggregation places.

Wang, Yu and Fesenmaier (2001) study, analysed needs of online tourists communities related to tourism organization marketing.

They have identified three main classes of needs: functional, social and psychological one. Functional needs include: transaction, information, entertainment, convenience and value. Social ones include: relationship, interactivity, trust, communication and escape among humans. Finally, psychological needs are include: identification, engagement, and sense of belonging, relatedness and creativity. In their work they pointed that since tourism is traditionally studied referring to geography location and space, it is noticeable that tourism-market organizations lack skill in how an online community can be used as a marketing tool. In fact, we cannot forget technological evolution about Internet since the last 15 years. They also predict growing of community concept as the Internet becomes more and more widespread with the new global economy. The network technology has allowed people to be more connected to each other.

People can obtain a lot benefits by joining themselves to the community depending on the different nature of communities and the various characteristics of community members (i.e. many people want to make efficient business transactions and interact with others people; many other rather want to have fun, meet fellow traveller and to express their own opinions; many other still want to develop a sense of belonging, to express their cultural and economic interests and establish relationships).

In 2004 Wang and Fesenmaier expanded their theory with a further study on modelling participation in an online travel community. In particular, were examined the relationship between members' needs and their level of participation in a virtual travel community. In their work they added to the user's needs identified in their previous papers, a hedonic need (including entertainment, enjoyment, amusement and fun). According to them, members participate in on line travel communities to satisfy four fundamental needs: functional needs, social needs, psychological needs and hedonic needs. In this latter work the authors also analyse the role of demographic differences in the behaving of tourism online consumer. It was hypothesized that users' needs in a VTC are not constant but can change with demographic characteristics such as gender, age, education and so on. For example female members usually attach more relevance to hedonic needs, while male members are more significant to membership duration. Other important results of their analysis are about differences according to different ages. For example they observed that groups who are aged 56 or more, versus young members are less attached to functional needs. The aged 20 or younger attribute greater importance than the older groups to social and psychological needs. Differences are also found between members with different education level and their respective needs.

All virtual travel communities have some common features. Communities mostly provide a warm, trusting, and supportive atmosphere. When members share information, they do it with great care and responsibility. They rely on each other more than they do on outdated travel guidebooks or on second-hand and static information from conventional travel literature. They also have to attract a lot of members and give them benefits and satisfactions to be successful. This aspect is very important for tourists because they need to solve a wide range of problems, starting from the pre-visit, the post-visit and of course the travel itself. Before the new technological and digital tools, the traveller could only trust a travel agency operator, but now s/he can find help to choose her/his destination, to solve the most common problems during the planning of the travel (medical suggestions or documents needed), and obviously buying the cheapest flight rate; every question asked by the traveller finds an answer from a user already experienced in that way. But not only, virtual communities are also widely used by tourists to enhance their journey;

in fact as seen before, mobile devices allow users to access the infinite Web without sitting in front of the computer at their home, and heavy guidebooks are now substituted by light mp3 reader or other mobile devices. Finally, tourism virtual communities are important after the travel itself, in fact if a traveller has made use of others' resources, it is now the time to contribute with his own by uploading pictures, writing a review or an itinerary, answering to other's questions, or by simply writing his personal blog to share with friends, but sooner or later someone will find his writing and they will find it useful.

Every user approach a community to search something for himself. This usually is the first step in a lot of other cases, and it is important because the user identifies himself with the community; after this first step her/his loyalty is guaranteed, and it is very easy that s/he will become also a content creator contributing to the growth of the community.

NEW SCENARIOS ABOUT TOURISM AND NEW TECHNOLOGIES

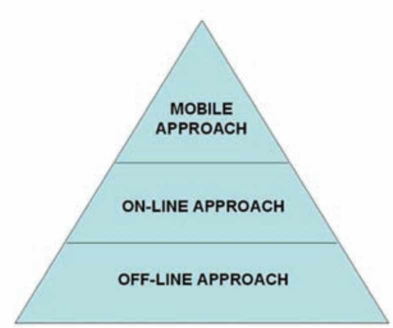
In the early section we analyse the technological and cultural changes trigged by Internet in the tourism fruition and the motivations that encourage the users to belong to a virtual community. All the different modalities used by tourists to plan, buy and do their own vacations, can be classified in off-line, on-line and mobile approach; we can represent them in a pyramid (fig. 5).

The bottom of pyramid is occupied by off line approach; they firstly appeared in chronological order and are the most used by people because they are largely widespread and accessible to the large part of the tourists. Moving upward we find the online technologies more powerful but less widespread and less accessible to users. On the top of the pyramid there are very advanced technological tools (such as mobile devices) but fewer diffused than the other, in fact the pyramid tightens.

It's easy to imagine that in the future this pyramid will probably completely reversed. The traditional paper guidebooks will probably be less utilized and available to tourists while online tools on advanced mobile devices will be largely accessible and commonly used.

The objects in the pyramid of Figure 6 are sorted to help us to simply keep in mind evolution of tourism tools. The basis of the pyramid contains "old economy tools" such as paper guidebook and travel agency. Indeed, they are the most common tools used by travellers, and are concrete objects based on an economy made of selling of products (the paper book itself) or services (travel agency). The next step is represented by the rising of the Web technology; in fact Internet has deeply changed everything, and of course economy too. Everything was concrete in the old economy has now become immaterial and "made of bits". People have Web pages to browse using a computer instead of such a book to read in several contexts. They contain forums, blogs, personal pages and wikis. Then, over the Web era, the top of the pyramid contains technological mobile devices. They are obviously the most advanced tools among the all previously described, and we also have to point that they have to be used jointly with the previous one. In fact, all Web applications will use mobile device so users will have access to information from anyplace and anytime.

Figure 6. Tourism technology evolution



In this way, what we have hypothesized about the future of this evolution (the previously described rotation bottom up) is already in progress, because mobile devices of every kind are used much more than other Web tools. This is causing tourism to benefit from use of mobile technologies about new services to travellers on the move.

Every new technology that we described, can be used to enhance tourist's experience, and they can also be a great chance for local administrations or local tourist promotion organizations to improve their tourism appeal and promotion between appropriate operators.

Mobile technologies give great opportunity to increase the value of territory and to develop sustainable tourism. The virtual community is in fact one of the most effective business models and provides great opportunities for both tourism organizations and customers. (Armstrong & Hagel, 1996).

Users more and more approach to mobile virtual communities to search new and unique contents, uncontaminated places, not commercials and far from mass-tourism. The new tools can allow promoting a sustainable tourism with respect of people and places.

People are searching more and more a high quality environment, but environmental resources on which tourism is based are limited. That's why it is very important to invest in a sustainable tourism. This kind of tourism is defined by the World Tourism Organisation as "tourism which meets the needs of the present tourists and host regions while protecting and enhancing opportunities for the future." A fundamental characteristic of sustainable tourism is that it creates safeguard and respect for the environment and local traditional culture. Moreover it recognizes the centrality of hospitality.

Tourism is a worldwide phenomenon that is very important for the socio-economic development of a lot of countries. It can contribute to the progress of a country but there is the risk it causes environmental degradation and loss of local cultural heritage.

Through digital mobile communication, the natural and cultural heritage that characterize the geographical area, could be communicated in an integrated way to tourist to guide him/her, also by geo-referenced information, toward their own fruition and knowledge with the other purpose, to contribute, through tourism, to the developing of minor tourist centres and their own neighbouring zones. The tourist communication can be made using a Personal Digital Assistant (PDA) or a Smart Phone and thanks to GPS system, it is so possible not only to localize the tourist on the territory, but also the cultural heritage of the location, by sending to users, on their wireless devices, the geo-referenced information with geographic route to reach that particular cultural heritage. The information sent to tourists can be personalized, integrated, complete, clear and multimedia: it can be communicated by text, but also map, video clip, 3D images and audio files.

Social innovations consist of tourists that can reach also places that are not necessarily promoted by tourist book guides or catalogues and they can be reached by messages concerning particular events at the particular time in that place.

Local administrations will have the opportunity, to promote and improve their own territory, to reinforce the sense of belonging and share memories and experiences. The promotion of these cultural aspects of a country will give the chance to have a positive impact on local economies and particularly on the tourism sector.

Promotion of a territory can produce added value to the economy of a place. To promote the local typical products, the local artisan products, to characterize thematic itineraries, fairs, that distinguish a country, is very important to improve the productivity, to create new job and new opportunities and to stimulate the development and innovation. The promotion of a territory is an important moment in the economic development of a community.

Thanks to the use of virtual communities, moreover, local people can enter in contact with people from all over the world and can attract tourists interested in culture and nature of their country.

The use of mobile technologies for sustainable tourism will remove the risk of cultural marginality or isolation because they represent an opportunity for cultural exchange and a possibility to integrate local knowledge into social, economic and cultural development. The importance of natural and cultural safeguarding is confirmed by increased demand of tourism, that joints attention for nature with the interest for intangible culture.

The intangible culture refers to a set of non-physical characteristics, practices, representations, expressions and skills that characterizes cultures, people and places.

The protection and safeguarding of cultural tradition: social practices, typical products, performing arts, rituals and festive events, language, knowledge and practices concerning landscape, has played an important role in the cultural politics and programmes at all levels (local, regional, national, European and international) in recent years. Furthermore people's interest to knowledge and their research of tourism that improves these aspects will remove the risk that a large number of cultural traditions can be lost.

Mobile technologies in the tourism sector represent an important opportunity to improve the vitality of a community. Its bring economic, environmental, image benefits. For example it can stimulate performing of traditional events and festivals that otherwise could be lost. Moreover mobile technologies can stimulate the development of the tourism in marginal regions and can reduce emigration from local areas. They can improve job and earning perspectives of the local population and improve the quality of the tourism activities and the related skills. Moreover they can improve the quality of life of the local population due to the creation of facilities and services, upgraded infrastructure, health and transport improvement, restaurants, food, and so on.

CONCLUSION

In this chapter we explained how mobile technologies have modified tourism sector and how they became important for the tourists themselves and for the global economy.

We have provided a description of the main social implications of the Web technologies earlier, and mobile later. We have analysed how users approach tourism Web applications; they firstly use (for example) Internet to buy low cost flight or to plan trips, later they start writing their blog, sharing their pictures or reviewing their trips, and in this way they start contributing to build a virtual community that will be used by other users for advices to buy their tickets, to plan or share their trip or just for fun. That's why all this, is now called social network.

The next step we have analysed is the coming of mobile devices that have revolutionized the way that tourists enjoyed their experiences. In fact, it is obvious that Internet, as described above, helps tourists before and after their trips, but now, thanks to mobile devices, the high potential of Internet is brought straight to their hands to facilitate the fruition of tourism. Mobile technologies can improve accessibility, information and service provisioning and safety for both tourists and tourism resorts.

Then we have hypothesized a pyramid representing on the bottom the old economy of a travel agency or a paper guidebook, in the centre all the new economy tools belonging to the Internet world, and on the top the mobile devices. Anyway analysing the future scenario we have also described the rotation of the pyramid because in the near future mobile devices will be more accessed and utilised than old economy tools, and what was rarely diffused in the past, will be commonly used.

Finally we focused on the issue of sustainable tourism a new way to travel respecting environment and traditions, by merging information sharing and new technological devices.

Thanks mobile devices several small places and particular events far from commercial routes can be promoted by local administrations and discovered by tourist, contributing both sustainable tourism and development of small place's economy.

REFERENCES

Antikainen, H., Rusanen, J., Vartiainen, S., Myllyaho, M., Karvonen, J., & Oivo, M. (2006). Location-based Services as a Tool for Developing Tourism in Marginal Regions. *Nordia Geographical Publications*, *35*(2), 39–50.

Armstrong, A., & Hagel, J. (1996, May). June). The Real Value of On-line Communities. *Harvard Business Review*, *74*(3), 134–141.

Beaulieu, M., & Cooper, M. (2001). Wireless Internet Applications and Architecture. *Addison-Wesley*.

Berger, S., Lehmann, H., & Lehner, F. (2003). Location-based Services in the Tourist Industry. *Information Technology & Tourism*, *5*(4), 243–256. doi:10.3727/109830503108751171

Buhalis, D. (1998). Strategic Use of Information Technologies in the Tourist Industry. *Tourism Management*, *19*(5). doi:10.1016/S0261-5177(98)00038-7

Desilets, A., Paquet, S., & Vinson, N. (2005). Are wikis usable? *WikiSym Conference*, Oct 16-18, San Diego.

Eriksson, O. (2002). Location Based Destination Information for the Mobile Tourist. In Wöber K.W., Frew A. J., Hitz M. (eds.). *Information and Communication Technologies in Tourism*. Springer-Verlag.

Eyefor Travel Research. (2007). European Online Travel Market Report 2007.

Maguire, D. (2001). Mobile Geographic Services Come of Age: ESRI Dives into Wireless Markets. *GeoInformatics*, *4*(March), 6–9.

Mele, M. (2007). Mercato turistico on line: è boom anche in Europa. http://www.ghnet.it/Article396.html

Poon, A. (1993). Tourism, Technology and Competitive Strategies, UK. *CAB International.* Wallingford, UK.

Rheingold, H. (1993). The Virtual Community: Homesteading on the Electronic Frontier. *Addison-Wesley* 57-58.

Schwabe, G., & Prestipino, M. (2005). How tourism communities can change travel information quality. *13th European Conference on Information Systems (ECIS)*.

Sigala, M. (2007). *Web 2.0 in the tourism industry: A new tourism generation and new E-Business models*. http://www.ba.aegean.gr/m.sigala.

Stipanuk, D. M. (1993). Tourism and Technology. Interactions and Implications. In *Tourism Management*.

Stockadale, R. (2006). Borovicka M., Developing an Online Business Community: A Travel Industry Case Study. *39th Hawaii Int Conference on System Sciences*.

Veijalainen, J., Virrantaus, K., Markkula, J., & Vagan, T. (2001). Developing GIS-Supported Location-Based Services. In Ozsu, T., Schek, H.-J. & Tanaka, K. (eds.), *Proc of the 2nd International Conf on Web Info Systems Eng* (WISE 2001), Kyoto. 3-6 Dec. (pp. 423– 432).

Wang, Y., & Fesenmaier, D. (2004). Modeling Participation in an Online Travel Community. *Journal of Travel Research*, *42*(3). doi:10.1177/0047287503258824

Wang, Y., Yu, Q., & Fesenmaier, D. R. (2001). Defining the Virtual Tourist Community: Implications for Tourism Marketing. *Tourism Management*, *23*(4), 407–417. doi:10.1016/S0261-5177(01)00093-0

Werthner, H., & Klein, S. (1999). *Information Technology and Tourism – A Challenging Relationship*. New York: Springer Verlag, Wien.

White, D. (2007). SPIRE Project – *Results and analysis of Web 2.0 services survey*, Version 1.0-2007. http://spire.conted.ox.ac.uk/trac_images/spire/SPIRESurvey.pdf.

World Tourism Organisation (WTO/OMT). (1997). International Tourism: AGlobal Perspective. *WTO Tourism Education and Training Serie*s, Madrid.

Zipf, A., & Malaka, R. (2001). Developing location based services for tourism – The service providers view. In Sheldon, P. J., Wöber, K. W. & Fesenmaier, D. R. (eds.): Information and Communication Technologies in Tourism, 83–92. *8th International Congress on Tourism and Communications Technologies in Tourism*, Montreal, Canada. Springer, 2001.

KEY TERMS

Mobile Interaction: The relation between users using mobile devices, it allows users to communicate and to access Web applications from any place.

Mobile Technologies: A parent category for mobile telephony, mobile computing, and miscellaneous portable electronic devices, systems, and networks.

Online Tourism: The purchases, the selling or the consultation about services tourism on Internet, both Web and mobile technologies.

Social Network: A community of people who share interests and activities, or who are interested in exploring the interests and activities of others. Each user create contents and uses others' contents.

Sustainable Tourism: A tourism which meets the needs of the present tourists and host regions while protecting and enhancing opportunities for the future.

Virtual Travel Community: A group of people that primarily interact via communication media such as forum, blog, chat room, instant messaging and that are interested in sharing information about tourism or travel.

Web Interaction: The relation between users and Web applications in order to search information about their own needs or for fun.

This work was previously published in Handbook of Research in Mobile Business: Technical, Methodological and Social Perspectives, Second Edition, edited by B. Unhelkar, pp. 293-303, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.10

Healthcare Quality and Cost Transparency Using Web-Based Tools

Jiao Ma
Saint Louis University, USA

Cynthia LeRouge
Saint Louis University, USA

ABSTRACT

This chapter explores the use of Web sites to provide patients with understandable information about the quality and price of healthcare healthcare transparency). Our first objective is to discuss patients' perceptions of empowerment and need for quality and cost information when choosing medical providers and facilities for healthcare procedures. To meet this objective, we address issues of patient awareness of sources of healthcare quality and cost information, perceived responsibility for managing healthcare costs, and knowledge of appropriate actions to exercise choice of providers. Our second objective is to investigate the potential of Web-based tools, which provide healthcare quality and cost information, to facilitate patients' decision-making processes regarding choice of provider for healthcare services, particularly common outpatient procedures. To meet this second objective, we use insights from user-centered design procedures (e.g., focus groups and in-depth interviews) associated with the development of a healthcare transparency Web-based tool.

DOI: 10.4018/978-1-60566-016-5.ch011

INTRODUCTION

On May 10, 2006, Carolyn Clancy, M.D., Director of the Agency for Healthcare Research and Quality, testified before the Joint Economic Committee of the U.S. Congress outlining the commitment of the Department of Health and Human Services (HHS) to provide Americans with understandable information about the quality and price of healthcare. This vision has four objectives:

- Promote quality transparency;
- Promote price transparency;

- Facilitate the greater use of health information technology; and
- Transform healthcare so its incentives support a consumer-oriented healthcare system.

The synergy of these objectives is an informed and empowered healthcare consumer with a panoramic view of his/her healthcare situation.

The Internet has done much to facilitate this view by providing an ever-expanding bounty of information about health prevention and maintenance; however issues such as the general understandability and accuracy of this information still remain. Interactive tools and features that provide communication channels (e.g., peer support groups) and personalized information are increasingly common supplements to content pages. Monitoring devices have also been an area of great advancement and promise.

It is evident that patients have an appetite for prescriptive and preventative healthcare information and tools. However, does this need also exist for healthcare cost and quality transparency? The first objective of this chapter is to better understand patients' perceptions of empowerment and need for procedure quality and cost information when choosing medical providers and facilities. The second objective of the chapter is to address the potential of Web-based tools to facilitate and enable the patients' decision-making processes. To meet these goals, questions of interest include:

- Do patients perceive quality and cost information as important to their choice of healthcare provider?
- Do patients feel empowered with quality and cost information to make decisions about where to have medical procedures (e.g., mammogram and bone density tests) or tests (e.g., spinal tap or allergy testing) performed?
- Would a Web-based tool that provides quality and/or cost information be compatible with the patients' decision-making processes?
- What factors should be considered in designing a useful and usable Web-based tool to provide quality and/or cost information for patients?

We review existing literature, assessment of existing healthcare transparency Web applications, and insights from user-centered design techniques from an on-going study involving developing such a tool to reveal issues, controversies, and problems associated with healthcare quality and cost Web-based tools, and discuss solutions. We adapt the constructs from a Theoretical Compatibility Model (Karahanna, Agarwal, & Angst, 2006) to use as thematic dimensions to organize our presentation of Web-based quality and cost tool compatibility, as well as provide insight within the landscape of existing information systems research.

BACKGROUND

In America's democratic consumer culture, consumers can easily obtain information about price and quality for most purchases using existing resources (e.g., mass media such as TV, print publications, Web sites) to assist with purchase decisions. A noted exception is healthcare where cost and quality information is still limited (Greenberg, 2006). In a consumer-driven system, healthcare buyers with financial incentives will demand quantitative and qualitative information on competing health plans, facility options, providers, and specific healthcare treatment. Plans and providers will be motivated to supply this information in order to manage cost and utilization as necessary to stay competitive.

Healthcare transparency is about disseminating information (quality and cost) in a patient-friendly manner to allow patients to make informed healthcare decisions based on value. Ideally, transparency creates incentives at all levels and motivates the entire system to provide better care for less money (HHS, 2007). Providers will improve when they see how their practices compare to others, and consumers benefit from making informed decisions. Additionally, it is theorized that healthcare transparency will encourage patients to become better healthcare shoppers by factoring quality and price into decisions about care, which will lead to a positive impact on management of healthcare costs at both individual and industry levels (Openshaw, 2006). It is through these ideals that some embrace healthcare transparency as an element in change and reform in America's healthcare system (Greenberg, 2006). A modern healthcare system and a democratic consumer culture demand a modern information strategy (Lansky, 2003).

The Internet plays and will continue to play a key role in consumer research, plan enrollment, ongoing administration, and customer service (Beauregard, 2006). According to Coulter (2005), alternative ways of accessing health advice are becoming more popular. For instance, a recent UK survey of people aged 45 and older found over 30% of respondents had searched for health information on the Web. In addition, the Agency for Healthcare Research Quality (AHRQ) recommends a wealth of Web-based resources for healthcare quality information and tips, as well as print resources (AHRQ, 2007) to assist consumers in making more informed healthcare decisions. As such, emerging Web-based tools and resources are a logical choice for providing patients with healthcare quality and cost information—for example, CompareYourCare hosts 22 tools (including three English/Spanish bilingual tools) that compare and rate the quality of healthcare, according to national treatment guidelines, and assist self-care by providing up-to-date information and skills to better manage an illness or maintain good health (http://www.compareyourcare.org/). This site was developed by the Foundation for Accountability (FACCT), and is owned by HealthGrades Inc. Among the efforts of making healthcare more "transparent," a handful of states like New York, Florida, Maryland, and Texas provide state-sponsored Web sites (e.g., New York State Healthcare Report Card at http://www.abouthealthquality.org/index/about) which enable patients to compare hospital and, physician prices (a range of charges or average prices) and performance.

The aforementioned states are using this type of "transparency" to force poorly performing providers out of business and push providers at the high end of the spectrum to voluntarily lower their fees (Lemov, 2006). Hospitals were found to respond positively to public reporting by immediately improving their quality scores during the preparation of disclosing quality data. Most of the hospitals were able to spot their failings from site information and changed policies as appropriate to meet the standards.

The objectives of these sites are not only to prompt organizational action, but also to influence patients' decision-making behavior. Site information about the cost and quality of care is offered to patients to accommodate their higher responsibilities in paying and managing healthcare costs. However, despite the free and easily accessible data, little empirical evidence has shown that patients have altered their behaviors in response to publicly reported quality or cost measures (Lemov, 2006). In fact, there is little elaboration on patient issues related to cost and quality transparency. Furthermore, there is no clear indication that patient input was considered in creating the design and content of these sites. Hence, this chapter focuses on trends, issues, controversies, and solutions for using Web-based tools to provide healthcare transparency from the patient perspective.

TRENDS, ISSUES, AND CONTROVERSIES

Quality and Cost in the Healthcare Decision Process

Do patients perceive quality and cost information as important to their healthcare provider decision-making process?

Service marketing literature indicates that customers attempt to evaluate both functional quality (i.e., delivery quality, which includes an assessment of reliability, responsiveness, assurance, empathy, and tangibles) as well as technical performance quality in deriving quality judgments. However, the entire quality judgment will be made on delivery quality elements, if someone lacks the ability to assess technical quality (Gronroos, 1984). In reference to healthcare, Bowers and Keife (2002) pointed out that functional quality traditionally has received more weight, because patients often lack the technological sophistication to understand details of their diagnoses and treatment and place paternalistic trust in their physicians to make the right decisions.

However, times are changing. The 1996-2000 FACCT study on patients and healthcare quality found that patients are beginning to seek outside information and use it to ask questions of their healthcare providers and participate in healthcare decisions (FACCT, 2000). About one-third of the patients in the FACCT study used the Internet for seeking health information. When selecting a physician, the FACCT study found that patients were interested in factual information, such as years in practice, malpractice record, hospital affiliation, and so on. About 28% of those studied accessed Web sites to read and post messages about a physician. Segmentation of the adult population in the FACCT study suggests that the most passive healthcare participants are older men with lower than average incomes with at least one chronic condition, while the most aggressive information seekers are women with higher-than-average incomes.

Recent trends also show a shift related to healthcare costs. American healthcare is moving toward consumer-driven health plans, which require significant copays or deductibles and/or utilize personal flex spending accounts (these accounts allow participants to allocate limited amounts of pre-tax dollars to cover healthcare costs). As early as 1986, Rosenstein (1986) called attention to the trend in making patients more financially responsible for medical care services by increasing their out-of-pocket costs. For patients with consumer-driven plans, cost is an important factor. In fact, Rosenstein predicted that cost will eventually become one of the major criteria used in healthcare selection. In response to patients' increased concerns regarding costs, changes in healthcare are under way. For instance, walk-in retail clinics, which target providing convenient access/limited service at a competitive price, are spreading nationwide. More than a dozen clinic operators plan to open thousands of clinics in stores such as Wal-Mart, CVS, Walgreen's, and Kerr Drug. Prices at these walk-in clinics are posted on an electronic sign as in a fast-food restaurant (Schmit, 2006).

Cost and quality information are not separate issues, but are intertwined in making an assessment of healthcare value. In a national survey of healthcare consumers in Taiwan on patient perceived quality and expensiveness, Cheng, Wei, and Chang (2006) found that patients indicating higher perceived quality (i.e., physician technical or functional interpersonal skill ratings) were less likely to report that charges were expensive. Using consumer cost theory, this study concluded that perceived value rather than the price or quality alone is the essence of competition in the healthcare market. Hence, when the quality of the healthcare services meets a consumer's expectation and the consumer considers the price to be worth the anticipated quality, then consumer satisfaction emerges and consumption of those services continues.

Consumer Empowerment

Do patients feel empowered with cost and quality information to make decisions about where to have medical procedures performed?

The traditional paternalistic approach of decision-making in healthcare assumed that (1) doctors and patients shared the same goals; (2) only the doctors were sufficiently informed and experienced to decide what should be done; and (3) patient involvement would be limited to giving or withholding consent to treatment. Evidence shows that higher engagement of patients in decision-making and active involvement in managing their healthcare leads to more appropriate and cost-effective use of health services and better health outcomes (Coulter, 2005). Consumer empowerment includes consumer activation and public disclosure of performance information (Bethell, 2000). Consumer activism refers to patients proactively seeking information about healthcare, health, quality, and cost (e.g., looking up information about a new prescription such as precautions and side-effects). Public disclosure refers to a critical mass of patients having timely access to relevant and understandable information about healthcare performance (e.g., obtaining information on healthcare professionals' backgrounds before agreeing to see them).

When it comes to quality and cost information, the healthcare industry has been described as an imperfect competitive market of uncertainty and information asymmetry (Cheng et al., 2006). This asymmetry impairs patient empowerment in the decision-making process. Healthcare might be the only industry where a consumer can walk through the door, obtain a service not knowing how much it costs, and walk out without paying. This point is highlighted by a recent study, which found patients were able to guess the price of a Honda Accord within $300 and estimate the tab for a coast-to-coast round-trip airline ticket within $37, but were off by $8,100 for a four-day hospital stay (Howley, 2006). Hence, public disclosure of healthcare quality and cost information provide information that is not presently known. This additional information may empower consumers to exercise judicious choices in healthcare decisions, such as where to have medical procedures performed.

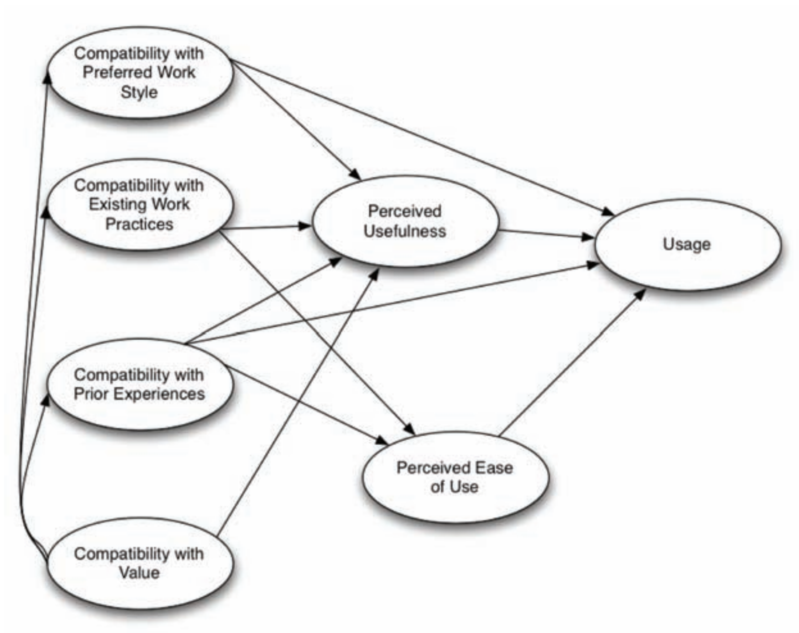
Web-Based Quality and Cost Tool Compatibility

Would a Web-based tool that provides quality and/or cost information be compatible with the patient's decision-making process?

While the trend is toward patient empowerment and healthcare information seeking, the question of compatibility between a Web-based tool and the patient's decision-making process is not yet sufficiently answered. Compatibility is a multi-faceted issue. Two general aspects of compatibility are normative/cognitive compatibility (referring to feelings or thoughts) and operational compatibility (referring to the compatibility of an innovation to what people do) (Rogers, 1995; Tornatzky & Klein, 1982). With respect to cognitive compatibility, patients' thoughts and feelings on how cost information factors into their healthcare decision model need to be considered. If cost information is not currently a major factor, a key challenge in controlling costs through cost transparency will be how to best reframe patients' thinking about cost as a decision factor. With respect to operational compatibility, the propriety of a Web tool used for this purpose and its design need to be further considered. Recent work in the information systems domain further breaks down operational compatibility into compatibility with preferred work style, existing work practices, and prior experience, as well as elements of technology acceptance (Karahanna et al., 2006).

There is little evidence in research, and to some degree in practice, to determine whether

Figure 1. Theoretical compatibility model (from Karahanna et al., 2006, pp. 789)



providing healthcare quality and cost transparency via the Internet is compatible with patients' existing needs and preferences. Compatibility in this context is complicated by the issues that what is deemed as relevant information is situational to patients (e.g., specific insurance coverage or health condition). The information infrastructure needs to offer information that people want, and distribute it when and where they need it. In addition, the information delivered by a Web-based tool needs to be personalized to meet each patient's and family members' specific needs (e.g., health needs, language, culture, information seeking behavior) (Lansky, 2002).

To begin to explore the compatibility issue, we adapt the compatibility antecedent constructs from the Theoretical Compatibility Model (Karahanna et al., 2006) to use as a dimensional (thematic) framework to disclose potential compatibility issues. Figure 1 shows the model from the Karahanna et al. study; the model exhibits some statistically significant relationships, along with theoretical relationships proposed for "compatibility with existing practices" that could not be tested. We use the adapted model as a priori prescriptive framework, as opposed to an assessment of a completed innovation in the previous research.

Table 1 provides a comparison of compatibility construct definitions used in the previous study with those used in this study. Some adaptation of construct definition is needed to suit the context of patients' volitional use of a Web-based tool. In the forthcoming section, we propose this model as a way to inform future design and acceptance of an innovation.

PATIENT COMPATIBILITY: CONSIDERATIONS TO ENHANCE EASE-OF-USE AND VALUE

What factors should be considered in designing a useful and usable Web-based tool for providing quality and/or cost information to patients?

To address the final question of interest, we continue the discussion using insights from user-

Table 1. Comparison of construction definitions between Karahanna et al. (2006) and current study constructs

Karahanna et al. (2006) Construct	Karahanna et al. (2006, pp. 787) Definition	Current Study Construct	Current Study Definition
Compatibility with Preferred Work Style	Captures the possibility offered by the technology of being consistent with a desired work style	Compatibility with Preferred Task Style	Patient's self-concept regarding the functionality (way) they would like to explore healthcare cost and quality information; cost and quality functional requirements
Compatibility with Existing Work Practices	Measures the extent to which a technology "fits" with a user's current work process	Compatibility with Existing Task Practices	Current practices patients use to find cost information or information seeking strategies used to find health information on the Internet.
Compatibility with Prior Experience	Reflects a fit between the target technology and a variety of users' past encounters with technology	Compatibility with Prior Experience	Existing knowledge or perceptions of fact regarding cost/quality information and experience in using healthcare cost comparison sites.
Compatibility with Value	Epitomizes the match between the possibilities offered by the technology and the user's dominant value system	Compatibility with Healthcare Cost and Values	Patient's dominant value system that affects factoring of healthcare cost and quality into healthcare decisions.

centered design procedures associated with the development of a particular healthcare transparency Web-based tool (sponsored by a healthcare insurer). We provide insights regarding patients' preferred task styles, existing task practices, prior experiences, and values to use as considerations in developing a transparency tool compatible with the patient perspectives. Insights from these procedures are not intended to draw final conclusions, but to inspire further work in research and practice in meeting patient healthcare quality and cost transparency needs through the use of Web-based tools.

User-Centered Design Methods

The user-centered design (UCD) methods that were used include focus groups and in-depth interviews (IDIs in-depth interviews (IDIs) with primary target users. Participants for both the focus groups and IDIs were randomly selected from a consumer sample provided by a professional marketing firm. The following selection criteria were used to define the sample:

- A variety of group and individual health insurance plans were represented
- A mix of males/females
- Variety in the age range between 18–55 years
- Participants were involved in, or responsible for, household healthcare decisions
- Insurance customers with an annual deductible under $500 and those with an annual deductible in excess of $500
- Participants were computer and Internet users

Members of the pool were assigned to either a focus group or an IDI. Two focus groups were conducted with the potential users of a Web-based tool. Each focus group consisted of eight-to-ten participants and lasted approximately one and one half hours. One focus group included insurance customers with an annual deductible under $500, and one group with an annual deductible in excess of $500. An experienced moderator and a panel of analysts conducted and analyzed the focus groups, respectively. Appendix A provides a general outline of the protocol used to explore beliefs and attitude structures related to healthcare transparency and the use of a Web site to provide this transparency. Exact questions and the flow of conversation evolved around the topics to best suit the characteristics and flow of the groups.

Twelve people participated in IDIs individually. IDIs are a nondirective, qualitative type of interview with an emphasis on listening to people talk in response to a minimal number of specific questions, which are primarily open-ended (Miller & Crabtree, 1999). This research technique provides value through qualitative insight into the belief and attitude structures, which may be used, particularly in business contexts, for strategy development. The IDIs were conducted by a professional interviewer and each lasted approximately 45 minutes. The objectives of these interviews that guided the discussion were: (1) to examine reactions to the concept and design of the Web-based tool, and (2) to extend exploration of focus group topics regarding healthcare transparency beliefs and attitudes. The IDIs began with a discussion of beliefs particularly related to cost transparency and migrated to concept evaluation. Projective techniques were used during the IDIs to allow respondents to engage in free-flowing and creative descriptions of their healthcare cost and quality information needs and how to meet these needs using the prototyped Web-based tool. The concept discussion was followed by a prototype review. The prototypes used in the IDIs were high-fidelity mockups of screen designs, which were a lifelike simulation of the final product with a refined graphic design. However, the back end of the product was simulated rather than real (Isensee & Rudd, 1996).

The Theoretical Compatibility Model in Figure 1 was used as a coding scheme for organizing insights from the investigation using the UCD techniques into compatibility dimensions. Two researchers participated in the coding process and worked together to reach final coding consensus.

Preferred Task Style

To make a decision on the choice of a provider or facility, the UCD data indicated that patients need quality information to accompany cost information, which aligns with the previous discussion of value. When it comes to quality information, participants stated they wanted performance metrics for the healthcare facilities. For cost information to be useful, it must be specific. In other words, participants indicated that patients need explicit and accurate cost information for obtaining the procedure of interest at specific facilities, instead of an average or a range of costs. Hence, the ability to specify parameters within the Web-based tool to customize information to individual situations is important. For example, some participants indicated they liked to be able to calculate their "out-of-pocket" costs based on their individual plan benefits. To meet this need, a calculator to determine "out-of-pocket" cost is an appropriate feature. As another example, some patients want to limit their search to facilities within a specific driving distance.

Multiple sources of data enhance usefulness, particularly regarding quality. To illustrate, some participants expressed an interest for a health transparency Web tool to include links to facility Web sites for further research regarding quality. In addition, people who have undergone similar procedures were considered as a valuable source of information. This was expressed by the desire for online discussion groups and other forums to communicate with other patients. As another alternative, a simple star ratings system was recommended to provide patient feedback for each facility that offers a specified procedure.

However, the UCD data recommended avoiding situations in which patients drown in data and cannot make use of the information. Healthcare cost information must address specific information needs, in order to be useful and meaningful to the patients. Participants mentioned the following factors, which need to be considered in specifying meaningful data regarding healthcare quality and cost transparency:

- Type of service anticipated by the patients
- Specific hospital/clinic providing the service
- Insurance carrier or coverage medium of the patient
- Benefit plan of the patient (deductible, co-insurance, out of pocket limit)
- Procedural information about procedures (such as side-effects and recovery time) is desirable and should be presented with intuitive graphics and imagery where possible.

In seeking to affect patient behavior, entities offering such tools should consider that the Web tool alone might not be enough for a patient to finalize a decision. Participants indicated a Web-based tool should afford users the opportunity to have personal contact via telephone with a "live person," either during or after using the site for questions, or to call a facility for confirmation once they have narrowed down their choices.

Compatibility with Existing Task Practices

The Web-based tool should function in a way that matches patients' natural information seeking behavior. This corresponds to the fact that the user interface needs to be intuitive, informative, and visually appealing, as expressed in various ways by participants. In reviewing UCD data to affirm these generalizations with specific examples, search, sequencing, and support considerations were mentioned.

A desirable search function should generate only effective and "on target" returns. In inquiring what patients may search for, the indication was that patients would seek information to better understand a specific procedure or learn more about a facility beyond the cost information for a procedure of interest. The name of a facility or procedure would be a common search parameter. Regarding the sequencing of information, proxies for quality or general descriptive information (e.g., bed counts, nursing care information) about a facility must be provided before patients can utilize the cost information in their decision-making processes, again aligning with concepts of value using consumer cost theory.

Support services in various forms are an existing practice with many technological tools. Like most users of complex information technology, healthcare patients feel most comfortable knowing a support person is available to help them, in case they encounter problems with the Web tool or have additional questions.

Compatibility with Prior Experience

Utilization of health transparency Web-based tools is a new experience. In recognizing this, participants indicated that public relations and marketing efforts are a must to promote awareness and, most importantly, the value of a health transparency Web-based tool.

Earlier in this chapter, we discussed the general information asymmetry that currently exists, that patients are not well informed regarding healthcare costs, and generally do not know where to look to find this information. Given this asymmetry, patients have little to no experience in accurately estimating healthcare costs and tend to make inaccurate estimates of the costs (Howley, 2006). The UCD data indicates patients may not only be unaware of a specific price or range for a procedure, but that patients may also be unaware that different facilities can have different costs for the same medical procedures or treatments (e.g., colonoscopy). In addition, even if armed with information regarding cost differences, participants doubted that they were empowered to select a facility for medical procedures. Many believed this decision was completely under the control of the physicians. If health transparency tools are to affect behaviors, the lack of knowledge regarding the patient's role in the decision process needs to be remedied through patient education. As such, health transparency Web tools may include such content as "next steps," "how to compare," or "discussing costs with your doctor."

Compatibility with Values

Consumer attitudes and expectations about healthcare are shifting toward increased clinical performance and public accountability. Some patients have become activists who demand and use information about medical treatments, healthcare products, and healthcare standards (Kizer, 2001).

On the subject of healthcare standards for a Web tool, participants indicated concern about the source of the information. Patients demand impartial and credible sources to help with their decisions. The Web tool prototype that participants reviewed during the IDIs was sponsored by a health insurance carrier. Remarks regarding the sponsorship indicated that trust in insurance carriers was generally low among the participants. Many participants believed that carriers bear the responsibility for healthcare costs, and hence, are driven by a "profit motive." Consequently, concerned participants suspected that data provided by the carrier might be biased toward increasing a carrier's profit. To offset perceptions of biased data, there were recommendations to add information provided by credible sources, such as independent ratings of facilities and providers. Examples given included performance metrics, backgrounds, customer evaluations, and activity information (e.g., annual numbers of services performed) for both physicians and facilities.

Regarding the decision process, the UCD data revealed that patients predominantly based their choice of care decisions on the quality of the physician performing the procedure, which is measured using three criteria: (1) past experience with that physician, (2) the physician's reputation and specialization in that procedure, and (3) recommendations of others (e.g., family and friends). The obvious caveat relative to this chapter is that the cost of a medical procedure is not currently a primary factor that patients consider when making decisions regarding personal or family medical procedures.

Part of the emphasis on quality is attributable to minimizing health risks. Participants pointed out that healthcare service was viewed differently from other professional services. Specifically, they indicated some healthcare procedures (e.g., a life-threatening surgery) could not be viewed as commodities. It was also expressed that when it came to a loved-one, particularly a child, even if a procedure was "simple," the focus was quality. Consequently, patients do not view these types of healthcare decisions as potential "shopping" experiences. Also, participants expressed that they had less desire to control costs than with other types of purchases. One focus group participant stated, "Healthcare is perceived [as] too important to be considered as a shopping experience." Participants indicated that they are more likely to research cost information on the Internet for autos, electronic products, and other types of professional services (e.g., funeral, legal service) than they would be for healthcare cost information.

The other reason for concentrating on quality in the healthcare decision process is attributed to the fact that most consumers are not paying directly for medical procedures, whereas they are for other products. Furthermore, patients often feel entitled to insurance coverage or some other form of cost coverage. Participants disclosed a long-standing perception that healthcare costs are someone else's problem. Consequently, UCD participants expressed no need to take responsibility in controlling overall healthcare costs. Underscoring this general position, participants overwhelmingly defined their costs of interest as "out-of-pocket" costs for a medical event, which resulted from copay or coinsurance. Hence, the only time the cost of the procedure may change patient behaviors is when patient out-of-pocket cost is impacted. When patients are not concerned about the cost of procedures until it hits their pocketbook, the potential of the Web tool to change behavior may be restricted. We make the prior statement with caution, because even in cases where out-of-pocket costs exist, UCD data does not reveal any clear indication that patients would use cost information from Web tool on a voluntary basis to change any care decisions.

Earlier in this chapter, we discussed recent trends that show a shift to a greater consumer burden regarding healthcare costs. Some participants in both the IDIs and focus groups showed some variance in their concerns about costs given the nature of the procedures and who were receiving the procedures (less concern with a child or loved-one). Based upon patients' existing values regarding healthcare cost responsibility and defining costs as their out-of-pocket expenses, could movement toward a greater consumer burden shift healthcare transaction processes toward "more of a shopping model?" Cosmetic surgery typically involves out-of-pocket costs and may provide a useful perspective in contemplating this question. Lee and Hoo (2006) argue that most physicians are compensated differently from the way the other professionals were. They further state that physicians would handle medical transactions differently, depending on how they were paid. For instance, a cosmetic surgery transaction has all the characteristics of a normal market transaction in which the seller has a financial interest in how all aspects of the transaction affect the buyer. The cost information of the typical cosmetic surgery is often readily available in three ways: (1) a package price in advance covering all services and facilities; (2) price comparison/shop-around prior to the surgery; and (3) a price that is lower in real terms than the price charged a decade ago for comparable procedures despite considerable technological innovations in the interim. Advertisements may even indicate a price range. There is ample evidence to suggest that consumers consider costs and "shop around" and that cosmetic surgeons compete with at least some consideration to costs. In contrast, in more typical physician-patient interactions, physicians are not paid to be concerned about all aspects of the transaction and therefore typically ignore the effects on the patients of a variety of costs, including cost of time, the cost of drugs, and other additional costs. In summary, differences exist in decision-making for elective medical procedures (e.g., cosmetic surgeries), as opposed to necessary ones (e.g., life-saving surgeries). According to information retrieval studies (Taylor, Cool, Belkin, & Amadio, 2007; Xu & Chen, 2006), costs constitute a relevant factor in information seeking process of elective "less threatening" medical procedures, however, they may not be as relevant or important in selecting essential procedures.

CONCLUSION

Although trends are moving toward patient empowerment and considerations of healthcare quality and cost, patients are not currently empowered when it comes to healthcare quality and cost information. In addition, patients perceived that they had limited responsibility for healthcare costs. The ultimate reason to look at compatibility issues early in the design process is to have some foresight to determine and enhance the potential for technology acceptance and outcomes. In aggregating literature and insights from the UCD data, what healthcare transparency Web tools have to offer may become more compatible with the patients' decision-making process as the healthcare industry continues with emerging trends toward patient information empowerment and increasing patient healthcare cost burden.

Currently, providing health transparency through a Web-based tool will likely influence the consumer healthcare decision processes the most in situations not involving extremely serious healthcare conditions and when the consumer bears the greatest cost burden. Web-based tools that provide understandable comparative cost information and quality indicators may best enable healthcare patients in the decision processes. Furthermore, such tools may also best serve their purposes by providing features that help the consumer deliberate (e.g., forums with peers, calculator, and further contact information).

The Theoretical Compatibility Model (Karahanna et al., 2006) proved to be a viable framework for analyzing the patient perspective and may be well suited to further explore and test consumer health informatics issues related to quality and cost transparency (e.g., acceptance of such Web sites and actual changes of behaviors resulting from using the Web site), as well as other consumer health issues (e.g., adoption of various forms of patient health records). Future work will include assessing existing and forthcoming healthcare transparency Web tools' ability to raise cost and quality awareness and encouraging the use of this information as a change agent for decision-making.

REFERENCES

Agency for Healthcare Research and Quality (AHRQ). (2007). *Improving healthcare quality: A guide for patients and families*. Retrieved May 26, 2008, from http://www.ahrq.gov/consumer/qntlite/

Beauregard, T. R. (2006). *Large employer views of consumer-driven healthcare: Current interest, barriers, and potential solutions.* Manhattan Institute. Retrieved May 26, 2008, from http://www.manhattan-institute.org/consumerdriven-healthcare/pdfs/Beauregard.pdf

Bethell, C. (2000). *Patient-centered care measures for the national healthcare quality report (defining patient-centered care).* Retrieved May 26, 2008, from http://www.markle.org/resources/facct/doclibFiles/documentFile_168.pdf

Bowers, M. R., & Kiefe, C. I. (2002). Measuring healthcare quality: Comparing and contrasting the medical and the marketing approaches. *American Journal of Medical Quality*, *17*(4), 136–143. doi:10.1177/106286060201700403

Cheng, S.-H., Wei, Y.-J., & Chang, H.-J. (2006). Quality competition among hospitals: The effects of perceived quality and perceived expensiveness on healthcare patients. *American Journal of Medical Quality*, *21*(1), 68–75. doi:10.1177/1062860605283618

Coulter, A. (2005). What do patients and the public want from primary care? *British Medical Journal*, *331*(7526), 1199–1201. doi:10.1136/bmj.331.7526.1199

Greenberg, J. (2006, July 18). *Will consumer driven healthcare put the brakes on rising costs?* New Hampshire Public Radio. Retrieved May 26, 2008, from http://www.nhpr.org/node/11210

Gronroos, C. A. (1984). A service quality model and its marketing implications. *European Journal of Marketing*, *18*, 36–44. doi:10.1108/EUM0000000004784

Health and Human Services (HHS). (2007). *Value-driven healthcare—transparency: Better care lower cost*. Retrieved May 26, 2008, from http://www.hhs.gov/transparency/

Howley, K. (2006, April). "I can't afford to get sick"—Think your insurance has you covered? Just wait till you need it most. *Reader's Digest*. Retrieved May 26, 2008, from http://www.rd.com/content/the-cost-of-health-care-in-america/

Isensee, S., & Rudd, J. (1996). *The art of rapid prototyping*. MA: International Thompson Computer Press.

Karahanna, E., Agarwal, R., & Angst, C. M. (2006). Reconceptualizing compatibility beliefs in technology acceptance research. *MIS Quarterly*, *30*(4), 781–804.

Kizer, K. W. (2001). Establishing healthcare performance standards in an era of consumerism. *Journal of the American Medical Association*, *286*(10), 1213–1217. doi:10.1001/jama.286.10.1213

Lansky, D. (2002). *Information for better quality care—the patients' perspective*. Foundation for Accountability. Retrieved May 26, 2008, from http://www.markle.org/resources/facct/doclibFiles/documentFile_526.ppt

Lansky, D. (2003). *A person-centered view of consumer information in the healthcare marketplace*. Federal Trade Commission/U.S. Department of Justice Joint Hearings on Healthcare and Competition Law and Policy. Retrieved May 26, 2008, from http://www.markle.org/resources/facct/doclibFiles/documentFile_595.ppt

Lee, P. V., & Hoo, E. (2006). Beyond consumer-driven healthcare: Purchasers' expectations of all plans. *Health Affairs 25*(6), w544-w548. Retrieved May 26, 2008, from http://content.healthaffairs.org/cgi/content/abstract/25/6/w544

Lemov, P. (2006, September). Healthcare: A dose of transparency. *Governing Magazine*. Kaiser Family Foundation. Retrieved May 26, 2008, from http://www.governing.com/articles/9hcare.htm

Miller, W. L., & Crabtree, B. F. (1999). Depth interviewing. In W.L. Miller, & B.F. Crabtree (Eds.), *Doing Qualitative Research* (pp. 89-108). Thousand Oaks CA: Sage Publications.

Openshaw, J. (2006, October 31). Checkup lines: Americans need to learn to shop for healthcare. *Market watch*. Retrieved May 26, 2008, from http://www.marketwatch.com/News/Story/consumers-need-become-better-health-care/story.aspx?guid=%7B09F2F091-A840-4D08-AAE3-C4CDD9A8C02C%7D

Rogers, E. M. (1995). *Diffusion of innovations*. New York: The Free Press.

Rosenstein, A. H. (1986). Applications of consumer behavior in medical care. *Journal of Professional Services Marketing*, *1*(4), 31–37. doi:10.1300/J090v01n04_05

Schmit, J. (2006, August 28). Could walk-in retail clinics help slow rising health costs? *USA Today*. Retrieved May 26, 2008, from http://www.usatoday.com/money/industries/health/2006-08-24-walk-in-clinic-usat_x.htm

Taylor, A. R., Cool, C., Belkin, N. J., & Amadio, W. J. (2007). Relationships between categories of relevance criteria and stage in task completion. *Information Processing & Management*, *43*(4), 1071–1084. doi:10.1016/j.ipm.2006.09.008

The Foundation for Accountability (FACCT). (2000). *Patients and quality: What do they know? What do they want? Results from FACCT consumer research 1996-2000*. Retrieved May 26, 2008, from http://www.markle.org/resources/facct/doclibFiles/documentFile_196.pdf

Tornatzky, L. G., & Klein, K. J. (1982). Innovation characteristics and innovation adoption-implementation: A meta-analysis of findings. *IEEE Transactions on Engineering Management*, *29*(1), 28–45.

Xu, Y., & Chen, Z. (2006). Relevance judgment: What do information consumers consider beyond topicality? *Journal of the American Society for Information Science and Technology*, *57*(7), 961–973. doi:10.1002/asi.20361

APPENDIX A: FOCUS GROUP PROTOCOL OUTLINE

- To discover the current state of cost information –
  - Do patients have an awareness of costs – if so, sources?
  - How/if patients currently use cost information?
- To discuss the future state of cost transparency –
  - Relevancy of information – what do they need to know, when?
  - Information access – how do they want to get the information?
  - Credibility of information – what makes it trusted?
  - What other factors must play in with cost – i.e. quality?
  - At what point in time does the shopping process start? Proactive, Reactive, not Active?
  - Roles of decision makers/influencers – family members, friends, third party sources, physicians, etc.?
  - Which care situations are open to choice/cost considerations, and which ones aren't?
  - What features are most important when a procedure needs to be done? How does this vary by situation?
- To explore how to encourage use of cost information as a change agent -
  - Discussion about plan designs – current vs. prototypes – what would encourage cost research/factoring into decision-making?
  - Incentives to factor price into decision-making?
- To determine what patients mean by quality of care.

This work was previously published in Patient-Centered E-Health, edited by E. Wilson, pp. 143-156, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.11
Exploiting Collaborative Tagging Systems to Unveil the User-Experience of Web Contents:
An Operative Proposal

A. Malizia
Universidad Carlos III de Madrid, Spain

A. De Angeli
University of Manchester, UK

S. Levialdi
Sapienza University of Rome, Italy

I. Aedo Cuevas
Universidad Carlos III de Madrid, Spain

ABSTRACT

The User Experience (UX) is a crucial factor for designing and enhancing the user satisfaction when interacting with a computational tool or with a system. Thus, measuring the UX can be very effective when designing or updating a Web site. Currently, there are many Web sites that rely on collaborative tagging: such systems allow users to add labels (tags) for categorizing contents. In this chapter the authors present a set of techniques for detecting the user experience through Collaborative Tagging Systems and we present an example on how to apply the approach for a Web site evaluation. This chapter highlights the potential use of collaborative tagging systems for measuring users' satisfaction and discusses the future implications of this approach as compared to traditional evaluation tools, such as questionnaires, or interviews.

DOI: 10.4018/978-1-60566-208-4.ch009

INTRODUCTION

Collaborative tagging is the process by which users add metadata to a community-shared content, in order to organize documents for future navigation, inspection, filtering, or search. The content is organised by descriptive terms (tags), which are chosen informally and personally by the user. The freedom to choose unstructured tags is the main distinctive feature of collaborative tagging systems, as compared to traditional digital libraries or other systems of content organization, where the creation of metadata is the task of dedicated professionals (such as librarians) or derives from additional material supplied by the authors (Bennis et al. 1998, Csikszentmihalyi, 1997). Like all socially-generated structures, tagging is an adaptable process; it takes the form best supported by the content, letting users decide the categorization of such content, rather than imposing a rigid structure on it. Collaborative tagging is most useful in an environment like the World Wide Web, where a single "content classification authority" cannot exist and there is a large amount of data content being continually produced by the users.

The widespread success of collaborative tagging systems over the last few years has generated a large collection of data reflecting opinions on, and evaluation of, web contents. In this chapter, we look into the possibility of exploiting this large database to evaluate the user experience (UX) of web sites. UX is a multi-faceted construct recently introduced into the HCI agenda to describe the quality of an interactive system (Garrett 2003; McCarthy and Wright 2005). This construct is used to indicate how people feel about a product and their pleasure and satisfaction when using it (Hassenzahl and Tracktinsky, 2006). Responses such as aesthetic judgments, satisfaction or frustration, feelings of ownership and identity are the most prominent aspects of user experiences investigated in this new, comprehensive, HCI research area (De Angeli, Sutcliffe and Hartman, 2005; Hartman, Sutcliffe and De Angeli, 2007; Norman, 2004). Normally, these responses are collected in formal evaluation settings via questionnaires and/or interviews. Collaborative tagging may offer an interesting alternative, one which is cheaper and less prone to experimental bias. In this chapter, we present a technique to extract semantics from tagging systems, and interpret them to describe the user experience when interacting with on-line content.

This chapter has the following organisation. Paragraph 2 reviews related works on collaborative tagging systems. Paragraph 3 describes three different techniques that can be used to extract semantics from tagging systems. Paragraph 4 reports a method to derive semantics differential attributes from collaborative tagging systems, 3, and its evaluation. Paragraph 5 summarizes the chapter, delineates future trends in the use of collaborative tagging systems for automating evaluation techniques and draws the conclusions.

BACKGROUND

Collaborative Tagging Systems (Golder et al., 2006; Mathes, 2004) offer their users the possibility to index contents for organizing web-related information, sharing knowledge and opinions. There is a growing number of successful web sites which include collaborative tagging, allowing users to index and share different types of contents. Del.icio.us (http://del.icio.us/), for example, specializes on bookmarking, categorizing and sharing URLs, Flickr (http://www.flickr.com/) allows users to tag photographs they own; Technorati (http://technorati.com/) is devoted to tag weblogs; and Youtube (http://www.youtube.com/) allows tagging videos. Other interesting examples are Snipit (http://www.snipit.org/), which offers the functionality of bookmarking sections of web pages, and CiteULike (http://www.citeulike.org/) or Connotea (http://www.connotea.org/) that allow tagging and commenting references to academic publications.

Collaborative tagging systems allow users to become active contributors in the classification of web-content. Because of this characteristic some authors refer to them as "folksonomy" (Mathes, 2004), short for "folk taxonomy", albeit there is still some debate whether this term is accurate (Golder et el., 2006). Users of collaborative tagging systems do not only categorize information for themselves, but they can also share their classification and browse the information categorized by other users. In fact, many collaborative tagging systems have features for sharing contents and their associated tags among users. They also, offer functionalities for keeping contents private, shared only within a pre-set list of users, or public (shared with everyone).Therefore, tagging is both a personal and a social activity. According to the number of people who can tag the same content and/or to the level of privacy of the tag (shared vs. personal).Collaborative tagging systems are distinguished in "broad" and "narrow" systems (Van der Wal, 2005). A broad tagging system is the result of one item being categorized by many people (Del.icio.us is an example). This can generate a very diverse set of tagging, as different users can enter their preferred terms, with obvious semantic and syntactic variations. There will be some terms that are used by many people to describe one item or many items which are described by the same terms. The concentration of terms can take advantage of power laws (like the Zipf distribution (Zipf, 1949; Newman, 2005)) to quickly see the preferred terms for an item or items.). It states that the frequency of the occurrence of a term is inversely proportional to its frequency class. Zipf has discovered experimentally that the more frequently a word is used, the less meaning it carries.

A narrow folksonomy is the result of one person categorizing one item (Flickr is an example). In this case, tags are private, but users could decide to share their own photos allowing others to view their tags and thus their categorization of contents. When the contents (and tags) are shared with other users a narrow folkosomy can approximate a broad one; nevertheless since the option of sharing contents and tags is leaved to the final user we cannot strictly rely on it.

This paper concentrates on broad collaborative tagging systems, where several users index and share different content. We regard the folksonomy produced by these systems as a result of collective intelligence and social creativity (Fischer 2006): different users contribute to the establishment and dissemination of knowledge. In this vision, collaborative tagging systems are not only important for their primary task (e.g., information retrieval), but they assume a fundamental role in the quest for understanding the user experience. People tag content with words which have both denotative and connotative meaning; these tags are a reflection of their opinion on the content, the service provider and the interface design. We believe that tagging systems act as social dynamics enablers representing the real ”vox populi”; in fact, users can take advantage of tagging information shared by others(Nov, 2007). Tagging systems leave the users free to express their own opinion without restricting them in a frame, such as a questionnaire. We believe that this method is more likely to capture the ecological perception of the web site audience. Collaborative tagging systems offer a lot of unstructured metadata (tags) associated to many different contents (web sites, photos, videos, etc.) that can be used for measuring the UX of these contents over the Internet.

Moreover, collaborative tagging systems allow detecting variations over time, by analyzing how tagging evolve. The goal of our research is to develop a methodology to extract meanings from collaborative tagging systems and to use this information in order to understand what people think about on-line contents. This methodology requires a two phased process: (a) detecting semantics from tagging systems; (b) interpreting the meaning of this information.

DETECTING SEMANTICS FROM TAGGING SYSTEMS

Information retrieval (IR[1]) from unstructured contents such as those produced by tagging systems) is a complex task. A major problem relates to the fact that no current tagging systems have synonyms control (e.g. "Mac" and "Macintosh" do not coincide in Del.icio.us). For this reason, in order to use the information contained in a collaborative tagging system, we need to use techniques extracting semantics from users' tags. In the following paragraph, we discuss three information retrieval techniques that can be used to extract semantic features from tagging.

Many information systems use keywords or key phrases to search or browse collection of documents for specific terms and information. Not only are keywords used for searching relevant documents but also to index and categorize the content. Relevant information is indicated by the authors of a document and is placed in appropriate sections for emphasizing them. Typical examples are the title, abstract and author's name written with bold or in appropriate places of the document. This approach is useful if employed within document collection explicitly managing this information, such as newspapers articles. Nevertheless this information is not available in general and providing them manually can be tedious or inapplicable depending on the amount of relevant keywords or terms we want to provide for each document.

IR algorithms were devised to address this problem, trying to automatically extract relevant terms and keywords from unstructured document collection. IR algorithms employ two different phases (Turney, 2002): keywords assignment, and keywords extraction. Usually, there is a training phase where an initial default list of relevant keywords is provided to the system, thus using a controlled dictionary. The wider is the list, the greater should be the number of documents used to train the system by manually indicating the keywords included in each document (chosen among the given list). These types of algorithms are called training-intensive, i.e. a big training set is required to obtain good performance.; On the contrary, keywords extraction does not need any training since the keywords are directly extracted from the body of each document by using the information learned from the training phase and some similarity measure. In the next section, we present a selection of three IR algorithms that can be used to automatically extracting semantics from collaborative tagging systems.

PMI-IR

The PMI-IR (Point wise Mutual Information – Information Retrieval) algorithm employs the technology of a search engine, such as Google Page Rank, or Yahoo, (Krikos, et al. 2005; Kraft et al. 2006)) to extract the frequency of searched keywords within a collection of documents. In general, the algorithm takes as input a word and a set of alternative terms for that specific word. The output is the selection of the terms whose meaning is the closest to the given word. That is to say, the algorithm finds the synonyms by analyzing the co-occurrences of the terms with the given keyword and among them.

This is exactly the case for tagging systems, where we have a collection of contents labeled with different words representing keywords for that collection and we would like to group words having the same meaning.

There exist different ways of measuring the co-occurrence between two terms, but the one used by PMI-IR algorithm is based on the Point wise Mutual Information (1), where problem represents the given word (tag in a folksonomy), {$choice_1$,..., $choice_n$} represent the n alternatives for problem and P(problem, $choice_i$), i=1,..,n the probability of the co-occurrence.

$$\text{Score}(choice_i) = \log2 \left(P(\text{problem}, choice_i) / P(\text{problem})\, P(choice_i)\right) \quad (1)$$

If problem and $choice_i$ are statistically independent, then the probability of co-occurrence is described by P (problem) P ($choice_i$). If problem and $choice_i$ are not independent (i.e., they tend to co-occur) the numerator in (1) will be greater than the denominator and the ratio will describe the independence rank between the two terms.

By considering that P (problem) is assuming the same value for each associated $choice_i$ and that the log function is monotonically increasing, equation (1) can be simplified as follows:

Score (choicei) = log2 (P (problem, choicei) / P (choicei)). (2)

The conditional probability value P (problem| $choice_i$) is assigned as a measure of how close the words are (synonyms). This measure can be computed, for instance, by using a search engine like Google page rank or Altavista advanced search. P (problem| $choice_i$) represents the number of documents returned by the search engine, called hits, when searching for problem and $choice_i$. The term which is the most similar to the problem is the one that maximizes the measure as shown in (3).

To clarify how this algorithm can be used for extracting semantics from tags, let us give an example. In the first instance we consider as co-occurring two words appearing in the same document, for example tags used in del.icio.us for categorizing the same web site; e.g. www.microsoft.com tagged with both the words 'explorer' and 'windows'. In this context, the score assigned to each $choice_i$ is computed as follows:

Score (choicei) = (hits (problem AND choicei) / hits (choicei)). (3)

The equation reported in (3) assigns as score the value of the ratio between the number of documents containing the two terms (problem and $choice_{i)}$ and the number of documents containing only $choice_i$

The tag which is most correlated to the problem is the one obtaining the highest score value computed as of (3). This is a reasonable is a reasonable measure of similarity among tags and a given term, yet is has some problems. In fact, a good similarity measure should include the totality of the tags included in a tagging system and not only within a basic set of $choice_i$ terms. Different tags can have different meanings depending on the interpretation of the author of the tags. Thus, this approach is suitable for narrow folksonomies where clusters of tags are created by some users who upload the contents using their own way of categorizing contents which are likely to have a narrow and well defined range of synonyms.

Collaborative Tag Suggestion

A new IR algorithm (Xu et al., 2006) has been recently introduced, which is based on tag suggestions for annotating documents in collaborative tagging systems. This method assigns reputation weights to the authors of tags, on the basis of the accuracy of words entered. The system suggests terms to use as tags for documents based on the words which are most frequently used by users with good reputation (good sense-making).

This algorithm still keeps into account the magnitude of the co-occurrences of terms but using a subset of terms used by certain experienced users. The objective is to evaluate which tags are relevant (keywords) for the documents in the folksonomy. This objective is achieved by a ranking among users indicating which ones participate positively (the most reliable) on the tagging process. The notation used for this algorithm is defined as follows:

- Ps (ti/tj; o): it is the probability that a user labels an object *o* with the tag *ti* knowing that the tag *tj* has already been used for the same object (i.e., document). To measure the correlation between the two tags on the object o, the algorithm considers the ratio between the number of users using both ti and tj, and the number of users using only tj.
- Pa (ti/tj): it is the probability that any object is labeled with the tag ti knowing that the tag tj has already been used for the same object. In this case the observation refers only to the tags and not to the objects. To

measure this correlation between the two tags, the algorithm considers the ratio between the number of users using both ti and tj, and the number of users using only tj.

- S (t, o): indicates the score of the tag t on the object o, computed by summing the number of users that labeled o with t.
- C (t): indicates the coverage of a tag, which is the number of objects labeled with t. The greater is the number of objects tagged with t, the less specific is the meaning of the tag t. In other words if t is used very often it is a generic term.

The algorithm works by iterating the selection of the tags ti for which S (t, o) is high and multiplying this by the inverse of C (ti). After selecting the ti with the maximum score, the scores of every other tag t' are changed according to the following statements:

- t' score S (t',o), is decreased removing redundant information, i.e. subtracting the value of the probabilities product of t' and ti used together. In this way the superposition of the suggested tags is reduced, as in S(t',o)= S(t',o)- Pa(t'/tj) S(/tj,o)
- t' score S(t',o), is increased if it co-occurs with the selected tag tj over the object o, as in S(t',o)= S(t',o)+ Ps(t'/tj;o) S(/tj,o). This procedure allows dealing with basic level variations of tags, normalising the score for tags like BLOG, BLOGGING, and BLOGS.

The drawbacks of this approach are related to the fact that there exist narrow folksonomies, like Flickr, where every object o (i.e. a document) is own by the user that uploaded it, or by users to whom the owner has granted access permissions. In these cases, we do not have access to all the information needed for running the algorithm over a wide number of tags. Thus the clustering process does not necessarily represent the users' opinion on the tagged topic; but a specific feeling about the shared content (e.g. a link to a web site, and thus the web site itself) can be detected by means of finding related words among a community of users sharing the same interests.

The Semantic Halo Algorithm

In our previous work we introduced a Semantic Halo technique in order to deal with word semantics in tagging systems (Dix et al., 2006). The basic idea consisted of using co-occurrences of tags to cluster their relationships and meanings.

The Semantic Halo is defined as a set of search results for a given tag made by a set of four features, labeled as **4A**:

- **Aggregation**. Representing all the tags linked or related to a given tag.
- **Abstraction**. It is similar to aggregation but it relates to a direction (increasing and decreasing), thus it contains two subsets:
  - Generalization, tags increasing abstraction with respect to the given tag,
  - Specialization, tags decreasing abstraction with respect to the given tag.
- **Ambience**. It is the context for a given tag. It includes all the possible tags appearing in the same context, and that will be useful for augmenting or refining the user query. This set is built from a basic context set.
- **Age**. It is a list of the Ambience feature elements over time. It helps in retrieving tags ordered by meanings given to them over time.

The algorithm was tested within the Del.icio.us community. In this environment, users submit their links to a website, adding some descriptive text and keywords, and Del.icio.us aggregates their posts with everyone else's submissions allowing users to share their contributions. The algorithm was implemented using Del.icio.us programming APIs (Application Programming Interfaces). This procedures allowed to collect results while the users where tagging. Because the Del.icio.us community is very large and active this test resulted in a quite complex but effective test.

For example, given the tag "university", which is quite general, our algorithm searched over Delicious for related tags and retrieved:

Ambience = {'open', {'learning', 'University'}}
Abstraction = {'online', 'education'} U {'colleges', 'high', 'degree', 'distance', 'Commons'}
Age = (('learning', 'University'), ('open'))[2]
Aggregation = {'soccer', 'gradschool', 'corps', 'indoor', 'course', 'masters', 'research institute', 'cites', 'cincinatti', 'peace', 'demographic', 'content', 'courses', 'innovators', 'urban', 'tournament', 'entrepreneurship', 'liverpool', 'york', 'community-college', 'schools', 'Illinois', 'abroad', 'Content', 'latino', 'Course', 'complexity', 'planning', 'Initiative', 'academiclibrary', 'enterprise', 'semantic-web', 'Education', 'grad', 'scholarship', 'teaching', 'college', 'school'}.

We can observe that the Ambience set is composed of two subsets, associated with two different contexts or meanings of the 'university' tag. The algorithm can solve also basic level variations[3] since the tag 'University' with the capital 'U' is strongly associated with the 'university' tag (without using any parser). The first part of the Abstraction set is related to generalization of the given tag, while the second part is specialization, thus providing a partition of the related tags in increasing and decreasing abstraction. The Age sequence is the ordered set of contexts (meanings) with respect to last updates. The Aggregation set lists all the related tags, and even if there are unwanted tags the majority (as shown in the example above) is clearly related.

The Semantic Halo algorithm is applicable in general to broad and narrow collaborative tagging systems but has the drawback of employing a clustering technique that can be less effective or precise in specific sub-domains originated by users' tags.

Summary

All the different algorithms presented in these paragraphs can be employed for extracting semantics from tags by automatically organising them in classes or synonyms. The designer can choose different algorithms or techniques depending on the characteristics of the considered tagging system.

PMI-IR is quite fast to compute and since it is a standard approach within the IR field, many implementations can be found. The drawbacks of this approach are related to the fact that tags have different meanings depending on the sense-making of the users. As a consequence, for retrieving useful semantics the algorithm should span over the entire collection of tagged contents (considering the different choice_i words). The Collaborative Tag Suggestion approach is very effective but it should be avoided when dealing with narrow tagging systems (Flickr for example) when every tagged content is owned by the user or shared with a specific subset of users granting permissions to them. Finally, the Semantic Halo can be used to extract semantics both from broad and narrow Collaborative Tagging Systems but it is less precise in specific sub-domains of tags (users' annotating contents in on specific domain or topic).

Choosing an IR algorithm is a first step for organizing unstructured content, which is a prerequisite for evaluating the UX over contents shared by tagging systems.

SEMANTIC DIFFERENTIAL IN COLLABORATIVE TAGGING SYSTEMS

This section describes a method to evaluate the information extracted from the tags in order to obtain a measure of the user-experience with web-sites. The evaluation phase in our approach is based on the elaboration of the concept of semantic differential introduced by the psychologist and communication scholar E. Osgood (1975). The original work of Osgood focused on the measurement of meaning, addressing issues of word semantics and psychological differences between words. In his influential research the author proposed a method (the semantic differential) to highlight individual differences in the attribution of meaning to words. The semantic differential measures people's reaction to stimulus words and concepts. Participants are invited to rate the stimulus with a bipolar scale. Each extreme of the scale is labeled by contrasting adjectives, such as bad-good. This technique has been frequently used in psychometrics to measure a number of psychological constructs, and more recently has been employed in HCI to build user satisfaction questionnaires. An example of opposite couples of adjectives used by Osgood methodology is shown in Table 1.

Osgood research has demonstrated that ratings on bipolar adjective scales tend to be correlated, and to cluster around three basic dimensions of response, which account for most of the covariation in ratings. These dimensions, labeled as Evaluation, Potency, and Activity (EPA), have been verified by factor analyses and replicated in an impressive variety of studies.

In our approach, there are no fixed couples of opposite adjectives but the information is extracted from the adjectives freely introduced by the user. The adjectives are then associated to one of the three dimensions evaluation, potency, and activity

Table 1. Opposite couples of adjectives used by Osgood

1.	Angular/Rounded,
2.	Weak/Strong,
3.	Rough/Smooth,
4.	Active/Passive,
5.	Small/Large,
6.	Cold/Hot,
7.	Good/Bad,
8.	Tense/Relaxed,
9.	Wet/Dry,
10.	Fresh/Stale.

Evaluation

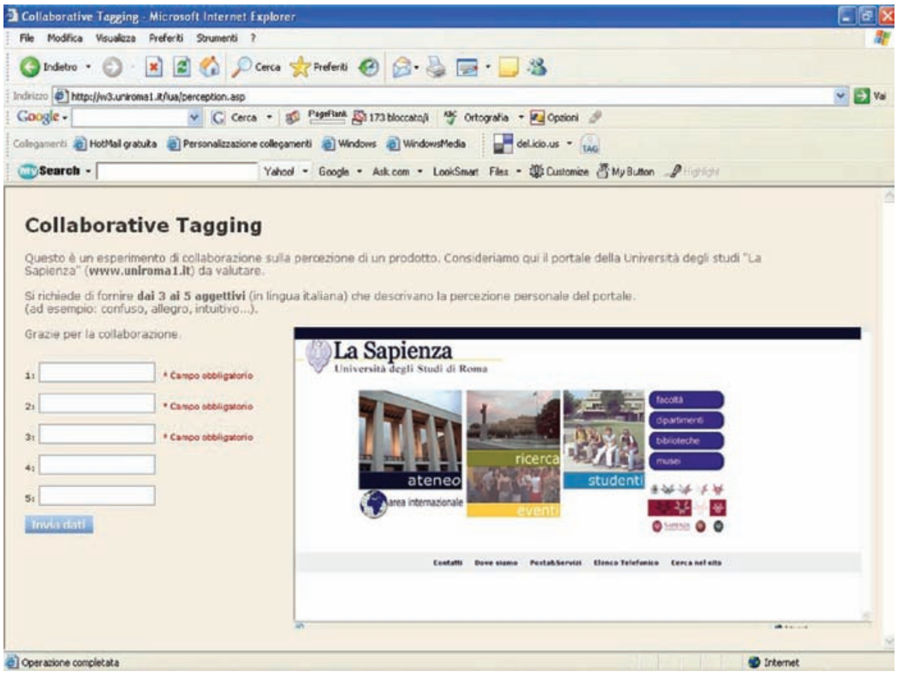
In order to test our methodology of semantic differential through collaborative tagging systems we built a basic collaborative tagging system. Users could add tags for categorizing contents, we explicitly asked users to use adjective as tags for categorizing the contents. Then by employing an IR algorithm we obtained a structured representation of the tags. Successively, we clustered structured tags into groups of adjectives. These groups of adjectives where used for applying the semantic differential technique and obtaining a measure of the UX on the web-site contents.

The objective of our evaluation study was to evaluate the experience of a community of users with respect to the Sapienza University of Rome Italian web portal (www.uniroma1.it). We choose this target because many users of the web portal were complaining about its features and usability. In fact, the Sapienza web portal has now been redesigned.

The community considered in this experiment is composed of 48 people. The majority of users (60%) were students. The remaining sample was split in 20% of administrative staff and 20% of academic staff.

Participants were invited to browse the web-site and tag it with their preferred set of adjectives. This

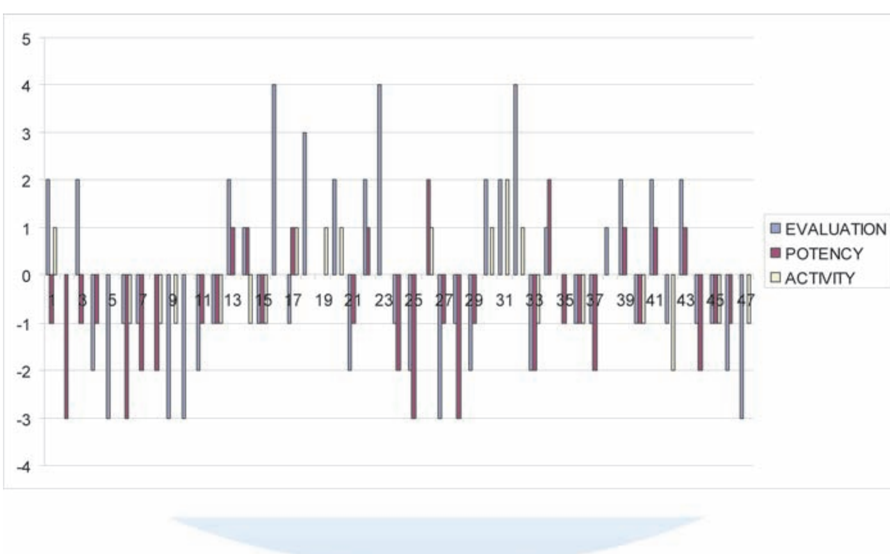
Figure 1. The custom collaborative tagging system including tagging labels and interested web site



system provided users with classic collaborative tagging functionalities, such as: presenting the document to be tagged and the text labels where the corresponding tags could be added. Figure 2 shows the tagging systems used for the experiment, even if displayed in Italian the tagging and web site area are clearly visible.

We collected around 162 tags from 48 individuals in a 2 weeks time-frame. As a first step we analyzed the frequency of each tag (adjective) as

Figure 2. A bar-chart view of the users with respect to the selected classes (evaluation, potency and activity)



presented in Table 2. Analysing the adjectives, it appears that there is a sort of binary distribution of the general tags among positive and negative evaluation. Looking at the frequency distribution (for example f= 11), we found that two very different tags (simple and dazed) are the most frequently used. This effect is evident for almost every couple of tags in Table 2. The effect was also evident in the complete dataset, even with less frequent tags. This let us hypothesize that even after the clustering process (assigning tags to the three classes: evaluation, potency and activity) the user perception would be split in two neat categories according to the overall binary perception: positive or negative.

We categorized the tags in the three classes according to the clustering proposed by Osgood in (Osgood et al, 1975), as shown in Table 3. We used the PMI-IR to automatically measure the distance between the selected adjective (Problem) and the couples of adjectives contained in the Osgood scale. This approach has been employed for the positive and the negative meanings of a tag. The three major factors for a tag to belong to a class are:

- **Evaluation:** Representing the overall feelings about the web site (adjectives like good or bad);
- **Potency:** Representing the expressive power and impact on the perception of the web site (adjectives like strong or weak);
- **Activity:** Representing the possibilities and functionalities (at informational level) offered by the web site (adjectives like: active or passive).

Table 2. Table showing the most occurring tags

Tag	Occurrence
Simple	11
Dazed	11
Clear	6
Sad	6
Comprehensive	6
Inconsistent	5
Intuitive	4
Useless	4
Poor	4

By using the clustering results we analyzed the experience of each user, expressed by the tags inserted in the system with respect to the selected classes (evaluation, potency and activity) and their positive or negative meaning. Figure 2 shows individual results for each user (1-47[4]).

Figure 3 suggests that, in general, the user evaluation of the web-site has been quite negative as most of the scores fell in the negative half of the scale. Potency is the weakest dimension. Furthermore this graph highlights that users with a positive evaluation of the web site focus their attention to that particular class (the Evaluation class), which deals mainly with strong feelings about a web site (adjectives like: good or bad, nice or ugly, etc.).

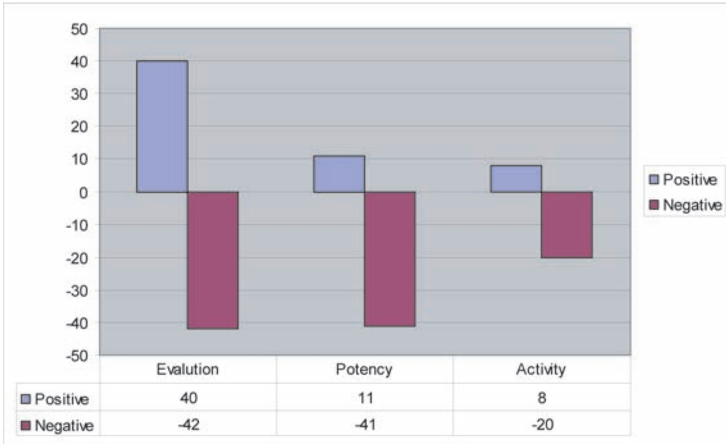
CONCLUSION

This chapter presented a UX evaluation approach consisting of three steps:: 1) select a collaborative tagging system containing the content for which we would like to evaluate the users' experience

Table 3. Classes assigned after the clustering phases and number of tags falling in their relative class; the third rows represent the number of tags per class with positive and negative meanings, (+n) stands for n tags in that category with positive meaning and (-n) for the negatives.

Evaluation	Potency	Activity
82 (+40) (-42)	52 (+11)(-41)	28 (+8) (-20)

Figure 3. A bar-chart view of the overall evaluation, potency and activity



(Del.icio.us for web sites, Flickr and YouTube for multimedia content, Technorati for blogs); 2) employ an IR technique to extract semantics for users' tags and group them together by cluster of synonyms or related tags; 3) use the data automatically extracted from the clusters of tags to detect the overall impression of users (UX) over the selected content (web sites, multimedia systems, blogs, etc.) by grouping tags according to the semantic differential technique. We reported an example of how this evaluation approach can be applied on an ad-hoc collaborative tagging system. Anyway, this procedure can be applied to a wide range of collaborative tagging systems. In our test we asked users to add adjective tags to keep it controlled but generally we will have plenty of available tags already inserted by users of collaborative tagging systems. What do you mean by that? You need to explain further by rewriting this sentence.

Our approach suggests the importance of collaborative tagging systems in the evaluation of the end users experience. It seems to be a promising and cost effective alternative to questionnaires or interviews.

Collaborative tagging systems are becoming increasingly popular on the Internet. There are many reasons why users are motivated in volunteering their time to support these on-line communities (Clary et al., 1998), adding information to collaborative tagging systems over the web; nevertheless such systems keep growing as a social phenomena. We can take advantage of this huge number of users to detect the user experience perceived by them when adding tags for categorizing a content of interest over the web.

REFERENCES

Bennis, W., & Ward Biederman, P. (1998). None of us is as smart as all of us. *IEEE Computer, 31*(3), 116–117.

Clary, E. G., Snyder, M., Ridge, R. D., Copeland, J., Stukas, A. A., Haugen, J., & Miene, P. (1998). Understanding and assessing the motivation of volunteers: A functional approach. *Journal of Personality and Social Psychology, 74*(6), 1516–1530. doi:10.1037/0022-3514.74.6.1516

Coleman, W. D., & Williges, R. C. (1985). Collecting detailed user evaluations of software interfaces. *Proc. Human Factors Society: Twenty-Ninth Annual Meeting*, Santa Monica, CA, (pp. 204-244).

Csikszentmihalyi, M. (1998). Creativity: Flow and the psychology of discovery and invention. *Perennial*, June 1997.

DeAngeli, A., Sutcliffe, A., & Hartmann, J. (2006). *Interaction, usability and aesthetics: What influences users‘ preferences? DIS 2006 Conference Proceedings, ACM*, (pp. 271-280).

Dix, A., Levialdi, S., & Malizia, A. (2006). Semantic halo for collaboration tagging systems. In S. Weibelzahl & A. Cristea (Eds.), *Proceedings of Workshop held at the Fourth International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2006). Workshop on Social Navigation and Community-Based Adaptation Technologies* (pp. 514-521), Lecture Notes in Learning and Teaching, Dublin: National College of Ireland.

Fischer, G. (2006). *Distributed intelligence: extending the power of the unaided, individual human mind.* In C. Augusto (Ed.), AVI (pp. 7–14). ACM Press.

Garrett, J. J. (2003). *The elements of the user experience: User centred design for the web*. London: Easy Riders.

Golder, S., & Huberman, B. A. (2006). Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, *32*(2), 198–208. doi:10.1177/0165551506062337

Hartmann, J., Sutcliffe, A., & De Angeli, A. (2007*).* Assessing the Attractiveness of Interactive Systems. *CHI 2007 Conference Proceedings* (pp. 387-396), San Jose, CA, USA

Hassenzahl, M., & Tractinsky, N. (2006). *U*ser experience - a research agenda. *Behaviour & Information Technology*, *25*, 91–97. doi:10.1080/01449290500330331

Kraft, R., Chang, C. C., Maghoul, F., & Kumar, R. (2006). Searching with context. In *Proceedings of the 15th international Conference on World Wide Web* (pp. 477-486) (Edinburgh, Scotland, May 23 - 26, 2006). WWW ‘06. ACM, New York, NY.

Krikos, V., Stamou, S., Kokosis, P., Ntoulas, A., & Christodoulakis, D. (2005). DirectoryRank: ordering pages in web directories. In *Proceedings of the 7th Annual ACM international Workshop on Web information and Data Management* (pp. 17-22) (Bremen, Germany, November 04 - 04, 2005). WIDM ‘05. ACM, New York, NY.

Mathes, A. (2004). *Folksonomies cooperative classification and communication through shared metadata.* Retrieved October 12, 2006, from http://www.adammathes.com/academic/computer-mediated communication/folksonomies.html.

McCarthy, J., & Wright, P. (2005). *Technology as Experience*. Cambridge: MIT Press.

Newman, N. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics*, *46*, 323–351. doi:10.1080/00107510500052444

Norman, D. (2004). *Emotional Design: Why we love or hate everyday things*. New York: Basic Books.

Nov, O. (2007, Nov.). What motivates Wikipedians? *Communications of the ACM*, *50*(11), 60–64. doi:10.1145/1297797.1297798

Osgood, H. M., William, C. E., & Murray, S. M. (1975). *Cross-cultural universals of affective meaning*. University of Illinois Press, Urbana.

Turney, P. (2002). *Mining the Web for Lexical Knoledge to Improve Keyphase extraction:Learning from Labeled and Unlabeled Data*, NRC/ERB-1096, July 19, NRC 44947.

Van der Wal, T. (2005). *Explaining and showing broad and narrow folksonomies*, Personal InfoCloud: February 2005 Archives. Retrieved November 22, 2007, from http://www.personalinfocloud.com/2005/02/.

Xu, Z., Fu, Y., Mao, J., & Su, D. (2006). Towards the semantic web: Collaborative tag suggestions. *In proceedings of the Collaborative Web Tagging Workshop at WWW2006*, Edinburgh, Scotland, May, 2006.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Cambridge MA: Addison-Wesley

KEY TERMS

Collaborative Tagging Systems: Collaborative tagging (also know as folksonomy, social classification, social indexing and other names) is the practice and method of collaboratively creating and managing tags to annotate and categorize content[5].

Distributed Intelligence: In many traditional approaches, human cognition has been seen as existing solely "inside" a person's head, and studies on cognition have often disregarded the physical and social surroundings in which cognition takes place. Distributed intelligence provides an effective theoretical framework for understanding what humans can achieve and how artifacts, tools, and socio-technical environments can be designed and evaluated to empower human beings and to change tasks[6].

Information Retrieval: Information retrieval (IR) is the science of searching for information in documents, searching for documents themselves, searching for metadata which describe documents, or searching within databases, whether relational stand-alone databases or hypertextually-networked databases such as the World Wide Web[7].

Semantic Clustering: Identifying and disambiguating between the senses of a semantically ambiguous word, without being given any prior information about these senses[8].

Semantics Differential: A type of a rating scale designed to measure the connotative meaning of objects, events, and concepts[9].

Usability Evaluation: Usability usually refers to the elegance and clarity with which the interaction with a computer program or a web site is designed[10].

User Experience: User experience, often abbreviated UX, is a term used to describe the overall experience and satisfaction a user has when using a product or system[11].

ENDNOTES

1 Searching a body of information for objects that match a search query, particularly a text or other unstructured forms (http://www.cs.cornell.edu/wya/DigLib/MS1999/glossary.html).

2 Round parenthesis are used in the mathematical sense that we are not enumerating a set here but we consider an ordered sequence in the case of Age feature.

3 Basic level variations are consider to occur when, having two words differing by the case or including or not a dash.

4 One user has been deleted from the sample because inserted tags as spam, due to the anonymous login to the system.

5 Collaborative Tagging Systems. (2007, Nov. 20). In *Wikipedia, The Free Encyclopedia*. Retrieved Nov 20, 2007, from http://en.wikipedia.org/wiki/.

6 Fischer, G. (2006). *Distributed intelligence: extending the power of the unaided, individual human mind*., AVI (Augusto Celentano, ed.), ACM Press, 2006, pp. 7–14.

7 Information Retrieval. (2007, Nov. 20). In *Wikipedia, The Free Encyclopedia*. Retrieved

Nov 20, 2007, from http://en.wikipedia.org/wiki/.

8 Semantic Clustering (2007, Nov. 20). In *Wikipedia, The Free Encyclopedia*. Retrieved Nov 20, 2007, from http://en.wikipedia.org/wiki/.

9 Semantics Differential. (2007, Nov. 20). In *Wikipedia, The Free Encyclopedia*. Retrieved Nov 20, 2007, from http://en.wikipedia.org/wiki/.

10 Usability Evaluation. (2007, Nov. 20). In *Wikipedia, The Free Encyclopedia*. Retrieved Nov 20, 2007, from http://en.wikipedia.org/wiki/.

11 User Experience. (2007, Nov. 20). In *Wikipedia, The Free Encyclopedia*. Retrieved Nov 20, 2007, from http://en.wikipedia.org/wiki/.

This work was previously published in Handbook of Research on Social Software and Developing Community Ontologies, edited by S. Hatzipanagos; S. Warburton, pp. 119-132, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.12

Identifying Users Stereotypes for Dynamic Web Pages Customization

Sandro José Rigo
Universidade Federal do Rio Grande do Sul (UFRGS), Brazil

José Palazzo M. de Oliveira
Universidade Federal do Rio Grande do Sul (UFRGS), Brazil

Leandro Krug Wives
Universidade Federal do Rio Grande do Sul (UFRGS), Brazil

ABSTRACT

Adaptive Hypermedia is an effective approach to automatic personalization that overcomes the difficulties and deficiencies of traditional Web systems in delivering the appropriate content to users. One important issue regarding Adaptive Hypermedia systems is the construction and maintenance of the user profile. Another important concern is the use of Semantic Web resources to describe Web applications and to implement adaptation mechanisms. Web Usage Mining, in this context, allows the generation of Websites access patterns. This chapter describes the possibilities of integration of these usage patterns with semantic knowledge obtained from domain ontologies. Thus, it is possible to identify users' stereotypes for dynamic Web pages customization. This integration of semantic knowledge can provide personalization systems with better adaptation strategies.

DOI: 10.4018/978-1-59904-974-8.ch014

INTRODUCTION

With the enormous quantity of documents that are now available on the Web, accessing and collecting the desired and relevant data has become a difficult task that produces low quality results. The Websites adaptation allows the minimization of this problem as an adaptive application generates Website content or the structure in accordance with a class of users. In fact, the personalization aspects are a critical factor for the successful user experience. As a personalization example, it is common now to find several customization options in an increasing number of Websites. The reasons for this are due to the diversity of users and its experience, intents, needs, preferences and even available equipment and software. The design of a Website with thousands of daily visitors will face hard time to fulfill these very different expectations. The personalization resources available can help users to have a more personal interaction, by observing their needs and preferences.

There are many different definitions for "Web personalization" in the literature. In a more general sense, it is considered as a set of actions that fine-tune the results of some user interaction, regarding this user or a set of similar users (Mobasher, 2005). The practical personalization depends on the context. For an e-Commerce Website it may be related to the set of products that are shown to the user each time he logs on, but for other applications it may refer to the interface organization, the navigational structure and content options. All approaches have their specific problems and some of them are hybrid, combining their better techniques (Middleton, 2004; Kleinberg 2004). It is important to notice that a superior result for the personalization requires not only an efficient approach to the analysis of the contents or users behaviors, but it is also dependent on the Website life cycle. The personalization application should be integrated with tasks such as content management, users profile management, adaptation strategies and interface generation. These tasks are well known in Adaptive Hypermedia initiatives.

Adaptive Hypermedia (Brusilovsky, 2004, 2001; De Bra, 1999) has as its objective the establishment of better user experiences by adapting hyper-documents and hypermedia to the users' needs, preferences and goals. Usability improvement is achieved with the construction of models that represent the users' objectives, preferences, previous knowledge, and skills. The use of these models, together with some complementary information as context, usage records or adaptation rules, allows the identification of possible topics of interest, restrictions and personalization options. In addition, domain information is very important in this process and drives the adaptation choices. This can be with respect to different aspects of a Website, such as its content or structure. Briefly put, this adaptation is based on the relationship between information concerning the application domain and information regarding the user profile.

One important topic in Adaptive Hypermedia systems research is the generation and maintenance of the users' profiles. Some approaches create the user profile from data obtained at the registration process, others incorporate the results of interviews and some perform automatic acquisition of information tracking the resources usage. In general, the profile based on the user identification tends to generate information valid over long periods. In some circumstances, short-term information can also be very useful and this kind of profile relies almost exclusively on the user interaction.

Web Usage Mining originates in prior Data Mining research with the purpose of automatic or semi-automatic discovery of Websites users' access patterns to generate information to be used by recommendation systems or by personalization systems (Mobasher, 2005). Analyzing the approaches to the generation of users' profiles by Web Usage Mining, a general pattern is identified and involves several stages (Markelou, 2005; Woon, 2005) that are briefly cited. The first is the acquisition of usage data. The second stage is dedicated to the pre-processing of data and the identification of access sessions amongst other necessary adjustments due to the Web environment (proxy servers, cookies or access errors, for example). At the end of the second stage, data is organized in appropriate formats for patterns mining where association rules and clusters can be generated or frequent pathways indicated. The third and last stage deals with the analysis and handling of these patterns in specific applications or contexts.

From the analysis of the obtained patterns, clusters identified or validated association rules, it is possible to generate complementary information to support the adaptation stage of Adaptive

Hypermedia systems. Nevertheless, it should be clarified that the patterns are obtained mainly with the access information present in the data for each user session. As mentioned, Web Usage Mining makes possible the capture and analysis of the behavior characteristics of Website users allowing the use of mechanisms directed toward personalization and adaptation (Aldenderfer, 1984; Brusilovsky, 2004). Despite being appropriate, this approach can be extended and improved by the use of semantic information associated with content access and navigation information.

An Adaptive Hypermedia application can discover better personalization choices by relating the semantic knowledge of a domain application, such as structural relations, with the usage information, such as navigation patterns. Some complex objects and some specific relation will not be treated with content-based or user-based techniques, as they have no representation in these systems. The representation of such complex objects and relations is possible with a domain ontology, which provides constructs for concepts and its relations definition. In a Website case, ontology provides the content concepts description, the hierarchies between them and the representation of some other existing relations. Domain ontology can be constructed by experts, manually. It can also be accomplished by using Machine Learning, Web Mining and Natural Language Processing techniques. The nature of the application can suggest the more suitable technique or combination.

The use of semantic knowledge, along with usage information, can lead to better knowledge discovery, by treating relations not applied in the other techniques. In general, this improvement takes two forms, which are the use of the semantic information in the pre-processing stage, enriching the pattern generation, or the use of the semantic information in the last stage, in combination with the adaptation itself (De Bra, 2004).

This chapter describes some possibilities for the acquisition of user profiles based on Web Usage Mining and domain ontologies. The main objective is to present the integration of semantic information obtained through the Website domain ontology with usage information obtained from the data gathered from user sessions. In addition, there is the intention to bring information to discuss if it is possible to identify more precisely the interests and needs of a typical user with these resources. The following sections provide some important background information on the Web personalization and Web Mining, ontology construction and semantic integration possibilities. The requirements for the semantic knowledge and usage information integration are discussed and finally some aspects of an Adaptive Hypermedia application based on the concepts of semantic application modeling are presented.

BACKGROUND

This section describes some concepts related to the main topic of this chapter. It provides the necessary background to the analysis of related works presented in the next section and to understand the integration approach described.

Web Personalization

The definition of Web personalization is found in the literature with some variations. In a general form, it is considered as a set of actions that adjust the results of user interaction, regarding this user or a set of users (Mobasher, 2000). In some cases, like an online bookstore, it corresponds to products indications. The scope for this kind of personalization can be very broad, ranging from items such as books or music to stocks, computers or cars. Another personalization example is the flexible organization of the user interface or the selection of contents. For example, the layout elements can be presented with more textual information, with differences in graphic elements, the options displayed in menus or the hyperlinks may be organized in more adequate manner for some user and the content may be more concise or have more details.

The approaches for the personalization can be grouped in content-based, collaborative-filtering or in some hybrid forms with both characteristics. Its differences rely on the strategies and information used to the personalization options generation.

In the content-based approach, the users' personal profiles represent mainly their interests. The Websites content is classified with respect to some subjects. Some metrics to evaluate the subject proximity to the users' interests are then applied for the generation of personalization. This approach can be found in several works, with small variations, but the principal aspect is the adaptation of a Website based on the preferences of the user (Lieberman, 1995; Mladenic, 1999; Mikroyannidis, 2004). Some advantages are found in this approach in situations as Web Information Retrieval, allowing the filtering of a large amount of pages based on the user profile. For example, a Website about movies can personalize the user navigation based on the previous shown interests and the options in the movies database. While this advantage can be relevant, it may also characterize some ineffective situations, when the user has a new interest, in an area not yet described in the profile. Another problem is that some useful semantic relations cannot be applied, as in cases of a more specific or more general approach for the same topic or in cases of different objects used in the same process and thus related.

The collaborative filtering techniques do not perform analysis in the content, but instead they focus on the preferences or activities associated with a specific user. These are then compared with all the other users and can lead to the identification of a set of users with common interests and preferences. There are several options to identify these relations, as the access to similar Web pages, the purchase of related items, the choice of similar options, the selection of similar feedback in ratings options and so on. Once the set of users with similar interests is defined, the personalization can be carried out by observing items not purchased or pages not accessed by a specific user, for instance. As the set of users is found to have the same preferences, it is assumed that one individual in this set can be interested in the same operations performed by the others. This technique, also known as user-based, can present problems in some situations, as in the publication of a new page or the release of a new product. Since the users do not have the necessary time to access, the item cannot be associated with some personalization action (Konstan, 1997; Balabanovic, 1997; Sugiyama 2004).

Some approaches using both techniques are known (Middleton, 2001; Kleinberg, 2004) and can be found as a way of reducing the limitations of each one. With both content and user preferences information the personalization system can be adapted to perform its tasks in a more efficient way.

Web Mining

Web Mining is defined as the discovery and analysis of useful information on the Web, with the objective of identifying behavior, characteristics, trends and navigation patterns (Cook, 2000; Kosala, 2000). There are three main areas of interest in Web Mining, described as Web Content Mining, Web Structure Mining and Web Usage Mining (Zaiane, 2000; Mobasher, 2005). Each of these is associated with some specific data collection originated in the records of Web Server activity, in the Website structure or in its content.

Web Content Mining is the process of extracting useful information from the content of Web documents. The Web content can be unstructured (plain text documents), structured (when dynamic pages exhibits content from databases) or semi-structured (HTML documents). The results can help information retrieval operations and personalization systems (Popov, 2003; LeGrand, 2002; Alani, 2003; Loh, 2000). The main advantage of this approach is the possibility to discover and classify documents and Web pages with respect to their content.

The Web Structure mining is the process of knowledge discovery driven by the Websites links structure. The topology of a site, its organization and the link structure are used to identify patterns. Some useful information extracted from these patterns can be, for instance, the identification of pages that represents collection of specific information, or collections of general information, with a large number of references, like in the concept of hubs and authorities. There are some examples of algorithms such as HITS (Kleinberg, 1999) and PageRank (Brin, 1998), that are based on this kind of data.

The objective of Web Usage Mining is to identify browsing patterns. This is achieved by analyzing the navigational behavior of a group of users. The information necessary to do this is available mainly in Web server log files. Web Usage Mining is carried out in well-defined stages, already mentioned in the text. These are the acquisition of usage data, pre-processing, analysis and usage. In the following sections, these stages are briefly described and discussed.

The first phase involved in the process of Web Usage Mining is concerned with the usage data processing. The processed data is extracted from the Web Server access log files or is generated from a script code included in the Web pages. One of the advantages of both forms of data collection is that they allow a Data Mining approach to the generation of user models for a specific Website, given the ability to obtain the data that is automatically generated when the pages of a Web site are accessed.

These access log files, created by Web Server software, were originally meant to aid debugging and to perform some simple statistics operations (Kohavi, 2001). The Common Log Format (Nielsen, 1995) is widely used, despite the existence of some improvements in other similar formats, like in the Extended Common Log Format and in other proprietary options. The Common Log Format is structured in text documents where each line represents a request or part of one. The main fields are the remote host identification, the remote user identification and login name, the date and time of the request, the exact request line received from the client, a code which indicates whether or not the file was successfully retrieved and the number of bytes actually transferred. The Extended Common Log Format adds two fields, the referrer and the user agent. The first indicates the URL accessed by the client browser before the request. The second indicates the browser software used in the request. In these formats, when some information is not available, it is replaced in the log with a minus sign ('-'). A few lines of a typical log file are shown in Figure 1.

Some problems can be observed. The verbose structure of these log files tend to be very expensive to process, because each single transaction made by the Web Server is stored in the file. Some of these are not relevant for the mining activity, for example, the retrieval of an individual image file, CSS (Cascading Style Sheet) or a script file. There is also a difficulty in the user identification, which requires some extra processing. Since the

Figure 1. Extended Common Log Format example

```
66.249.64.47 - - [13/Feb/2005:04:15:13 -0200] "GET /cursos/intercambios/apresentacao/corpo.htm HTTP/1.0" 304 - "-"
"Googlebot/2.1 (+http://www.google.com/bot.html)"
10.21.213.93 - - [13/Feb/2005:04:15:20 -0200] "GET /_imagens/capa/banners/ban_extravest.jpg HTTP/1.1" 200 2968
"https://www1.unisinos.br/" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.2; .NET CLR 1.1.4322)"
```

user identification and its sessions are important information, the mining systems need to use some heuristics. The date and time, along with the remote host identification are applied to separate the session log information.

To overcome some limitations observed in the log files processing, there is another approach that uses some specific script code, embedded in the pages displayed to the users. This process allows the recording of the real actions performed, without cache or proxy problems, and in real time, while the log file approach involves an offline processing step (Peterson, 2005). The preprocessing phase is simplified with these systems, because the transaction data is already stored in an appropriate form. However, this is achieved with some additional costs in each page view. These costs are due to the scripts embedded in the pages. Some are executed when the page is generated, while others are executed at the page restitution by the Web browser, by client-side script languages such as JavaScript (Netscape, 1998).

The tasks involved in this first preprocessing phase starts with data cleaning, when the log file is examined and some irrelevant entries are removed. Examples are the entries related to software robots doing crawling activities, or related to the structure of the pages as in the case of frames utilization. The next task is the user and session identification. Since the log files can be employed without user identification, there exists the necessity of processing the records to identify users, in general with information such as IP (Internet Protocol) numbers, date and time of the access. When some user, an individual accessing the site not a specific person, is identified, there is also the necessity of session identification, because it can be of importance for some mining process to identify different sessions of a user. There are also peculiar situations to deal with, as in the case of the use of cache mechanisms in the client software (the Web navigator), in the occurrence of errors or in the case of missing stored information in cookies.

After the processing and adjustments in the original data available, different techniques are applied in the knowledge discovery, such as sequential pattern mining, association rule mining, clustering and classification. The most frequent are the association rule and the sequential pattern mining. The first relates items and are used to identify groups of pages visited in similar ways by users. The second allows the most employed sequences of pages to be found. The other techniques are applied to group users into similar interest sets and to identify users in predefined interests groups.

Each technique mentioned can present results that are adequate to different tasks in the final phase, the adaptation or personalization. More details are described in the following sections.

Semantic Web

The Semantic Web initiative has the objective of solving some deficiencies observed in the traditional Web and implementing some improvements to the present possibilities of automated Web content processing. The first task in this direction is the description of documents in a more structured manner, allowing software agents to automatically manipulate such documents. In less structured languages, such as HTML, it is not possible to automatically perform tasks that depend on the documents content. As stated by Berners-Lee et al. (2001), these structured documents along with domain ontologies and inferences mechanisms can overcome the actual weakness of the Web. Documents in the Web are easily accessed by humans, but are not available to some automatic use.

One of the main requirements is the use of URI (Universal Resource Identifier). Another is the use of Unicode (Unicode Consortium, 2006) for codification, in order to ensure platform interoperability. The proper separation between structure and content in the documents is achieved with the use of XML language (Freitas, 2003). If some metadata pattern, such as RDF (Resource

Description Framework) (Herman, 2007), is utilized, it is also possible to describe information about the document. This can be very useful and helps to annotate the semantics of documents with information that can be automatically processed enabling a large number of new applications. When this metadata are described in accordance with some user's community standard, they provide the means for a consistent terminology to be at the disposal of several applications. As an example, the "Dublin Core Metadata Initiative" (Dublin Core, 2007) can be cited. It was originated from an open organization engaged in the development of metadata standards that created the "Dublin Core Metadata Element Set" which is widely known and used in resources descriptions. In this metadata set, elements as "creator", "contributor", "coverage", "date", "subject" and others are described with a precise meaning. Its adoption by communities allows precise information exchange, by documents with metadata annotation.

For some operations, the minimal metadata annotation can be insufficient. Descriptions that are more effective can be done with ontologies, in which a set of concepts and relations belonging to a particular domain may be shared. In this case it is possible to achieve more effective treatment of the documents information, with diverse objectives, as Information Retrieval, Electronic Commerce, Distance Learning or Data Integration, to name a few possibilities (Hendler, 2002; Nilsson, 2003). There exist some specific languages for the ontology description, as the OWL (Ontology Web Language) (Herman, 2006) which have the objective of precise concepts and relations description. According to Heflin (2004), this language supports the ontology description and integration, along with inference and query operations. The ontology creation can be done manually, by an expert in the specific domain, or automatically, using some Machine Learning techniques (Fensel, 2001; Fensel, 2002). Some principles, already indicated by Grubber (1993), must be adopted, including coherence, clear description of terms or facilities for ontology extension. The ontology editor usually verifies some other principles, as its correction.

The possibilities for ontology application depend on the knowledge acquisition operation, which can be a difficult task. Some Websites have a large number of pages and the manual creation of an ontology that describes the Website can be unfeasible. For this situation, it is useful to apply automatic creation mechanisms.

Adaptive Hypermedia

Several Adaptive Hypermedia systems were developed by different research groups and address different application areas. The best known are those for education, information retrieval and tourism, library and museum support. Some systems have identification options that connect the users with the profile information. Others are driven by non-invasive techniques and try to get the information for the profile generation in an automatic way (Dolog, 2004). The large volume, the diversity in formats and the great rate of information generation and update makes it hard to treat manually in an adequate form to the different users. In addition, the great number of users and diversity in interests and preferences makes it difficult to generate an efficient and usable interface in a system without adaptive options (De Bra, 2004).

The research in the Adaptive Hypermedia field has the objective of improving the users' satisfaction while using these systems. This usability improvement is achieved by the construction of models that can represent the knowledge, skills, objectives and preferences of the users. Besides the user modeling, some specific techniques are observed in the interface construction and in the usage recording. Some complementary information, as the application context, usage data, adaptation rules, allows for the identification of possible topics of interest and useful adaptations (Brusilovsky, 1996; De Bra, 1999).

To make possible the adaptation, the documents should be related to the domain model and its concepts. These can be more general (broad concepts), can represent groups with topics about a general subject or can describe specific information about some topic. The possible relations will be at the system disposal and the contents can be related to the domain model (Wu, 2002).

The system's generic operation can use these descriptions, in different tasks. The first one is the recording of the users' behavior. In Internet systems, this behavior can be associated with sequences of page accesses, for instance. A second task is to apply some processing method over the user model to classify the content information, regarding its profile. The third one is to combine this information to generate the interface, according to the identified possibilities.

The user profile can be composed of information that is valid over long or short time periods. Normally data with a long period of validity requires the identification of the user to be correctly acquired. In the context of educational systems, it is desirable and even necessary that the system user be identified to allow the update of the profile of the accessed information. In other circumstances, however, this identification could be undesirable for the users, and even be unnecessary given that it is also possible to obtain good results when the adaptation of a Web site is based on a class of users, represented by stereotypes, rather than on specific users. The user profile can be generated based on knowledge or behavior. The knowledge-based approach, that tends to make use of static models, can apply tools as interviews, tests and questionnaires. The behavior-based approach employs data from the users' interaction. These data can be from different periods and are applied in order to extract useful patterns (Middleton, 2004; Kobsa, 1993).

The adaptation in Adaptive Hypermedia systems refers to the contents and its presentation form - information is presented with different details. In addition, the information can be shown in an interface with more text or more images, some specific color configuration or with auxiliary media, as sound, video or animations (Christopher, 2002). Also the new devices' capabilities requires that some specific information initially applied to only one context will be sent now to a diversity of devices, with varied capabilities in memory, display and processing power (Petrelli, 2005). The possibilities of integrating sensors as input to these systems also can be very useful because of the effective interest delimitation that it allows the generation and short cycle of recording, and the inference and adapting process (Zimmerman, 2005).

RELATED WORKS

The Web Mining process can be related to the discovery of knowledge in sources such as the content, the usage records or the structure of Websites. This knowledge can be analyzed and, if considered useful, applied in adaptation or personalization tasks. Some details about the information obtained with these sources can be helpful to identify limitations observed in traditional systems using Web Mining. These limitations are described and related to possible solutions, with the use of some complementary semantic knowledge.

When the content of the pages is treated like a bag of words, it is difficult to the mining process to identify a relation between different pages dealing with the same concept but described with synonyms or hyponyms. As an example, one page can present the term "exercises" and another page can use the term "learning activities" referring to the same concept. Another example is the case of pages with the terms "car" and "gear-box", which, in this approach, will have no relation (in fact, they are composite objects). These situations are treated in several different ways; in some research works, such as Loh et al (2000), concepts are used to describe the contents of documents (which may be Web pages). Concepts are higher-level abstractions that represent ideas, objects and events. They are described by a set of words or even by semantic networks containing synonyms, quasi-synonyms, lexical variations, plural, verb derivations, semantic related words, proper nouns, named entities and abbreviations, multi-words, lexical compounds or noun phrases. Each concept has only one set as a descriptor, but one term may be present in more than one descriptor set. Thus, associated to each term in a concept there must be a weight, describing the relative importance of the term in this concept.

Another limitation found in the traditional process is related with the usage data acquisition and treatment. In this case, the access patterns computed with data from the Web server logs or some other acquisition form can also have some important information for the mining process that is not correctly treated. Since the traditional systems work with the page view concepts, the common results are a cluster of pages or a frequent access pattern. However, the information is restricted to access only. None of the possible relations between the page views is taken into account. As stated before, some interesting conclusions can be obtained from the page views relations analysis. For example, a cluster of visited pages can be used to discover the relations between its pages.

The integration of semantic knowledge, as a way to overcome these limitations, can be found in works that are referred to "Semantic Web Mining", as described by Stumme (2002). The main objective is the integration of domain knowledge with the mining process. Web Usage Mining makes possible the capture and analysis of the behavior characteristics of Website users, for mechanisms directed toward personalization and adaptation (Koutri, 2004; Mobasher and Dai, 2005). This approach is improved with semantic information associated with the Web usage information. The semantic information can be used in the pre-processing stage, enriching the pattern generation, or in the adaptation process (Stume, 2002; Eirinaki, 2006).

The construction of models that are able to represent the knowledge, abilities, goals and preferences of users can be seen in Adaptive Hypermedia systems (Christopher, 2002; Petrelli, 2005; Wu, 2002). Many of these systems maintain identification interfaces and profile characterization of each user, while others use non-invasive techniques and aim to automatically obtain data for the generation and maintenance of the user model (Dolog, 2004). Morales (2006) describes an interesting approach aimed to acquire user models by a specific subsystem that applied semantic Web technologies. The system developed to model learners' behavior is closely coupled with a Web-based educational system and all the user actions are treated as events and related to content elements. This treatment has the objective to create beliefs about the learner that can be validated and stored in the model. In Cantador (2006) there is the description of a strategy to automatically cluster users' profiles based on an ontology that describes domain concepts. As these concepts are used, the system can generate several layers of clusters, each representing some group of users with particular interests.

Some interesting results may be achieved by the collaboration between systems. Since different systems can have partial information about the users, it is interesting to have some form of mediation or collaboration. The systems can benefit from enriching the stored User Model information. Some works are known in this field, as in Berkovsky (2006) and Musa (2005), which suggests the use of resources as Web-services and mediation approaches. Dolog (2004), Nejdl (2003) and Arroyo (2006) present a more detailed discussion about the possibilities of interoperability in personalization systems. The user model is implemented in diverse forms, using the semantic Web resources, as RDF for metadata description. In addition, this metadata annotation can be related to standards for learning modeling, such as PAPI

(IEEE LTSC, 2001) and IMS LIP (IMS, 2001). The application of semantic Web resources in the description of elements for the user models is proposed in Ounnas (2006), where some known standards as PAPI, IMS LIP and FOAF (FOAF, 2000) are studied and an extension to FOAF is proposed. From this extension, it is possible to relate information in these models.

In addition to user modeling there are known techniques for the construction of interfaces in a flexible manner and following the usage of these interfaces. These models and complementary information such as the application context, usage data covering user or user group interactions, and adaptation rules, amongst others, permit the identification of possible topics of interest, access restrictions and adaptations of content and format (DeBra, 2004; Brusilovsky, 2004). One example in this direction is the GLAM system (Jacquiot, 2006), that uses a layered model, in order to facilitate the adaptation. The main objective is to implement navigation adaptation, provided by means of actions selection.

The integration of usage information in combination with semantic information produces better results, as reported in the work of Mobasher (2002), where semantic information contained in the ontology of a Website, together with the usage data, is applied in the analysis and generation of clusters and association rules. Thus, the clusters and association rules generated allow correlating the relevant details of each section of the Website. The case study used is a film Website, and with this treatment, the user's choice of a page containing a film description can be associated to several possible actors of user interest, which would not be possible without the semantic description of the pages.

The usage information is also related with some structural data or annotation information. An example can be seen in the work described by Bateman (2006), where the annotation problem is addressed. This work suggests the collaborative annotation approach (CommonFolks) together with document annotation, in an e-learning context. The RDF and LOM patterns are applied. Another example is described in Bechofer (2006), related to conceptual browsing. Ontologies in OWL Language automatically relate hyperlinks from different Websites. This allows the use of the hyperlink structure of Websites in order to discover interesting relations. The metadata and reasoning components can dynamically relate resources.

The problem of content adaptation to users' preferences can also be treated with semantic knowledge, as demonstrated by Aroyo (2006). In this work, ontologies about a domain application allow inferring some specific relations associated with time or lexical relations. This facilitates the recommendations to the user, as it allows the conceptual navigation. Another example is the "Poncelet Project" (Habel, 2006) that applies an ontology describing the concepts in the educational material at the student's disposal. The ontology classifies the resources and relates it to concepts. Then it can provide multiple paths to different students. In addition, it may help in the administration of the resources.

Other works deal with the processing of specific characteristics involved in this process, such as the use of Description Logic techniques as an aid to the processing of semantic information (Esposito, 2004). There are also approaches to the more specific usage of semantic information (for example, similarity) in mining, as in the case of the *Semantically Similar Data Mining (SSDM)* algorithm (Vieira, 2005), for the mining of association rules taking into account synonym information. Mechanisms for the treatment of sets of data accessed by the user to construct a conceptual map with the objective of revealing their interests can also be observed (Zhong, 2006). There are methods for the identification of users based on the Web usage with the integration of this information with semantic information (Zhou, 2006; Jin, 2005).

In some works the employment of usage information integrated with semantic information is associated with the use of clustering techniques that take into account the set of concepts identified in a group of pages that has been previously reported as a commonly followed path with associative rules or as a cluster (Eirinaki, 2003; Esposito, 2004; Mobasher, 2005). In other cases, the ontology capabilities are used to provide semantic bridges between data resources, as RSS feeds. This can be seen in Conlan (2006), as domain ontologies are applied to personalize the exhibition of news items.

AN INTEGRATION APPROACH

The main purpose of this chapter consists in the description of the integration of Web usage information with semantic information. This integration makes it possible to obtain user classes that are associated with well-defined behavior, observed in the usage of Websites and, finally, to employ this information to generate adaptations without the need to identify specific users. An experiment that illustrates this approach is described. An open source Web Content Management system is used to implement the Web usage data acquisition and to generate the structure adaptations. The pre-processing and the Web Usage Mining steps were implemented independently, and a domain ontology provides the semantic description of the application. Details of the integration process are described below.

Web Usage Mining and Semantic Information Integration

The approach described here intends to be more complete, including not only the usage information adaptations. To accomplish this objective, the process also involves some semantic information regarding the Website structure and some complementary relations, such as content type, precedence and requisites. This information is maintained in a domain ontology, which is described below.

The usage information considered in this approach consists on the frequent sequential paths of the Website users. Considering a set $P = \{p_1, p_2, p_3, ..., p_n\}$ as the set of *n* pages in a Website, then the user access in a session allows the generation of a non empty set $L = \{l_1, l_2, l_3, ..., l_m\}$, where each l_i belong to P. A frequent sequential pattern is the set of repetitive accesses, observed some defined limits of occurrences. The identification of frequent sequential patterns in this work is implemented as an additional stage, using the algorithm known as Spade with the improvements described in the literature (Zaki, 2001; Leleu, 2003).

In the application used for the experiment, the published pages have a specific code that allows the recording of access information. This data is processed in a way that describes the path taken by each user in their visit to the Website (Oliveira, 2006). Figure 2 shows an example of the format and data applied in the access recording process. This format allows the recording of date, time and browser, the page URL, IP number and access parameter. The access origin is recorded in the “userid” element, generated by a cookie created at the first user access. The element “adapt” distinguishes the access between normal or suggested pages.

When the adaptation is based only in the usage information, these frequent sequential patterns are typically consulted at each user interaction in order to verify if the user path have some similarity with the considered patterns. In the case of a correct match, the system assumes that this user can have some interest for the subsequent pages in the pattern and they are suggested as structure adaptations. Figure 3 shows two different access patterns in a simplified Website structure: the first (a) is composed by the pages indicated by the set with continuous line. The second (b) is identified by the dashed-line set. When using the access information only, it is not possible to identify that the first represents a browsing over general

Figure 2. Usage data example


```
<acess>
<ip>201.37.126.43</ip>
<page>/cms01/index.php</page>
<parameter>34</parameter>
<agent>Mozilla/5.0 (...) Gecko/20050717 Firefox/1.0.6</agent>
<date>12/11/2006</date><time>13:04:23</time>
<userid>f4b3173f4a4efb248f6f200c5ce678e3</userid>
<adapt>0</adapt>
</acess>
```


topics and the second represents the navigation inside a specific topic. This information is very important since it allows us to improve the adaptation process.

The ontology allows the definition of concepts, relations and restrictions regarding some application domain, which can be more general or specific. In this work, a more specific approach was chosen. In this case, the ontology is also known as a "domain ontology" and one of its advantages is the possibility of having a more precise mapping of the important concepts and its relations, given the target domain. This choice demands the ontology to be reviewed and rewritten to each (new) application domain.

An experiment to allow the validation of this approach was developed in the educational field. The domain ontology created had as objective to describe relevant concepts to the educational field. The same approach can be applied to different areas, with specific ontologies. The ontology was manually constructed by application domain experts using the Protégé ontology editor (Protégé, 2007), with the OWL language. This representation form facilitates posterior manipulation. In this case, part of the information described in the ontology can be seen in Figure 4 and represents the content available on a Website with educational material. There it is possible to observe the relations between "topic" and "course". The relations "part_of" and "composed_of" indicates the type of the composition. The relations "has_requisite" or "is_requisite_of" indicate dependency between the topics of one course. The relation "contain" allows qualification of each component in the Website. The relation between the Website and the ontol-

Figure 3. Some access patterns and its interpretation

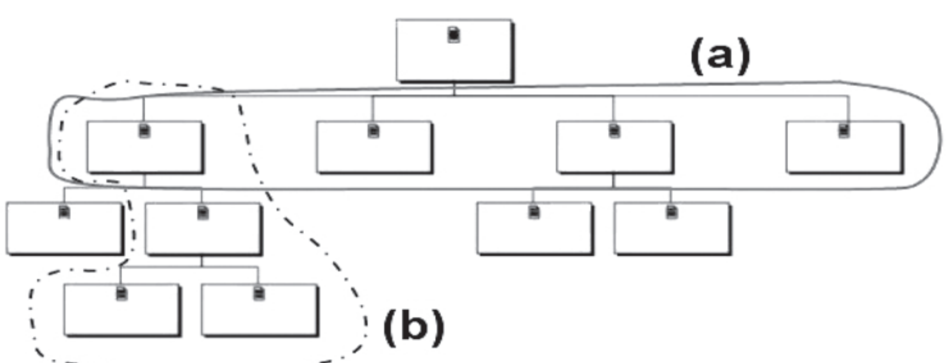
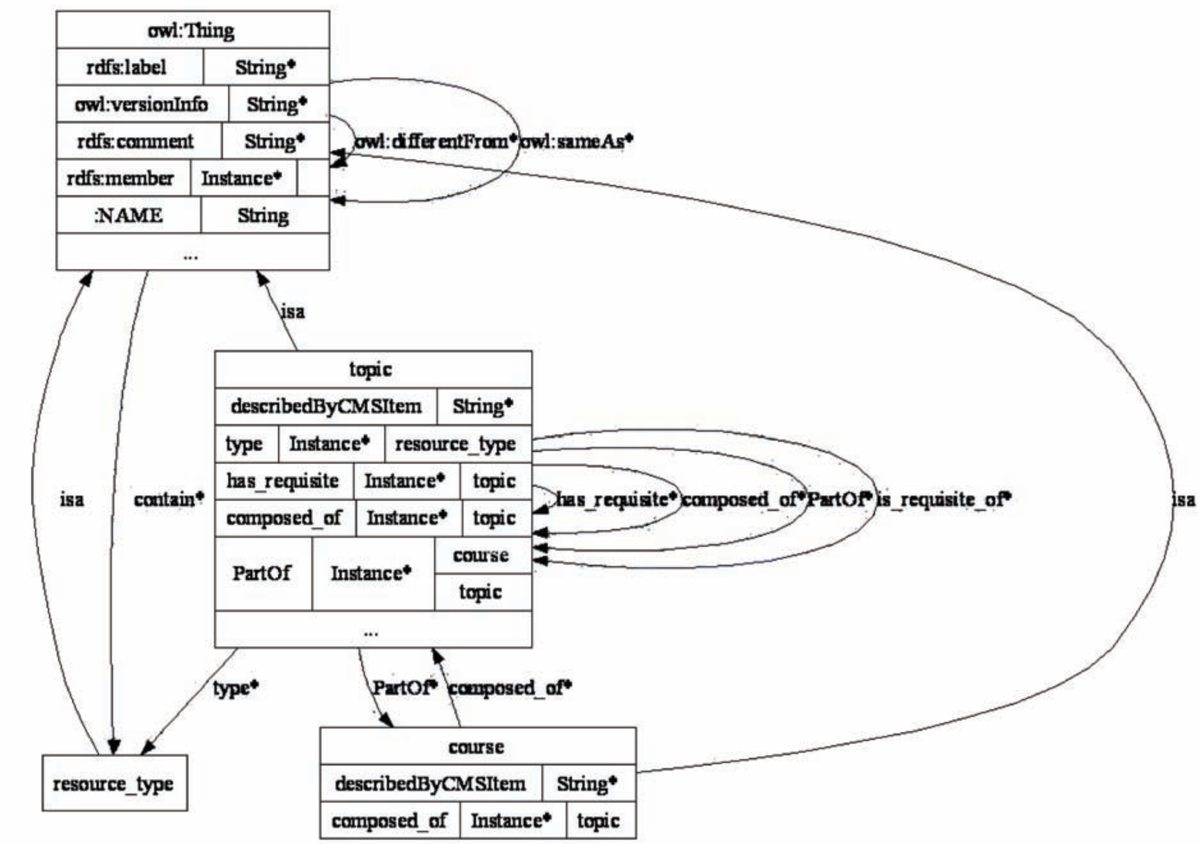


Figure 4. Part of the domain ontology used in the experiment



ogy is established by the semantic annotation of the Website elements. Each page of the Website is described as an ontology instance, along with the necessary relations. These instances are used in the integration process along with the usage information, as described below.

The instances in the ontology can be manipulated by inference mechanisms or by query languages, such as SPARQL language (Seaborne, 2007), which was the case in this work. The SPARQL language verifies the occurrence of interesting relations in the ontology instances. As an example, it is possible to identify, given a specific instance, all the relations associated to it. Also it is possible to recover, given a specific property, all the instances related to it. Finally, it is possible to discover all the relations that exist between two known instances.

To illustrate how the SPARQL language and the semantic annotation were used in this work, we will use Figure 5 that illustrates the description of one element of the application domain ontology.

Figure 5 shows part of the OWL representation of the instances in the ontology. The items "*T05_ACTIVITIES*" and "*T05_01_MENU*" are identified respectively as "*ID_24*" and "*ID_25*", based on the RDF ID element. This identification relates the elements to the corresponding pages in the Website. The "*composed_of*" relation defines the hierarchy between the items. The property "*describedbyCMSItem*" allows the semantic annotation of the contents, as they are stored in the Web Content Management System applied in the experiment. The relation "*part_of*" identifies the topic described as "ID_24" as part

Figure 5. Part of the domain ontology instances

```
<topic rdf:ID="ID_24">
  <part_of rdf:resource="#DATABASE"/>
  <composed_of>
    <topic rdf:ID="ID_25">
      <part_of rdf:resource="#ID_24"/>
      <describedbyCMSItem>24</ describedbyCMSItem >
      <rdfs:comment >T05_01_MENU</rdfs:comment>
     </topic>
  </composed_of>
  < describedbyCMSItem >24</ describedbyCMSItem >
  <rdfs:comment >T05_ACTIVITIES</rdfs:comment>
</topic>
```

of the course titled “Database”. This information can be accessed using the following SPARQL statements.

Figure 6 illustrates some possibilities for the identification of relations using the SPARQL language. The topic identifiers in the ontology match the browsing parameters used in the Website navigation. This allows the integration between the access information and the ontology information, in queries performed with the previously identified frequent sequential patterns. The first example (a) identifies all relations and instances associated with the two topics indicated (“ID_24” and “ID_25”). The second example (b) recovers all the relations between these two topics. The outcome of these queries allows the identification of the context that is not accessible from the usage patterns alone. This context is applied to identify users’ stereotypes, which are then associated to specific adaptation rules.

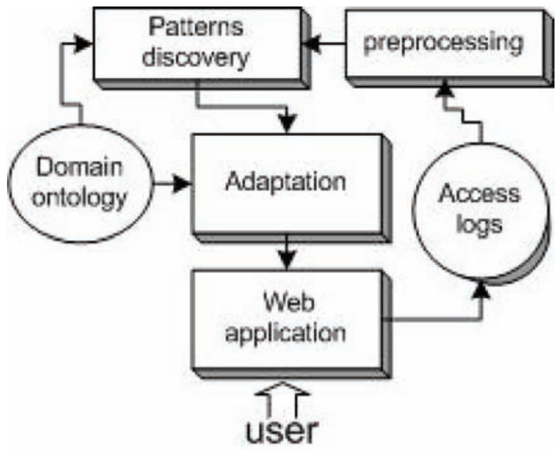
The result is the identification of more interesting patterns, related both with the usage and with the Website structure. By associating these patterns with specific rules, adaptations that are more expressive can be reached. Some examples of these results are the identification of users looking for specific topics, general view of the Website or complementary contents.

Figure 7 illustrates this process in a general form. The adaptation stage receives information from the domain ontology as well, in addition to the frequent sequential patterns already obtained from the processing of the access log data. With these two sources of information, it is possible not only to identify the page access sequences, but also to identify the concepts present in each page and their relationship. From the interaction with the

Figure 6. Part of the SPARQL queries

```
a) PREFIX v:<http://www..../.../...owl>
     SELECT ?x, ?y WHERE  (v:ID_24, ?x, ?y),
           (v:ID_25, ?x, ?y)

b) PREFIX v:<http://www..../.../...owl>
     SELECT ?x WHERE (v:ID_24, ?x, v:ID_25)
```

Figure 7. Integration of semantic information with Web Usage Mining



user, the Web application collects the usage data, storing them in the 'access logs' component.

The pre-processing integrates the usage information, making it available to the next stage, the patterns discovery. After their validation, these patterns are available for the adaptation component, which interacts with the Web application generating adaptations of the structure of the Web pages. This action takes into account the existence of a domain ontology for the application in question.

Structure and Content Adaptation

The information originated in the observed frequent sequential patterns is added to the original structure of the Website. Following the information already collected and processed, the system has a description of frequent sequential paths and specific relationships derived from the domain ontology. Based on the behavior observed from a user session the information is employed as complement to the original structure of the site and is published in specific areas of the interface. Thus, the pages accessed by the user are correlated by frequent sequential patterns. The identification that a certain number of pages accessed during a user session belong to a pattern can be assumed to indicate that the user in question is part of a group of users of the Website that share a specific content interest, and found this set of pages.

The identification of frequent sequential paths is applied in this work to minimize the need for specific rules for the generation of adaptation. In some systems, and particularly in those directed toward education, it is necessary to identify users and even to record the actions they make on the system. As consequence, in these cases it can be easy to specify the adaptation mechanism using rules (Paramythis, 2005). In the experiment that we have carried out, the identification of the users and their details is avoided, as was the use of this information to generate rules. An argument to justify this approach can be found in the behavior of users of Websites with domains different to those directed toward education, where in general there exists no desire to provide identification information.

The pre-processing of usage data and the generation of patterns is carried out periodically. During the user sessions, the system detects, from a recent access history that includes only accesses made in the current session, any coincidence of the observed behavior with the behavior patterns previously established and, if there exists such coincidence, carries out the associated adaptation.

In these cases, as covered in the experiment, the adaptation was composed of alterations applied to the structure of the resulting page with the addition of new navigation possibilities derived from the patterns that have been established. Content adaptation could be carried out by the addition of material related to the content being accessed on the basis of some relationship given in the ontological description. The semantic annotation of the content permits the identification of specific situations, from the identification of the type of complementary resources.

Experimental Results

Adaptive Hypermedia systems can pose several problems when the question is the performance and quality evaluation. The adaptations implemented can be associated with quality attributes and with efficiency parameters. In this case, since the main objective of this work is the identification of the useful Website adaptations, based on users' stereotypes related to a set of users, some tests were conducted in order to evaluate the quantity of generated adaptations and the quantity of accesses to these suggested adaptations.

The experiment in discussion was carried out over a period of six months during which the material was available for access, with the necessary information for the adaptation being generated.

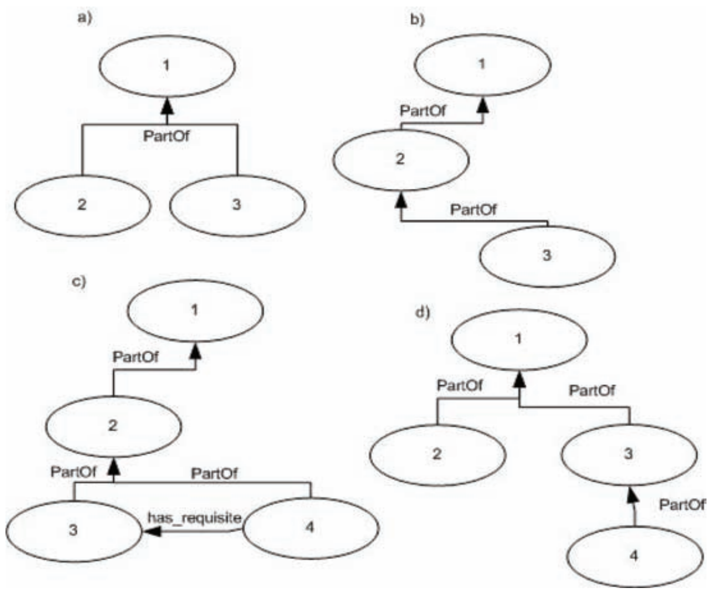
The results obtained indicate that some frequent patterns were related to specific behavior. One of these cases is associated to the overall navigation, where the user accesses the main topics available in the Website structure. This information is obtained from browsing frequent patterns that returns, when integrated with the semantic information described in the domain ontology, a relationship with an upper concept, usually the Website initial content. The kind of relation verified in the ontology in this case is mostly the "part_of". Another frequent case is the situation in which the items in the frequent pattern are related mostly with the "part of" relation, but in a way that an antecedent and subsequent item are associated. In this situation, the behavior detected is described as a navigation in which the user accesses the related and internal items of one specific topic.

Some of these situations are identified below and can be used as examples of the improvements obtained with the approach. The analysis of different frequent sequential patterns, with the same number of elements, allows the identification of different contexts. These contexts can only be distinguished when the domain ontology relations are used together with usage information. Some examples are summarized in Figure 8. In this figure, it is possible to identify elements that correspond to frequent patterns, with a number indicating their access order. The arcs between the elements indicate the ontology relations found for the items. Comparing the items "a" and "b", it can be seen that the first item ("a") is the representation of a browsing in the same level of the Website, the equivalent of a general view of the contents in this level. However, the second item ("b") indicates the access to more detailed information in one specific Website topic.

Based on this identified context, different adaptations procedures can be chosen. Another example of different contexts discovery can also be observed (items "c" and "d"). In the first case ("c"), the browsing started in a more general level and was directed to a more detailed level. In the second case ("d") there is more activity browsing in a general level, followed by the choice of a more detailed level.

The suggested adaptations are monitored and the access to these items can be compared with the normal Website items. In this case, the results indicate a useful set of adaptations generated. The proposed method can generate valuable information by relating the Web usage data and the semantic information. It is also possible to insert of new relations in the domain ontology, in a way that can be appropriated and effective to different application domains.

Figure 8. Semantic contexts obtained



CONCLUSION

This chapter presented an approach for the acquisition of user stereotypes based on Web Usage Mining and domain ontologies. In this case, the domain ontology describes important relations for the application. The relations in the ontology are combined with the usage patterns obtained by Web Mining techniques. The integration of semantic information with usage information was described and some related works were presented, in order to illustrate the possibilities of better identification of the interests and needs of a typical user with these resources.

The cognitive overload observed in the process of Internet users searching for information can be related to the difficulties of automatic acquisition of needed information. As the Internet standards were developed to human usage, there is the necessity of more adequate resources to the structured and more formal description of the documents and contents. In addition, there is the necessity for formal mechanisms to documents annotations. This allows the use of metadata and gives support to a great number of possible applications that would benefit from this information. The operations that become possible with a structured and properly identified (annotated) document are far more interesting than those possible with non-structured or un-annotated documents. Some difficulties in this case are observed in the annotation process, which can be manual and dependent on users' effort, or can be automatic and rely on Text Mining techniques. Resources such as ontologies and inference mechanisms can improve this scenario, with even better possibilities, as they can describe domain applications concepts and its relations. In this case, the knowledge described in ontologies can be associated with the existing documents for better manipulation.

These semantic resources are also used in a great number of experimental applications, but there are also diverse industrial initiatives. The Adaptive Hypermedia applications are among those that benefit from these resources. This can be observed in several aspects, as in the application descriptions, interface generation, user and context models construction or adaptation mechanism.

Some of the tasks of Adaptive Hypermedia applications are carried out with information generated from fields as Web Usage Mining. Specifically the user stereotype can be enriched by observing significant access patterns, related to the pages of a Website, obtained from the user navigation pattern. In this sense, these patterns can be treated as evidences of specific needs or goals, and then used to identify classes of users. It has also been shown that the use of a domain ontology, in which the pages can be associated to specific concepts or stages of repetitive processes on the Website, is more general than just the use of access information, without the related semantic information. That can justify the integration of Web Mining, ontology description and semantic integration possibilities. The requirements for the semantic knowledge and usage information integration are discussed and related with Adaptive Hypermedia application. There are even better possibilities for improvements when the application has a more formal description, such as an ontology, and associated to specific models that help, for example, in the identification of concepts associated to the application domain or to stages of routine tasks.

The handling of a user model and the direct identification of a specific user makes adequate results possible, because information concerning their interests, knowledge and goals are taken into account. Despite that, it is considered important to also evaluate the possibilities that arise from the data collection being uncoupled from a specific user. With this procedure, repetitive behavior is identified, which can also serve to identify goals and interests, with no specific user identification. In many application areas, it is not feasible or desired to obtain the user identification.

REFERENCES

Alani, H., Kim, S., Millard, D., & Weal, M. Hall, W. Lewis, P. & Shadbolt. (2003). *Automatic Extraction of Knowledge from Web Documents*. Workshop on Human Language Technology for the Semantic Web and Web Services, 2 Int. Semantic Web Conf. Sanibel Island, Florida, USA.

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster Analysis*. Beverly Hills, CA: Sage, 1984. 88 p.

Aroyo, L., Bellekens, P., Björkman, M., Broekstra, J., & Houben, G. J. (2006). *Ontology based personalization in User-adaptive systems.* In 2nd International Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces (WPRSIUI'06), Dublin, Ireland.

Aroyo, L., Dolog, P., Houben, G.-J., Kravcik, M., Naeve, A., Nilsson, M., & Wild, F. (2006). *Interoperability.* In Personalized Adaptive Learning. Educational Technology & Society, 9 (2), 4-18.

Balabanovic, M., & Shoham, Y. (1997, March). Content-based collaborative recommendation. *Communications of the ACM*, *40*(3), 66–72. doi:10.1145/245108.245124

Bateman, S., Brooks, C., & Mccalla, G. (2006). *Collaborative Tagging Approaches for Ontological Metadata in Adaptive E-Learning Systems*. Proceedings of the Fourth International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL 2006). pages 3-12.

Bechhofer, S., Yesilada, Y., Horan, B., & Goble, C. A. (2006). *Knowledge-Driven Hyperlinks: Linking in the Wild*. Sean Bechhofer. AH 2006, 1-10. 'http://dx.doi.org/10.1007/11768012_1', pages 1-10,2006.

Berkovsky, S., Kuflik, T., & Ricci, T. (2006). *Cross-Technique Mediation of User Models*, in proceedings of the AH Conference, 2006.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web . *Scientific American*, (May): 28–37.

Brin, S., & Page, L. (1998). *The anatomy of a large-scale hypertextual Web search engine*, Computer Networks, 30(1-7): 107-117, Proceedings of the 7th International World Wide Web Conference (WWW7).

Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction*, *6*(2/3), 87–129. doi:10.1007/BF00143964

Brusilovsky, P. (2001). Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, *11*, 87–110. doi:10.1023/A:1011143116306

Brusilovsky, P., Karagiannidis, C., & Sampson, D. (2004). Layered evaluation of adaptive learning systems. *International Journal of Continuing Engineering Education and Lifelong Learning*, *14*(4/5), 402–421. doi:10.1504/IJCEELL.2004.005729

Cantador, I., & Castells, P. (2006). *Multilayered Semantic Social Network Modeling by Ontology-Based User Profiles Clustering: Application to Collaborative Filtering*. 15th International Conference on Knowledge Engineering and Knowledge Management - Managing Knowledge in a World of Networks (EKAW 2006). Podebrady, Czech Republic, October 2006.

Christopher, D. (2002). *Staff: The hypercontext framework for adaptive Hypertext*. Conference on Hypertext and Hypermedia. Proceedings of the thirteenth ACM conference on Hypertext and hypermedia. Maryland, USA. Pages: 11 – 20.

Conlan, O., O'Keeffe, I., & Tallon, S. (2006), *Combining Adaptive Hypermedia Techniques and Ontology Reasoning to produce Dynamic Personalized News Services*, Proceedings of the Fourth International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2006), Dublin, Ireland (2006).

Consortium, I. M. S. (2001). *Learner Information Package Specification*. Retrieved september, 20, 2007, from http://www.imsglobal.org/profiles/index.html.

Cook, D. J., & Holder, L. B. (2000). *Graph-based data mining*. IEEE Intelligent Systems . *Los Alamitos*, *15*(2), 32–41.

De Bra, P. (1999). Adaptive hypermedia: From systems to framework. *ACM Computing Surveys*, *31*(4), 1999. doi:10.1145/345966.345996

De Bra, P., et al. (2003). *AHA! The adaptive hypermedia architecture*. Conference on Hypertext and Hypermedia. Proceedings of the fourteenth ACM conference on Hypertext and hypermedia. Nottingham, UK, Pages: 81 – 84. 2003. ISBN:1-58113-704-4.

De Bra, P., Arroyo, L. & Chepegin, V. (2004). *The next big thing: adaptive Web-based systems*. Journal of Digital Information, V(5)N(1).

Dolog, P. (2004). *Identifying relevant fragments of learner profile on the semantic Web*. In Proceedings of SWEL'2004 — Intl. Workshop on Semantic Web for eLearning, Intl. Semantic Web Conference 2004, Hiroshima.

Dolog, P., Henze, N., & Nejdl, W. (2004). *Personalization in distributed e-learning environments*. In Proceedings of WWW2004 --- The Thirteen International World Wide Web Conference, New York, May 2004. P.85-94. ACM Press.

Dublin Core. (2007). *The Dublin Core Metadata Initiative*. Retrieved september, 20, 2007, from http://dublincore.org/.

Eirinaki, M. (2003). *SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process* (2003). In Proc. of the 9th SIGKDD Conf., 2003.

Eirinaki, M., Mavroedis, D., Tsatsaronis, G., & Vazirginannis, M. (2006). *Introducing Semantics in Web Personalization: the role of ontologies*. M Ackermann et al. (Eds.): EWMF/KDO 2005, LNAI, pp.147-162,2006. Springer Verlag, Berlin Heidelberg 2006.

Esposito, F., & Lisi, A. (2004). *An ILP Approach to semantic Web mining.* In P.Buitelaar, et al (Eds.), Notes of the ECML/PKDD, 2004 Workshop on Knowledge Discovery and Ontologies, 139-144, Pisa, Italy.

Fensel, D. (2001). Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce. Springer-Verlag.

Fensel, D. (2002). Ontology-Based Knowledge Management. *IEEE Computer*, *35*(11), 56–59.

FOAF. (2000). *The Friend of A Friend Project*. Retrieved september, 20, 2007, from http://www.foaf-project.org/index.html.

Freitas, F. L. G. (2003). *Ontologias e a Web Semântica*. XXIII Congresso da Sociedade Brasileira de Computação. JAI. Campinas, São Paulo, Junho de 2003.

Gruber, T. (1993). *What is an Ontology?* Retrieved May, 01 2007, from http://www.ksl.stanford.edu/kst/what-is-an-ontology.html.

Habel, G. Reyes, M. L., Magnan, F. & Reyes, G. (2006). *General Poncelet meets the Semantic Web: A concrete example of the usage of ontologies to support creation and dissemination of eLearning contents.* Workshop on Applications of Semantic Web Technologies for e-Learning (SW-EL@AH'06), June 21-23 2006, Dublin, Ireland.

Heflin, J. (2004). *OWL Web Ontology Language Use Cases and Requirements*. Retrieved May, 01, 2007 from em http://www.w3.org/TR/Webont-req/, 2004.

Hendler, J., Berners-Lee, T., & Miller, E. (2002). *Integrating Applications on the semantic Web*. Journal of the institute of electrical Engenieers of japan, Vol 122(10), October, 2002, p. 676-680.

Herman, I. (2006). *Web Ontology Language*. Retrieved september, 20, 2007, from http://www.w3.org/2004/OWL/.

Herman, I. (2007). *Resource Description Framework*. Retrieved september, 20, 2007, from http://www.w3.org/RDF/.

IEEE LTSC. (2001). *Learning Technology — Public and Private Information*. Retrieved september, 20, 2007, from http://edutool.com/papi/drafts/08/IEEE_1484_02_02_D08_PAPI_rationale.doc.

Jacquiot, C. Bourda, Y., Popineau, F., Delteil, A. & Reynaud, C. (2006). *GLAM: A Generic Layered Adaptation Model for Adaptive Hypermedia Systems*. AH 2006: 131-140.

Jin, X., Zhou, Y., & Mobasher, B. (2005). *Task-Oriented Web User Modeling for Recommendation.* Proceedings of the 10th International Conference on User Modeling (UM'05) Edinburgh, Scotland, July 2005.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46*(5), 604–632. doi:10.1145/324133.324140

Kleinberg, J. M., & Sandler, M. (2004). *Using Mixture Models for Collaborative Filtering*. Proc. 36th ACM Symposium on Theory of Computing, 2004. STOC'04, June 13.15, 2004, Chicago, Illinois, USA. Copyright 2004 ACM 1581138520/04/0006.

Kobsa, A. (1993). *User Modeling: Recent Work, Prospects and Hazards*. M. Schneider-Hufschmidt, T. Kühme and U. Malinowski, eds. (1993): Adaptive User Interfaces: Principles and Practice. North-Holland, Amsterdam, 1993. Copyright © North-Holland.

Kohavi, R. (2001). *Mining e-commerce data: the good, the bad and the ugly*. Proceeding of the 7th ACM SIGKDD Inernational COnference on Knowledge Discovery and Data Mining, San Francisco, California, 8-13.

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., & Riedl, J. (1997). Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, *40*(3). doi:10.1145/245108.245126

Kosala, R., & Blockeel, H. (2000). *Web Mining Research: a survey*. Sigkdd explorations. ACM SIGKDD, Vol 2, Issue 1, July, 2000.

Koutri, M., Avouris, N., & Daskalaki, S. (2004). *A survey on Web usage mining techniques for Web-based adaptive hypermedia systems*. Adaptable and Adaptative Hypermedia Systems. Idea Inc. Hershey, 2004.

LeGrand, B., & Soto, M. (2002). *XML Topic Maps and Semantic Web Mining*. Semantic Web Mining Workshop, Conference ECML/PKDD 2001. Freiburg, Germany. January, 2002.

Leleu, M. (2003). GO-SPADE: Mining sequential patterns over datasets with consecutive repetitions [Springer Verlag Berlin Heildeberg.]. *LNAI*, *2743*, 293–306.

Lieberman, H. (1995). *Letizia: An Agent That Assists Web Browsing* (1995) Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95).

Loh, S., Wives, L., & Oliveira, J. P. M. (2000). Concept –based knowledge Discovery in texts extracted from the Web. *SigKDD Explorations*, *2*(1), 29–30. doi:10.1145/360402.360414

Markellou, P., Rigou, M., & Sirmakessis, S. (2005). *Mining for Web personalization*. In Web Mining: Applications and Techniques. Anthony Scime (ed.), Idea Group Publishing, p.27-49.

Middleton, S. DeRoure, D. & Shadbolt, N. (2001). *Capturing knowledge of user preferences: Ontologies in recommender systems.* In Proceedings of the ACM K-CAP'01, Victoria, Canada, 2001. ACM Press.

Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems*, *22*(1), 54–88. doi:10.1145/963770.963773

Mikroyannidis, A., & Theodoulidis, B. (2004). *A Theoretical Framework and an Implementation Architecture for Self Adaptive Web Sites*. In Proc. Of IEEE/WIC/ACM International Conference on Web Intelligence (WI'04). 2004. Beijing, China, p.558-561.

Mladenic, D. (1999*). Machine learning used by Personal WebWatcher. Proceedings of ACAI-99 Workshop on Machine Learning and Intelligent Agents*, Chania, Crete, July 5-16, 1999.

Mobascher, B. (2005). *Web Usage Mining and personalization*. Practical Handbook of Internet Computing Munindar P. Singh (ed.), CRC Press.

Mobascher, B., & Daí, H. (2005). *Integrating Semantic Knowledge with Web Usage Mining for Personalization*. In Web Mining: Applications and Techniques. Anthony Scime (ed.), Idea Group Publishing.

Mobasher, B., Cooley, R., & Srivastava, J. (2000, August). Automatic Personalization Based on Web Usage Mining. *Communications of the ACM*, *43*(Issue 8), ▪▪▪. doi:10.1145/345124.345169

Mobasher, P., & Dai, H. (2002). *Using Ontologies to Discover Domain-Level Web Usage Profiles*. Proceedings of the 2nd Workshop on Semantic Web Mining. Held at PKDD'02, Helsinki, Finland, August 2002.

Morales, R., Van Labeke, N., & Brna, P. (2006). *Towards a Learner Modelling Engine for the Semantic Web*, International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL). AH2006.

Musa, D. L., & Oliveira, J. P. M. (2005). Sharing Learner Information through a Web Services-based Learning Architecture. *Journal Of Web Engineering, Princeton, New Jersey*, *4*(3), 263–278.

Nejdl, W. Dolog, P. (2003). *Challenges and Benefits of the Semantic Web for User Modelling* In Proc. of AH2003 workshop, WWW 2003.

Netscape Communications Corporation. (1998). *JavaScript Guide*. Retrieved september, 20, 2007, from http://wp.netscape.com/eng/mozilla/3.0/handbook/javascript/.

Nielsen, H. F. (1995). *Logging in W3C httpd*. Retrieved september, 20, 2007, from http://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format.

Nilsson, M., Palmer, M., & Brase, J. (2003). *The LOM RDf binding-principles and implementation*. Technical report, Information system institute, University of hannover, germany, 2003. 3rd Annual Ariadne Conference, 20-21 November 2003, Leuven, Belgium.

Oliveira, J. P. M., & Rigo, S. J. (2006). *Mineração de uso em sites Web para a descoberta de classes de usuários*. In: CLEI 2006.Santiago, Chile, 19-25 October 2006.

Ounnas, A. (2006) Towards a Semantic Modeling of Learners for Social Networks. In *Proceedings of International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL) at the AH2006 Conference*, pp. 102-108, Dublin, Ireland.

Paramythis, A. Stephanidis, C. (2005). *A generic adaptation framework for Web-based hypermedia systems*. In Adaptable and adaptive Hypermedia Systems. Anthony Scime (ed.), Idea Group Publishing, p.80-103. 2005.

Peterson, E. T. (2005). *Web Site Measurement Hacks*. O'Reilly (2005), ISBN 0-596-00988-7.

Petrelli, D. (2005). User-Centred Design of Flexible Hypermedia for a Mobile Guide: Reflections on the HyperAudio Experience. [Springer.]. *User Modeling and User-Adapted Interaction*, *15*, 303–338. doi:10.1007/s11257-005-8816-1

Popov, B., et al. (2003). *Towards Semantic Web Information Extraction*. Second International Semantic Web Conference (ISWC-03). Sanibel Island, Florida, October 20, 2003.

Protégé. (2007). *The Protege Plataform*. Retrieved september, 20, 2007, from http://protege.stanford.edu.

Seaborne, A., & Prud'hommeaux, E. (2006). *SPARQL Query Language for RDF*. Retrieved september, 20, 2007, from http://www.w3.org/TR/rdf-sparql-query/.

Stumme, G., Berendt, B., & Hotho, A. (2002). *Usage Mining for and on the Semantic Web. Next Generation Data Mining.* Proc. NSF Workshop, Baltimore, Nov. 2002, 77-86.

Sugiyama, K., & Yoshikawa, K. H. (2004). *Adaptive Web Search Based on User Profile Constructed without Any Effort from Users* WWW2004, May 17–22, 2004, New York, New York, USA. ACM 1-58113-844-X/04/0005.

Unicode Consortium. (2006). *The Unicode Standard*. Retrieved september, 20, 2007, from http://www.unicode.org.

Vieira, T. P. (2005). *SSDM: a semantically similar data mining algorithm*. XX Simpósio Brasileiro de Banco de Dados - 2005 - Uberlândia, MG, Brasil.

Woon, Y., et al. (2005). *Web Usage mining: algorithms and results*. 2005 In Web Mining: Applications and Techniques. Anthony Scime (ed.), Idea Group Publishing, p.373-394.

Wu, H. (2002). *A reference architecture for adaptive hypermedia applications*. Eindhoven: Technische Universiteit Eindhoven, 2002. ISBN 90-386-0572-2.

Zaiane, O. R. (2000). *Web Mining: Concepts, Practices and Research*. In:SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, SBBD, 15., 2000,João Pessoa. Tutorial... João Pessoa: CEFET-PB; Porto Alegre: PUCRS,2000. p. 410-474.

Zaki, M. (2001). *SPADE: An eficient algorithm for mining frequent sequences*. Machine Learning, 42, 31-60, 2001, Kluver Academic Publishers. 2001.

Zhong, N., & Li, Y. (2006). Mining Ontology for Automatically Acquiring Web User Information Needs. *IEEE Transactions on Knowledge and Data Engineering*, *18*(4), 554–568. doi:10.1109/TKDE.2006.1599392

Zhou, Y., & Mobasher, B. (2006). *Web User Segmentation Based on a Mixture of Factor Analyzers*. Proceedings ECWeb'06. Krakow, Poland, September 2006.

Zimmermann, A. (2005). Personalization and Context Management. [Springer.]. *User Modeling and User-Adapted Interaction*, *15*, 275–302. doi:10.1007/s11257-005-1092-2

KEY TERMS

Adaptive Hypermedia: Approach to automatic personalization.

Domain Ontologies: Description of concepts and relations regarding some knowledge field.

Personalization: Process that adjust the results obtained by users when accessing Web systems.

Semantic Web: Set of resources intended to improve the actual possibilities of Web applications.

User Profile: Set of information regarding user preferences, necessities and knowledge.

Web Systems: Any application designed to be used on the Web.

Web Usage Mining: Set of techniques to generate patterns and discover knowledge from the web usage data.

This work was previously published in Handbook of Research on Web Log Analysis, edited by B. Jansen; A. Spink; I. Taksa, pp. 284-306, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.13
Improving Online Readability in a Web 2.0 Context

John Paul Loucky
Seinan Jogakuin University, Japan

ABSTRACT

This study describes a task-based assessment (TBA) approach to teaching reading and writing online. It then analyzes key factors emerging from the results of implementing this approach with graduate engineering students in Japan. It is argued that these factors should be considered when designing or assessing any online reading or writing course for ESL/EFL students. The findings of this study are especially relevant to task-based approaches and technical or pedagogical innovations which can help foster more effective and enjoyable learning for teachers and students in blended learning environments. It is hoped that this discussion can serve as a model of what can be done to enhance online EAP/ESP/ETP courses, as well as any other online reading or writing course being designed for speakers and readers of languages other than English. The goal in this chapter is to summarize research aimed at integrating some of the most useful Web sites for English language learning into a user-friendly system for optimal online vocabulary development — which could be self-monitored by students as well as tracked by teachers via a course management system.

DOI: 10.4018/978-1-60566-190-2.ch021

INTRODUCTION

The emergence of new types of electronic media such as blogs, wikis, mobile phones and social networking sites is having a profound effect on the way people communicate. This is especially true of written communication and therefore as a consequence also greatly affects the way people read and consume information. The high levels of familiarity that today's students exhibit vis-à-vis these technologies is set to have profound effects on the ways that foreign languages are taught in a Web 2.0 context. If students of English as a Foreign Language cannot comprehend the high level of vocabulary and technological jargon found online, messages will not be understood and learning will be impeded. Given the challenge presented by the new landscape of Web 2.0 communications, there are two main objectives in this chapter:

1. To examine how best to assess and improve the readability of any website or application. Furthermore to indicate how a series of critical linkages can be formed to better integrate listening, glossing and translation so as to empower learners to better comprehend any Internet application or location.
2. Using the Virtual Language Education Links Library, known as the World CALL Language Links Library (Loucky, 2008), this chapter aims to identify which kinds of language learning sites and Web 2.0 functions are most helpful to Japanese graduate students vis-à-vis improving their online English reading and vocabulary skills. This World CALL Directory (found at www.CALL4ALL.us) is a Free/Open Source Language Education Resource Repository. Its aim is to serve as a Virtual Encyclopedia of all major language learning links, Web dictionaries and Computer-Assisted Language Learning organizations in the world.

In an age where multi-literacy and foreign language literacy in multimedia environments are becoming increasingly more important, teachers need to be able to understand and use more CALL technologies for efficient reading and vocabulary assessment to produce effective and enjoyable language development. The need for quick, easy and reliable readability checking for English reading texts has recently become more pronounced. Whether assessing print, online fiction or nonfiction texts for either Extensive or Intensive Reading (ER or IR) use, language learners and teachers are in real need of helpful, user-friendly ways of assessing the reading levels of texts.

For over 50 years, readability formulas have been used to help guide students to books at their appropriate level of reading and interest. Briefly stated, it is very important for ESL/EFL teachers to be able to individually test their students to determine their actual reading instructional level. EFL teachers using extensive reading are still in a quandary about how to arrive at commonly understood reading levels for various publishers, who do not use a uniform system. Before deciding what reading methods or materials to use, English reading teachers need to realize that three different kinds of reading levels are most crucial to understand. These three categories of reading ability are 1) Frustration Level, to be avoided; 2) Instructional Level, which Intensive Reading and classwork may be done at; and 3) Independent Level, which is appropriate for Free or Extensive Reading. Table 1, adapted from Loucky (1996, p. 301) and (Ekwall, 1976, p. 267), illustrates what these different levels involve.

Three other types of reading levels should be considered and compared when trying to best match texts or books to students:

1. The readability level of a text or a book's grade level: a particular text's reading difficulty level has most commonly been measured by a given readability formula (e.g. Flesch Reading Ease, Flesch Kincaid Grade Level).
2. The interest and maturity level of a book or text's ideas and content (usually grouped by

Table 1. Reading level criteria

Reading Level	Word Recognition	Comprehension %
Free or Independent	98 - 99% or more	90% or more
Instructional	95% - 98/99%*	75 Ideal (51 - 89%)
Frustration	90% or less	Under 50%

Note. The asterisk indicates not more than 1/20 unknown words.

Lower Grades, Middle Grades and Upper Grades).
3. The individual reading and maturity level of each student.

Ekwall's (1976) classic reading education text listed approximately 50 standardized reading tests, all based on using such formulae for estimating American students' average independent reading level. A learner's independent reading level may be defined as the most difficult level of text s/he can comprehend alone without using a dictionary or another's help. Loucky (1994, 1996, 1997, 2003, 2006a) has used these kinds of tests to assess several thousand Japanese college students' English reading levels. Based on nearly two decades of research, it has been possible to identify consistent patterns that are useful for language teachers to know, especially those wanting to tailor the level of required or free readings to individually appropriate levels. Once each learner's independent reading level has been determined and compared to the average for particular grade levels, each student can be guided to materials that are at his or her appropriate instructional level. This is normally defined as 1-2 grades above their free or independent reading level. Frustration levels beyond that should be avoided at all costs. Today there are many proponents of stress free-reading, or fluent, independent reading, whereas instructional level reading may be reserved for practice of particular reading skills during Intensive Reading classes.

Palmer distinguished extensive from intensive reading (1968, p. 137). Intensive reading tends to teach reading as a set of component skills and usually refers to careful or close reading (or translation) of shorter, more difficult foreign language texts with the goal of deeper and more detailed understanding. Texts are studied intensively in order to introduce and practice reading skills that are distinct. By contrast, Bamford and Day (1997) characterize free or extensive reading as being:

generally associated with reading large amounts with the aim of getting an overall understanding of the material. Readers are more concerned with the meaning of the text than the meaning of individual words or sentences … Extensive reading as an approach to teaching reading may be thought of in terms of purpose or outcome … It can also be viewed as a teaching procedure, as when Stephen Krashen (1993) terms it free voluntary reading, or when teachers give students time for in-class Sustained Silent Reading (SSR) — a period of 20 minutes, for example, when students and teacher quietly and independently read self-selected material … No matter how sophisticated the teaching profession's understanding of and ability to teach the reading process, until students read in quantity, they will not become fluent readers. (n.p.)

The two main elements that determine a student's reading rate are the difficulty or readability of a text and the purpose of reading it. Readability formulas have generally been a combination of two factors: 1) a measure of a text's word difficulty level, and 2) a measure of a text's sentence complexity. Raygor and Raygor (1985) have graphed readability estimates showing that a text's grade level is based on both its total number of sentences and its number of long or difficult words. Readability formulas measure a text's word difficulty and sentence complexity as follows: "Word difficulty is measured by word length or frequency. … Sentence complexity or syntactical difficulty is usually measured using the average number of words in the sentences" (p. 192). Two other important principles deduced from readability studies are also evident:

1. Readability formulas use objective measurements to analyze text and predict which materials can be comprehended by individual readers as long as they are used to assess both text and learner appropriately.

2. Students generally show the most reading improvement if they regularly practice reading within a range of difficulty that is neither too challenging (known as the *frustration level*) nor too easy (their *independent reading level*). While all readability formulas are based on analyzing some aspects of a text or book's difficulty, they cannot indicate the suitability of a particular text's content or literary merit for particular learners. The choice to read is usually a decision best left to educators and parents in consultation with the learners themselves.

Space constraints do not permit a thorough discussion of a new type of technology affecting reading, portable digital devices such as Amazon's Kindle or Sony's e-Reader, though they will clearly affect the way texts and books are read in the years ahead. A number of thinkers have already predicted that only devices that are wired to the Internet will have a long-term appeal, as the web will enable them to integrate and use free online glossing, storage, review and translation tools. Proprietary devices such as Kindle are currently limited by copyright protection, and can only read Digital Rights Management (DRM) content from Amazon. As Amazon mainly sells publishers' books, their central interest will be commercial rather than educational, thus creating a natural conflict of interest in terms of the real costs of overheads and royalties (for author and agent). All of this will continue to drive the price of e-books too high for most normal consumers. Many now believe that globalization includes the ideal of making generic knowledge freely available to as many people as possible, and making computers as cheaply available as possible, as evidenced by the MIT-sponsored One Laptop Per Child Foundation (OLPC). Those supporting these general educational ideals would no doubt agree with OLPC's five core principles:

1. Child ownership
2. Low ages
3. Saturation
4. Connection
5. Free and open source (OLPC, 2008, n.p.)

With such a contrast of principles and features available, and the substantially cheaper online costs and greater benefits available through open source online materials, it is not hard to predict who will win the next Information Revolution.

Accelerated Reader is another online commercial learning information system designed to help teachers manage and monitor their learners' reading practice. Renaissance Learning offers a set of teaching practices online; information on judging the suitability of books; articles on readability and how to use it in the classroom. Most Extensive Reading (ER) is done at the free or independent reading level. However, there is not only a need for more careful, individualized testing of what exactly constitutes each student's free-reading level, particularly when it comes to foreign or second language readers. Those having non-European native scripts are often more challenged and frustrated than native readers would be, due to having even more differences between the expectations of that second language system, as compared with reading their own native text. In addition, the complexities and different skills required to read online text fluently make it a rather different species of reading, whether in L1 or L2. Thus, the need for a strategy to better assess online as well as print texts in more uniform ways is becoming more apparent to reading teachers worldwide.

LITERATURE REVIEW

First, it is important to define Web 2.0 with reference to O'Reilly. Then we will contrast this definition with emerging definitions of nascent

Web 3.0. From O'Reilly's definition we can see that rather than merely being a new technology, Web 2.0 is characterized by a new mindset, or attitude towards the use of the Internet:

Web 2.0 is the network as platform, spanning all connected devices; Web 2.0 applications are those that make the most of the intrinsic advantages of that platform: delivering software as a continually-updated service that gets better the more people use it, consuming and remixing data from multiple sources, including individual users, while providing their own data and services in a form that allows remixing by others, creating network effects through an "architecture of participation," and going beyond the page metaphor of Web 1.0 to deliver rich user experiences. (O'Reilly, 2005)

Secondly, there are distinct differences both in how technologies are seen and used in successive generations of the Web, just as there are clear differences between reading print and online reading. Although Web 1.0 took people to information, Web 2.0 is involving people in information and knowledge construction, following a constructivist philosophy, creating what has become known as "an architecture of participation" (O'Reilly, 2004). While typical definitions characterizing this new style of online participatory information-gathering and meaning-construction generally range from positive to almost ecstatic hype, there are definite dangers and downsides for businesses, parents and public institutions to be aware of.

Devo (2008) explains both the upside and downside of emerging Web 2.0 tele-communication applications, which have been spreading like viruses. These include wikis, blogs, mash-ups, folksonomies, social bookmarking and podcasts:

Neatly labelled Web 2.0, zealous users appear to see bilateral and multi-lateral discourse with others, using collaborative and social networking, as not so much good as utterly compulsive. For the business world, there is a real concern emerging on the coat tails of the social explosion — a concern that employees are now spending so much time networking, that business could become a secondary consideration. (n.p.)

So what do we mean by Web 2.0? Devo (2008) summarizes Smee, marketing director of the Web Technology Group, who notes that this catch-all term for online social networking is still much misunderstood:

"In a way, the Web 2.0 label is a buzzword and there are lots of different interpretations of what it means," she says. "My personal view is that it is what Web 1.0 was always meant to be, which is simply to enable knowledge sharing. Putting user-generated content in the way of blogs on to the web is just the next step. It is an evolutionary process and not a case of yesterday we had Web 1.0, today we magically have Web 2.0." (n.p.)

Others have called this basic change of focus seen in many Web 2.0 applications a major paradigm or head-shift, an embracing of a freer, more two-way "E-democratic" mutual sharing of information by both user and provider. As such, it is seen as being most useful for e-Learning and sharing of discussion on social, political economic and human rights issues, as well as the full gamut of human discourse. Space limitations prohibit a full discussion of these issues, but groups like Involve (www.involving.org) go into detail about past failures and future hopes from using these new participatory online technologies more intelligently and democratically (Bryant & Wilcox, n.d.).

Comparing three generations of Internet website design and usage, it is possible to contrast Web 1.0 (read-only web) and 2.0 (read-write web) with what is now perceived as the Internet's future, Web 3.0 — a term which refers to the emergence of users who can modify substantial parts of the site or web-based resource. Others foresee Web

3.0 as an evolution of Internet use and interaction where it becomes a database in which information is accessible by various non-browser applications, not only by different browsers as at present. What is important for improving online reading and language learning in such an environment is to make sure that whatever browser or generation of Internet is being used, learners have instant access to a wide variety of glossing and translation engines, Text-to-Speech listening support, review test generation and language development programs that are needed to maximize their target language vocabulary learning recognition and use. Some others have seen the term Web 2.0 as just a marketing term, contrasting it with a more 3-Dimensional Web, which leverages various artificial intelligences, the Semantic and Geospatial Webs, into what Berners-Lee called a Giant Global Graph (GGG) (Dignan, Perlow, & Steinert-Threlkeld, 2007). He sees Web 3.0 as more of a "Social Graph," representing its third great conceptual leap — from net to web to graph (n.p.). Probably the best comparison of these second and third generations of Internet use is in Spivack's (2006) article called "The Third Generation Web is Coming." There he discusses its eight major characteristics, as well as offering the clearest definitions and distinctions between the first three generations of the web.

Finally, we can gain a good prediction of how the Internet will develop in the third decade of the Web (2010–2020), during which Spivack (2006) suggests that several major complementary technology trends will reach new levels of maturity simultaneously. His expanded definition of Web 3.0 envisions the third-generation of the Web as being enabled by a convergence of several key emerging technology trends. He predicted these new features of Web 3.0 would include:

1. The transformation of the Web from a network of separately siloed applications and content repositories into a more seamless and inter-operable whole.
2. Ubiquitous connectivity, broadband adoption, mobile Internet access and mobile devices.
3. Network computing, software-as-a-service business models, Web services inter-operability, distributed computing.
4. Open technologies, open APIs and protocols, open data formats, open-source software platforms and open data (e.g. Creative Commons, Open Data License).
5. Open identity, OpenID, open reputation, roaming portable identity and personal data.
6. The intelligent web, Semantic Web technologies such as RDF, OWL, SWRL, SPARQL, GRDDL, semantic application platforms, and statement-based datastores.
7. Distributed databases, the "World Wide Database" (enabled by Semantic Web technologies).
8. Intelligent applications, natural language processing, machine learning, machine reasoning, autonomous agents. (Wikipedia, 2008a, n.p.)

According to O'Reilly and Battelle (Wikipedia, 2008b), an architecture of participation where users can contribute website content creates network effects. Thus, in order to most effectively harness the power of the Internet for language education following a Web 2.0 paradigm, teachers need to learn to leverage the power of its "Long Tail" to develop "an architecture of participation where users can contribute website content [that] creates network effects" (n.p.). Since data becomes a driving force in Web 2.0, and even more so in Web 3.0 models, language teachers need to know and focus on which vocabulary (or lexical corpus and collocations) and grammatical structures their learners need to be exposed to in order to reach higher levels of fluency.

How can this be done most effectively to enhance online vocabulary and related reading comprehension development and language learning? This can be accomplished by integrating various programs needed by language teachers and learners into a more seamless whole, as is being done at some more innovative CALL sites. WordChamp.com, for example, combines many automatic functions, such as auto-glossing, auto-archiving, audio and visual enhancement, auto-uploading and test generation to provide a complete Course Management System (CMS) for courses. It also established peer-to-peer communication between users from 137 language backgrounds from any point in the world, using both an internal Instant Messenger system and the possibility of file-sharing. Learners' or teachers' vocabulary files can also be uploaded online or use mobile devices easily.

So perhaps the best way to enhance language learning using the Web is by finding and using good programs like these and by building more open source language learning communities online, which encourage maximum active participation and collaboration in the exchange for authentic communication between learners and speakers/readers of various languages. This means teachers need to embrace the web as a platform and aim to use its strengths (global audiences and collaborative learning, for example). Rather than fight or ignore the Web, teachers and researchers should aim to build applications and services around its unique features, especially its ability to enable users to both create and share content across various networks and boundaries.

USING READABILITY ENHANCING PROGRAMS

Chun (2006) examined CALL technologies for L2 reading, and compared the effect of providing some type of glosses upon improvements in vocabulary acquisition and reading comprehension. As she stated, results from CALL studies should always specify participants' L2 language proficiency, and cannot be generalized to all L2 learners. Chun noted various implications for online reading instruction from Grabe's (2004) reading research. Her rationale for having language learners use electronic and multimedia glosses is that:

> *They aid readers in performing the bottom-up function of recognizing and/or understanding individual lexical items, which in turn frees up working memory capacity and allows more of the reader's attention to go toward the top-down processes of reading comprehension. … Online glossing is thought to provide fast and easy access to the meanings of unknown words and to compensate for insufficiently automatic lower level processes and thus allows the reader to attend to higher level processes.* (Chun, 2006. p. 70)

There is still a lack of extensive, quality research about how to improve both skills and assessment of reading online, and more generally, how to improve the readability of web pages for learners from various backgrounds. Taking students to the Web should serve the double purpose of helping them to learn to read better either in their native (L1) or target foreign language (TL/L2), while simultaneously helping them to improve their acquisition of essential electronic literacy skills needed to cope with content and/or academic courses. While some of this delay seems caused by resistance to educational and technical change, instructors also seem to still lack clear pedagogical or theoretical models of reading online. Better understanding and application of Web 2.0 and Web 3.0 technologies can certainly help to design more effective models for successful interactive online reading and language learning communities. Another example is *Qnext*, a site which promises to be one of the fastest growing phenomena since *Facebook*, as it enables users to integrate and communicate with any Instant Messenger program, and share any and all files online with anyone else anywhere, free of charge.

Indeed, the educational community does seem to be a bit slow in making the transition from traditional text-based reading to online reading, which requires the teaching and learning of different perceptual approaches in both L2 text comprehension, as well as in lexical acquisition and processing strategies. Two major book readability grading systems already exist online:

1. Renaissance Learning's program which features the "Accelerated Reader" system with computerized quizzes and record tracking for more than 22,000 titles, also known as ATOS.
2. Touchtone Applied Science Associates' (TASA) Depth of Reading Power (DRP) program. TASA Literacy Online uses a scale of 0-100 in their own measure of text level and student reading level. They call these levels Degrees of Reading Power (DRP). They have also designed and used tests of vocabulary in context called Degrees of Word Meaning (DWM). Perhaps the best part of this vocabulary level testing scheme is that they provide a brief Conversion Table, which helps teachers convert these DWM vocabulary level scores into an estimated size of reading vocabulary.

Degrees of Word Meaning scores range from 850 (the equivalent to knowing over 157,000 words), to less than 300 (indicating that such a test taker knows 100 or fewer English words). Their products for educational assessment are numerous and include the Degrees of Reading Power (DRP) tests as well as online programs and steps for estimating both reading levels or the readability of any text or book.

Renaissance Learning is a commercial educational program for schools, with readability measures available for approximately 30,000 books. Anyone can estimate the level of any book by selecting three 150-word passages and emailing them to the site. Even more interesting is that users can enter three such samples in MS Word and freely analyze their level via Word Count with Readability measures activated.

Many reading teachers are looking for these kinds of helpful services to assess any text's readability. To help meet this need, Loucky (2005) developed an integrated English for Advanced/Specific/Technical Purposes online course, combining various online Reading Labs, for Japanese students. Its final Listening-enhanced Step for those having Natural Voice Reader or other text-recognition software includes:

1. Listen to the text read electronically.
2. Learners should try to understand its meaning phrase by phrase, paragraph by paragraph in whole sense units, not just reading word by word.

Where reading on screen differs most, however, is in the areas that Taylor (2005) notes. First, much more skimming and scanning is used. In fact 79% of Web users were found to be using these skills rather than reading word for word. This finding has huge implications for teaching the skills most needed for efficient online reading. Clearly both web writers and language teachers wanting to use online resources most effectively need to be aware of these major reading differences, as well as demonstrate sensitivity to the foreign readers for whom reading L2 texts designed for natives is often impossible or highly frustrating. Some studies, such as those done by Sun Microsystems Science Office have claimed that "Reading from a computer screen is 25% slower than reading from paper" (Nielsen, Schemenaur, & Fox, 1994). Many differences in reading rates have also been noted between on-screen reading using a monitor versus the printed page. Some of these differences may be due to differences in the text delivery system, in the layout, number of columns and length, etc. Other distractions to the reading task online may be caused by the need to use scrolling, hyperlinks or pop-up ads. All these factors may impede one's reading on a digital screen. Once practiced and proficient at reading online, however, probably the opposite is true for fluent readers. Much more research comparing skills, speed and accuracy levels of online versus print text reading is needed on readers at various levels, both native and non-native speakers, before such general claims can be accepted.

Teachers trying to use CALL or e-learning as well as web writers need to write in clear chunks and make text scannable at a glance, since an eye-tracking study done by the Stanford and Poynter Institute found that online readers often focus narrowly upon headlines and summaries (78% of their eye attention was here). Since online readers must use more skimming and scanning to get the gist and locate relevant information quickly, and Web distractions can make their speed 25% slower, these principles recommended by Taylor (2005) can help improve the readability of web copy.

Online authors must aim to write web materials clearly and succinctly in summary style. As Taylor (2005) states, a Web materials writer should think like a graphic artist, treating "each page like a painting that is framed by the computer on this electronic canvas [where there] are elements that you, the artist, must weave together, linearly, to form a coherent whole that can be accessed with little or no reading. … Web writing places a premium on good organization of content and devices [navigational bars or buttons] that clarify the content's organization to the web user" (n.p.). Other principles advanced by Taylor also include:

1. Invert the pyramid of information, using journalism's major headlines and summary first style.
2. Compress information and be more concise, so reduce word count by at least 50%.
3. Make one paragraph carry one major idea.
4. Make each page's text stand alone, since users can enter through hyperlinks from various places.
5. Provide needed hyperlinks, using keywords as titles/headings. In other words, make useful and relevant links to helpful resources, both within a site as well as to other sites.

Some important new principles come into play with a Web 2.0 approach to information dissemination. These include a user-centered mentality, where more readable and comprehensible information should be made available to users, when, where and in forms that they need and can readily use. Since Web 2.0 enables more people to share and author information by means such as text and audio file-sharing, podcasting, photo-sharing, blogging, etc., such information can become more personally relevant and meaningful to specific user groups. On the other hand, copyright, expensive monolithic publication systems and "Information Gate-Keepers" will tend to be by-passed or disintegrate, while information-sharing technologies give users more direct access to publically available, Web-visible learning resources. But how can Internet resources be used most effectively to improve online reading and language development? This is the key question under discussion in this chapter.

Having established a better understanding of the major differences between reading print text versus on-screen text, teachers and web writers can implement better solutions for the special needs of L2 readers. Beside these clear reading differences, others have only become apparent in recent research. As Taylor (2005) noted, "In both cases, it's essential for web writers to be aware of the differences between the world of linear text flowing like a river, and the fragmented world of hypertext on a pixel screen. The most successful web writers have honed in on the key differences between writing for print and writing for the web" (n.p.).

Finally, we must remember that the Web is rigorously democratic, in that the user is in much more control of an online, interactive learning experience, than when reading print text, or other more passive, non-responsive mass media. The Internet is a self-access mode of learning, but a majority of learners may not be "self-starters," and especially foreign language learners can be quite intimidated and feel threatened by L2 online materials. Thus we need to provide a variety of levels (with both authentic and simplified text), entry points, and plenty of multi-media and bilingual assistance to aid and encourage their language learning.

Many of these factors are considered by Coll (2002), Loucky (2002, 2005, 2006a, 2006b), and Akbulut's (2006) studies of learning in hypertext environments. Online reading seems to discourage word-for-word reading, since print readers can hold the entire document in their hands at once, whereas a web document must be called up one page at a time, either by the action of scrolling or by using hyperlinks. Even better for discouraging single word reading, however, are reading pacers, some of which can be set by the learner to at least three different speeds to adjust for their own comfort and ability level. This type of on-screen reading function is essential, for at least some Web reading, especially for lower level readers. It is available, for example, when using Eichousha's *Reading Skill Trainer* software, or *Rocket Reader* online.

Another type of help for enhancing online reading called, Visual-Syntactic Text Formatting (VSTF), has been tested by Walker, Schloss, Fletcher, Vogel and Walker (2007). This method transforms block-shaped text into cascading patterns to help readers identify grammatical structures. It has helped increase reading comprehension and the efficiency of reading online texts while reducing eye-strain among college readers. This VSTF method also helped increase high school students' academic achievement and long-term reading proficiency by more than a full standard deviation over randomized control groups in one academic year. This new method has been made feasible through computer-executed algorithms and electronic displays by integrating converging evidence from educational, visual, and cognitive research.

There do not seem to be many rigorous online studies yet of Web reading done by non-native speakers, using both speed and accuracy tracking, and also eye cameras to track and monitor perceptual movements such as regressions when reading online versus on paper. If such a comprehensive public grading system could be made available, it might indeed become a useful standard for both web-based and paper-based ER materials, but one must also recognize what a large project it would be, requiring the input of various educational bodies to achieve wider acceptance and use. In the meantime, traditional reading level tests that are online should be used and compared to see which provide the most helpful and consistent results for learners from various language backgrounds.

METHOD AND MATERIALS

Programs for Improving Vocabulary Accessibility with Online Glossing

Two major bilingual glossing programs — Rikai.com and Wordchamp.com (See Burston, 2007) — were used in a graduate reading course, alongside two online vocabulary level checkers. Rikai.com provides glosses from English to Chinese, Japanese, or Spanish, and can do auto-archiving of all target words looked up for later printing and review. Wordchamp.com will be explained in more detail below. The first vocabulary checker used was a simple, author-designed Vocabulary Knowledge Scale, known as the Dual Assessment Vocabulary Instructor-Evaluator (DAVIE). The second vocabulary checker used is called Vocab Check. In researching the use of these tools, three main research questions were identified:

1. Our primary and most basic research question was to find out: What is the relationship between using new types of Web 2.0 technology for CALL? Specifically, how can this more interactive, mobile World Wide Web of educational and social networks be used most effectively to produce the much-promised transformation of learning? In what ways

can Web 2.0 and 3.0 transform and improve language learning?

2. Secondary research questions applying this new technology to vocabulary, reading and language learning were: Do web 2.0 technologies contribute to the development of both intensive and extensive reading skills? If so, how? Why, or why not? Which online resources do EFL learners find most helpful in learning and using new English vocabulary and grammatical structures? Language educators and researchers need to ask: What are the implications and uses of web 2.0 for language education with reference to its innovative audio-visual, participatory and assistive technology? How can these new online technologies best be harnessed to increase language learning rates, proficiency and enjoyment? How can language teachers effectively guide students in the use of such online resources and socially interactive programs, in ways that are ethical, enjoyable and educational, so that their use does not degenerate into classes of online dating or worse?

Participants

Two groups of students were involved in this study. The first group consisted of a class of Japanese engineering students on a master's course (M = 38, F =1). The second consisted of two classes of English and Applied English Major students. Of the latter, one class had 15 students from the National Taiwan Normal University; the second class was from St. John's University and had 37 students. Thus, the total number of Chinese English students in this study was 52. The total of Japanese and Chinese in this study was 9, of whom 56 completed the English surveys summarized below.

Students taking the online reading course in Japan were Master's candidates in the new Department of Applied Science for Integrated Systems Engineering at a national university in Kyushu. Students' vocabulary and comprehension level and total estimated reading level were computed at the start of the semester course relative to American norms. A "Course Survey" and a "Website Evaluation" were also given at the end of this one-semester course. Average class reading levels for all 39 Japanese students relative to native reader norms in America (Loucky, 2003a) were assessed at the start of the fall semester: the average vocabulary level was grade 3.93, equivalent to the start of fourth grade level in the USA. The average reading comprehension level was 3.02, hindered by this low vocabulary level. The average expected reading grade level was the middle of third grade, or 3.51. Students wrote brief reports on each reading including a) a summary paragraph, b) impressions paragraph, c) five free comprehension questions and answers of their own, and d) constructed complete sentences for each new word they had listed. These were each printed or emailed, corrected by the teacher and returned for oral interviews.

DURATION AND DATA COLLECTION

Students had ten weeks from a fifteen-week one semester course in which to write at least five emails to Distant Learning Partners (DLPs) on their own outside of class, constructed around five general themes. They could choose words freely from pre-arranged Semantic Field Keyword (SFK) groups relevant to five academic disciplines, to help guide their writing. Each of these students was randomly assigned a keypal in the other country to write to online. Students had an average of two weeks for each email exchange, after which they were to print and submit it for credit. Partners could give each other peer-corrections, as well as making any corrections on their own after getting quick markups from teachers of where errors might exist in grammar, wording or usage. Students simply received credit for all email exchange letters handed in, and rough markups so they could make corrections and resend if they wished to do so. They were awarded a grade from 60-100% based on how well they completed email assignments in terms of their purpose, theme, and use of proper lexis and grammatical structures. Only printed versions were checked, though drafts could be written. Sending corrected versions to their partners was encouraged, but not checked or enforced.

Japanese students did all of their readings for this study online. In addition to online writings, blogs and use of a bilingual program, the Taiwanese students also used a writing text called *Steps to Writing Well* (Wyrick, 2005). Blogs did not work well enough to enhance these Chinese students' English this semester, so peer comments were encouraged and collected in addition.

ONLINE COLLABORATIVE WRITING PROCEDURES

Students were assigned five topics to write on, and told to use questions or statements for each of them, depending on their purpose and content. These five Collaborative Writing Exchange Topics Using the Semantic Field Keyword Approach (Loucky, 2004) within a Task-Based Language Teaching approach (Willis, 1996) were:

- **Topic 1:** Interview your new distance-learning partner (Using terms from Unit 1: Scientific Experimentation).
- **Topic 2:** Tell "My Life Story" (Using terms from Unit 2: History).
- **Topic 3:** Interview your new distance-learning partner (Using terms from Unit 3: Psychology).
- **Topic 4:** Tell your view of man, or how you think people and human civilizations came to be and where you think the world and humanity is going. (Using terms from Unit 4: Anthropology).
- **Topic 5:** Describe the culture and traditions of your people and country so a foreigner could better understand your nationality. (Using terms from Unit 5: Sociology).

These topics were first shared using the three phases of Willis' (1996) Task-Based Instructional Framework:

1. Pre-task Introduction
2. Task Cycle
3. Language Focus

For each of these five writing topics (different for each of five academic discipline areas and themes assigned), they were given 36 sets of Semantic Field Keyword groups of similar meaning-related words, but students could choose which of these they wanted to use and in what order. Their motivation was greatly heightened for writing in EFL since these were cross-cultural "Collaborative Writing Exchanges" between them as Japanese engineering graduate students and Taiwanese undergraduate English students. Data collection mainly consisted of recording students' averages for email exchanges done, pre-and post-test data for Japanese students on their knowledge of the first Unit of Semantic Field Keyword groups, and survey results for all who chose to complete English course surveys.

RESULTS

Pre- and Post-test class average scores for the Japanese graduate engineering students for Semantic Field Keyword Approach Unit 1-1 Sample provided interesting results. The pre-test raw scores were 22.86/60, equivalent to 37.92% organized correctly. Similarly, post-test raw scores were 33.73/60, or 56.11% organized correctly. The Learning Rate for Unit 1-1 was therefore 18.19%, a very good rate for a short-term study.

Taiwanese students did not use the LEARN Online Reading Lab program. Being higher-level English education majors, they both did some limited blogging and peer-correction of other written essays. Researchers only had direct control over the Japanese student's instructional material, but Taiwanese email exchange classes cooperated fully on five SFKA writing exchanges.

Survey of Online Reading and Collaborative Writing

On the "Survey of Online Reading and Writing Collaborative Course," 38 Japanese Graduate Engineering students and collaborating Taiwanese students completed an English survey and generally answered the questions very positively. 97 Japanese and Chinese participated in this study and 38/45 Japanese students completed surveys. However, just 18 Taiwanese students completed English surveys. The Survey Questions were as follows (*N=56*):

What have you learned from using the pre-organized, bilingual Semantic Field Keyword Approach online and doing Collaborative-Writing Exchanges using some of these words within assigned grammatical or topical frameworks?

Japanese Students' Answers: 13/37 or 35.14% were very positive. Chinese Students' Answers: 16/18 or 88.89% of answers were positive.

Which topic did you find the most difficult to write about? Why?

Various answers were articulated, for example:

Japanese Students' Answers: 1 each said the "Freed Hostage Trio" or the "Sake Story"; 4 said: "Kagawa's story because I didn't know him at all" (1 due to its difficult vocabulary); "Also his life was so busy it's hard to tell all that he did!" 2 said: "The first story in Japanese about Scientists." 2 said: "Manjiro/Neejima Joe." 3 said: "Pearl Harbor." 1 wrote: "(PH) Story, as I don't know about war"; 1 said: "PH as I didn't know those words. A third wrote: "Pearl Harbor, because the story is very long." 6 said: "SFKA Topics 4 & 5, Anthropology and Sociology — We don't think about it usually; I don't know much about my country to explain it to others." 2 said: "All SFKA Topics. All, since written in a language I don't know well."

Chinese Students' Answers: More than half wrote: Anthropology, "because the words and subject are complex or difficult (abstract)." Human civilization or Academic subjects, "since I am not professional in those fields"; "Because one's view of man is the most difficult topic, with many new words, so it took longer than others."

Do you think SFKA word lists improve your writing? Why or why not?

Japanese Students' Answers: 100% or 38 were positive. Chinese Students' Answers: A. 55.56% positive. B. 2 negative (5.25%). C. 33.33% or 6 gave no answer.

Do you think using the pre-organized, bilingual Semantic Field Keyword Approach online is a good way to help you increase your English vocabulary? Why or why not?

Japanese Students' Answers: A. 21/37 (56.76%) were positive. B. Negative: Only 1 (2.7%). C. Neutral/No opinion expressed: 15 (40.54%) gave no opinion.

Chinese Students' Answers: A. 11/18 (61.11%) positive. B. 1 Negative. C. 2 Undecided. D. 1 Recommendation given: One said: "It'll be better if SFKA could show us some example sentences." E. 2 (11.11%) Neutral/No opinion expressed.

Do you think our email exchange successful? Do you think your writing improved after this *exchange? State the reasons for your opinions clearly please.*

Japanese Students' Answers: A. 22 (57.89%) positive. B. 8 (21.05%) negative. Including 1: "No, because I couldn't get emails." 1: "No, since words or topics were limited." 1: "No because one sent me none, the other's English was too difficult for me." C. 1 Neutral said: "Not sure if successful, but it improved our English!"

Chinese Students' Answers: 7/18 (38.89%) of Taiwanese said: A. "Yes, it's a good way." B. 5/18 (27.78%) gave "No" answers. C. One (5.56%) gave it a 50% rating. Reason: "Because our keypal was Japanese … If the nationality could be European [with higher language proficiency many others wrote] that'll be better."

If you took the V-Check or used the WordChamp.com website to test your vocabulary level online, did it help you? If so, please tell your impression or opinion about how it was helpful or motivating to you. (Japanese students used both V-Check and the WordChamp.com website. Chinese only used V-Check).

Japanese Students' Answers about V-Check: 14 (36.84%) were positive, 2 mixed, the rest (58%) gave no specific response to this question. However, when ranking websites used in class for the usefulness, 24/38 (63%) of them chose WordChamp.com as their first or second favorite site.

Japanese Students' Answers about WordChamp.com: A. 18 Positive (47.4%). B. 2 Negative, "No, it was too slow on our LAN."

Chinese Students' Answers about V-Check: 8/18 (44.44%) were positive. Five wrote mixed opinions, with complaints or suggestions to improve V-Check.

Did you gain any new ideas or strategies for improving your English vocabulary, reading or writing strategies and skills? Yes or No. If Yes, please tell which particular skills and strategies did you learn that you will use the most or may help you most practically in the future?

Japanese Students' Answers: Six (16%) said: "yes." 4 wrote: "Reading in English improved. I will study English harder!" Two said: "Especially I gained Vocabulary skills; I gained practice in writing English sentences!"

Chinese Students Answers: Half (9/18) were positive.

If you took the V-Check to test your vocabulary level online, what was your impression or opinion about it? Was it helpful to you? Motivating?

Japanese students were given a survey about their online reading practices and learning, shown above. Chinese students instead were asked: *What do you feel about peer-commenting? Did you benefit from your classmate's comments? Why or why not?* Taiwanese students used and appreciated the ability to develop their English by doing blogging and peer-correction online.

Japanese students' responses to a mini-survey on websites and reading methods used in class and two scientists studied using them will be summarized here, called Question 9 on their survey.

Rank 1-5 which of these Websites "Most helped you to improve your English skills." Also "Put a check" on the right of any Website that you used if it helped you to learn new English vocabulary, reading/writing skills and strategies online (Rank#, then # of Students). Rank 1-5; Circle with "O" if this Website helped you to improve.

A. (www.call4all.us) #1-18; #2-7; #3-4; #4-2; #5-1; #7-1.

O. 26/38 students said this website helped them to improve their English. Numbers show how they ranked each site.

B. (www.WordChamp.com) #1-10; #2-14; #3-2; #4-5; #5-2. O-25/38 "Helped my English".

C. (www.Rikai.com) #1-6; #2-10; #3-13; #4-1; #5-3; #6-1; #7-1. O-26/38 “Helped”.
D. Online Dictionaries — CALL4ALL’s Dictionary page, or which online dictionary did you use most? #1-2; #2-8; #3-7; #4-11; #5-3 #6-1. O-18/38 “Helped my English.”
E. (www.learn.com Site) #1-2; #2-3; #3-5; #4-6; #5-10. O-19/38 “Helped.” 1-X “Did not help.”
F. Online Reading Labs (assembled at R-Reading Page of CALL4ALL.us). #1-2; #2-5; #3-6; #4-3; #5-6; #6-5. O-17/38 “Helped my English.”
G. OTHER English-Japanese Web Dictionaries Used: 4 Listed. 2 used ALC (SPACE); 1 Sanseido.net; 1 MSN Encarta Dictionary.

To summarize the above ranking of websites used, one can see that Loucky’s course website was the most highly favored, with 68.42% (26/38) saying it “Helped me improve my English,” and 47.37% (18/38 students) choosing it as their #1; 7 students as their #2; 4 students as their #3; 2 students as their #4; 1 student respectively as his #5; or #7 choice. WordChamp.com, found helpful by 65.79% of these students (25/38), was second in popularity, with 10 students choosing it as their first choice, 14 as their second, and 9 others choosing it as their third to fifth choices. Third in popularity was Rikai.com, found helpful also by 68.42% (26/38 learners), and chosen as their top by 6. It was second choice for 10, and third choice for 13 learners, with another 6 ranking it #4-7 among their choices. 32 students viewed CALL4ALL’s Dictionary page as being helpful to them (84.21% using it) 17 ranking it in their top 3 choices. Fifteen others placed it in their top fourth to sixth choices. When asked, “Which online dictionary did you use most?”, only these other online dictionaries were listed, all of which are included on CALL4ALL’s Dictionary page as well: 4 listed ALC (SPACE); 1 Sanseido.net; and 1 MSN Encarta Dictionary. Finally, Japanese student responses to the mini-survey on reading methods and scientists will be summarized here. Asked on their final Reading Survey section:

Which way of reading did you prefer?

92.10% (35/38) chose: a) reading online with bilingual glossing support, vs. only 7.89% or 3/38 chose: b) reading printed handouts or textbook articles?

An overwhelming majority (92.1%) prefer having bilingual glossing support for their online readings. Teachers and E-Learning developers should always keep this in mind.

Which way of reading was easier for you to learn the meaning of new words?

65.79% or 25/38 chose: a) When reading online with bilingual glossing support, or 28.95% or 11/38 chose: b) when reading printed handouts or textbook articles? So approximately 66% said it was easier for them to learn new words having such bilingual glossing.

Do you think you learned how to use strategies for reading more effectively?

32 chose: a) When reading online with bilingual glossing support, or 3 chose: b) When reading printed handouts or textbook articles?

84.21% state they learn to use a larger number of reading strategies more effectively when reading online with such bilingual support.

When reading about scientists using the LEARN Website, which way did you read?

A. Einstein Story: 9 read only and 14 read and listened to it
B. Edison Story: 8 read only and 16 read and listened to it

Which way of reading did you prefer doing in this class?

A. Online Reading without Listening support: 4 (10.53%)
B. Online Reading with Listening support: 31 (81.58%)
C. Offline Reading of printed texts only: 3 (7.89%)

Again it was clear that about 82% of these learners prefer reading online with listening support. This is a significant finding with potentially far reaching implications for CALL and Extensive Reading to analyze.

Based on your reading, who do you think was smarter?

7 answered a) Einstein and 8 b) Edison as opposed to c) with 23 and "both the same." Only 1 answered d) another scientist.

Why do you think so? Please give your specific reasons for your opinion here.

23 said "Both." Example answers included: "Both were great, so I can't decide." "Because I think all scientists are great." "These two are too smart to compare!" "Both were geniuses." "Both, because everyone knows both." "Both contributed to the world's development." A number said, "I respect both. They are not comparable, since Einstein theorized, but Edison manufactured." "Both are great scientists, both very smart." "We can't compare them." 7 chose Einstein, one saying, "Einstein, since his IQ score was 300!" 8 chose Edison, one saying "because he not only invented new products, but also a system to earn money!" "I like Edison … most popular inventor for our lives."

When designing language learning websites three major parameters of subjective enjoyment and objective effectiveness as well as technological efficiency should all be considered. In order to do so, students' improvement during this one semester course was assessed by two measures: a) average performance and participation in written reports and twelve online articles, and b) overall performance during three sessions using the Online Reading Lab articles. Their performance when reading these articles was assessed in three ways: 1) by the average number of stories read, 2) by their average speed when doing these timed readings, and 3) their average percentage of comprehension for all stories read during each session.

A majority of students reported that using the teacher's website (www. CALL4all.us) made the course very enjoyable and efficient for them. Students always did the reports unless absent, often making up written reports with much diligence, resulting in an overall class average of 76.75% on these homework reports, which were graded based on their grammatical accuracy, completeness of reporting and word study indicated. Objective test results — 59% average online comprehension despite this EFL class averaging just 3.5 in their total reading grade level — also showed a good level of improvement in learners' average vocabulary and grammar use levels, clearly supporting the effectiveness of such a blended online course. Thirty-five students completed an average of 18 online readings in a mean time of 6.78 minutes per reading. Since these readings were designed to be read in just five minutes, it became apparent that these graduate engineering students need more work on learning the essential core vocabulary required to read at a higher level with greater speed.

These were Japanese average comprehension scores for all readings done using Balsamo's Online Reading Lab on each of three days, as well as students' total overall average. As one would expect, from an initial average score of 54.19%, their comprehension scores increased to 63% and 60.5% on two subsequent days. Each time they were encouraged to try to read ten online articles on topics in areas of their choice. Students' total overall "Online Reading Averages" when doing timed online extensive readings on topics of their choice were as follows: 1) Average Comprehension, for Day 1: 54.19; 2) Average Comprehension, for Day 2: 63; and 3) Average Comprehension, for Day 3: 60.5. The Total Average Comprehension was 59.39% over three days using this online reading lab.

In sum, both objective and subjective assessments showed that a large majority of these students improved markedly, and enjoyed this course, which blended assigned online readings with integrated four skills English language development activities (written reports and paired interviews based on online readings) as described above. The course was not long enough (just one semester) to measure reading gains by grade level.

Students wrote brief reports on each online reading including: a) a summary paragraph, b) impressions paragraph, c) 5 free comprehension questions and answers of their own, and d) constructed complete sentences for each new word they had listed. These were each printed or emailed, corrected by the teacher and returned for oral interviews, emphasizing oral and written correction of grammar errors. All reports received a grade as they accounted for 80% of the semester grade. Consequently, assignments were taken seriously and done regularly by almost all students. Final class average for ten of these reports required was 78%, a figure close to Japan's A level for 80% and above.

This five-month semester course emphasized developing online reading skills using bilingual glosses and regular, blended and balanced integration of CALL with all four communication skill areas as described above. It was necessary to try to balance an intensive reading approach to cover higher level technical articles assigned by other engineering teachers, with an extensive approach using an online reading lab. The students' general surveys (*N=38*) showed an appreciation for both approaches, and improvement in their speed and comprehension during second and third sessions using the online reading lab as follows. Using Balsamo's Online Reading Lab, they averaged reading 18 stories over three weeks, at an average speed of 6.78 minutes. While average comprehension scores were close to just half (54.18) during the first week, they improved to 63 and 60.5% during weeks 2-3.

DISCUSSION

We have been able to develop a multi-purpose language learning site including an Online Reading Lab (ORL) and succeeded in fully integrating practice in all four communication skills with it for a graduate level course. Since the learners' average vocabulary level (grade 4.0) was comparable to that of undergraduate freshmen engineering students at the same national engineering university in Kyushu, Japan, such a course using only the Online Reading Lab's easier articles could be more successful in the future.

Technical articles would be skipped and simpler Rikai.com articles used instead, especially ones having instant online bilingual glossing available. The following resources and services were provided by this course and website:

1. Interesting, authentic online reading materials (copyright free).
2. Comprehensible input facilitated by instant bilingual glossing and other web dictionaries.
3. Comprehension questions on each article were available for each timed, online Reading Lab article. Learners wrote their own questions and answers for online articles, chosen and assigned by ten other engineering professors, to enhance and ensure their mental and linguistic interaction with each text. These were followed up with

oral/aural practice using these same questions after being checked for grammatical accuracy by the teacher.

4. Feedback was given by writing brief summaries, impressions and comprehension questions for each of these 10-12 academic articles.

The high levels of learner enjoyment and clear effectiveness of this type of CALL-based ESP online reading course suggests that many more courses should strive to have a web presence, especially reading and writing courses. This study also showed the benefits of giving end-user surveys and interviews, as well as objective post-tests and ongoing monitoring and assessment of students' learning, in order to improve such courses with added feedback. This online ESP course blended with interactive, communicative language learning activities both in-class and out also revealed that making parts of an online reading course available at all times on the Web and demonstrating it in class can ensure that students do use it effectively. Not only do language learners use such a website when it is intentionally and effectively integrated into regular class use, but they also seem to greatly enjoy and benefit from using it, as reported on their course surveys averaged by the school, and demonstrated in their online reading reports.

These were the results for assessing just the first article on hurricanes from Balsamo's online Reading Lab:

1. In less than 2 seconds, so much linguistic and lexical data can be generated for any text such as this, either inputted from any online text, text file or scanned text that one must summarize only the word data types, as it generated seven pages of data. They included this information about word families, types, tokens and percentages; a color-coded text showing word bands clearly with different colors. In addition, Token Lists for various Word Bands were all printed out. The AWL File produced at level 10: for Hurricances and Tornadoes article showed these academic words in bold print: similar, temporary, area, normally, predicted, ignored, considerably, normal, enormous, encounter, and capable. In this program each level includes all the previous levels, so band 10 includes 1-10. By providing such color-coding and word frequency bands, teachers can help students to focus on how to study the words they most need to learn in communicative and effective ways.
2. AWL only highlights ten levels of academic words within similar bands by bolding them. This is very helpful for quickly focusing both teachers' and learner's attention on essential vocabulary for understanding that text, for example here shown for the first Academic Word List level. This program will identify core academic vocabulary in a text, using the Academic Word List. It does look easier to print and much more manageable for teachers and students who are not linguists than the Vocab Profiler, whose advantage is its ability to assess both easier General Service List (GSL) words, as well as AWL words, focusing learners' attention on words above their present level.

English Vocabulary Profilers

Other linguistic data important to note and summarize here are these facts, which can be edited from an excellent function provided by Cobb's Vocab Profiler site called "Edit/print-friendly table." It is important to note that while our Target Story was reported to have only about 3% (2.68%) AWL words, 15.05% of the text are off-list words, which must be known to comprehend the story or read it fluently with adequate understanding. Since no more than 1 in 20 running words or 5% should ideally be unknown even for native readers (Ekwall, 1976), encountering these close to 18% yet unknown AWL and Off-List Words would make even this short article incomprehensible or frustrating for a majority of Japanese college students. Most undergraduates possess an average of only about 2,500 words, with graduates averaging about 3,500 known words, among thousands of learners studied repeatedly at seven colleges over ten years (Loucky, 1996; 1997; 2003a; 2003b).

Alternatively, one may use the AWL Highlighter to work on vocabulary found in the Academic Word List, but this has only 570 words (Coxhead, 2000). Thus off-list words needed by students would not be covered here, making the Vocab Profiler a much more versatile instrument, especially once learners have mastered these AWL terms. As an example, when inputting our target Pearl Harbor Story text into it, CAVE allows one to choose which of the AWL Sub-lists to scan for. At the highest level 10, these 35 words were highlighted (and at times repeated) in bold type by this vocabulary search engine: intelligence (information), objective (aim), military (adjective form), preliminary, accurately, plus, exploit, primary, principal, intervention, converts, ignore(d), encounter, assistance, conference, distributing, published, involved, committed, eventually, volunteered, found, finally, eventually, relevant, dynamic, attitude, liberated, motivation, seeking, purchase, despite, traditionally, drama, substitute. The advantages of using CAVE first are that it is more narrowly focused just on helping one to identify AWL terms needed by sub-lists, without distracting users by any other linguistic data, many of whom would be overwhelmed by Vocab Profiler's excessive data.

3. **Flesch Reading Ease:** First a percentage of passive sentences is shown as 36%. Then a reading ease score of 54.1 out of 100 is given.
4. **Flesch-Kincaid Grade Level:** Finally, the most important reading level perhaps for teachers to know and pay attention to when assigning online reading tasks is this, since it determines a text's estimated grade level of difficulty. The Hurricane text was assessed as being at a grade 10.2 (2nd month), so that students at a level of more than 6 months to a year lower than that should not be asked to read such texts for free reading. Generally speaking, such texts could be used for instructional reading for students reading at 1-2 levels below that, from about grade 8-9 level. Students reading at less than that level would tend to become frustrated with such texts, mainly due to their not knowing over 5%, or 1/20 running words. Two conditions could reduce their learning burden, to enable learners reading at lower grade (such as grade levels 5-7) to endure such texts without undue frustration: a) allowing and instructing them to use online or portable bilingual/bilingualized dictionaries, or b) if they have a very strong interest and background knowledge in the field of a particular text. Otherwise avoid the frustration level. Other lexical and linguistic data displayed at the same time by this program for this reading text for example were these facts. This text had 565 words, 9 paragraphs, 30 sentences; averaging 5 per paragraph, 18.4 words per sentence, with 4.6 characters per word.

Interestingly, reading pacers differ. The one used at Balsamo's online Reading Lab is basically a five-minute countdown speed-reading stimulator. Robb's (2008) reading lab site provides another timing device, which ideally should be part of all online reading or language learning labs. It measures total time on task. Learners and teachers can thereby get clear measurements of either free-reading or study times. Adding adjustable pacers and levels of text difficulty to all Web pages intended for E-Learning — along with a choice of either bilingual, monolingual glosses or both, as well as instant Text-to-Speech services — would be even more ideal, especially for lower level language learners. Adding listening support should be done wherever possible for language learners, since Extensive Reading alone is known to be too slow to allow more than incremental vocabulary development to occur (Rory, 2005). These options are already all available through references and links integrated at the CALL4ALL website (Loucky, 2008), which serves as a free Web 2.0 Virtual Language Education Library of various websites and applications useful for learning or teaching 120 languages.

CONCLUSION

More innovative Web 2.0 technology enables us to manage increasing amounts of semantically rich metadata and to deliver software and services for information management to users from most language backgrounds worldwide. There is a clear drive towards connectivity happening globally today, so that almost anyone with access to a computer can connect with millions of people around the world to collaboratively create, share and consume all forms of digital content at virtually no cost. Using Web 2.0 educators can now manipulate and share enormous quantities of data, so that people all over the world can more easily connect, talk to one another and exchange ideas, provided of course that they can overcome linguistic, cultural and vocabulary barriers.

Due to the potential for sophisticated mass collaboration that this technology wave provides, people are now able to make almost unlimited connections across the planet. However, often a majority of users are not English native speakers or readers, so they require many more means of support to enhance comprehensibility. Clearly such text analyzers, summarizers, glossing and translation engines to simplify text, as well as multimedia and TTS listening support options have great relevance to the needs of many users of Web 2.0, whether it be for online language learning, social collaboration, rating or tagging shared content, collaborative filtering of news, blogs, or other recommended content or to help improve the comprehensibility of any other shared Web browsers, program applications, components or recombinations.

In summary, one may assess vocabulary and reading levels not only for print but online for text from any of these Web components in the following ways, the first three of which are free and described by Loucky (2008) with links from the Reading and Readability page. Enter any text at any of these three programs to find out its reading level:

1. AWL URL: (http://www.nottingham.ac.uk/~alzsh3/acvocab/awlhighlighter.htm).
2. Vocab Profiler URL: (http://www.lextutor.ca/vp/eng/).
3. Word Spelling/Grammar Checker (explained in Office 2007 Word Help).

When Microsoft Office Outlook and Microsoft Office Word finish checking the spelling and grammar, you can choose to display information about the reading level of the document, including readability scores according to the following two tests: a) Flesch Reading Ease, and b) Flesch-Kincaid Grade Level.

Alternatively, enter any book title and or its ISBN to find out its grade level:

1. By using TASA's Depth of Reading Power, on a scale of 1-100.
2. By Reading Renaissance's ATOS, by school grade levels (relative to U.S. norms).

Loucky (2008) has been able to develop a multi-purpose language learning site, including several Online Reading Labs (ORLs) and succeeded in fully integrating practice in all four communication skills using it with graduate level Japanese engineering students. Since the learners' average vocabulary level (grade 4.0) was comparable to that of undergraduate freshmen engineering students at the same national engineering university in Kyushu, Japan, such a course using only the Online Reading Lab's easier articles could be more successful in the future. Technical articles would be simplified, by having the *WebReader* instant online bilingual glossing feature (www.WordChamp.com) made available for them, along with TTS listening support.

The following resources and services were provided by Loucky's (2008) online course and website:

1. Interesting, authentic online reading materials (copyright free).
2. Comprehensible input, facilitated by instant bilingual glossing and other web dictionaries.
3. Comprehension questions on each article were available for each timed, online Reading Lab article. Learners wrote their own questions and answers for ETP articles, chosen and assigned by ten other engineering professors, to enhance and ensure their mental and linguistic interaction with each text. These were followed up with oral/aural practice using these same questions after being checked for grammatical accuracy by the teacher.
4. Feedback was offline and done personally with the teacher, orally or in writing brief summaries, impressions and comprehension questions for each of these 12 academic articles.

The high levels of learner enjoyment and clear effectiveness of this type of CALL-based ESP online reading course suggested that many more courses should strive to have a web presence, especially reading and writing courses. This study also shows the benefits of giving end-user surveys and interviews, as well as objective post-tests and ongoing monitoring and assessment of students' learning, in order to improve such courses with such added feedback. This online ESP course blended with interactive, communicative language learning activities both in-class and out has certainly shown that making parts of an online reading course available at all times on the Web and demonstrating it in class can ensure that students do use it effectively. Not only do language learners use such a website when it is intentionally and effectively integrated into regular class use, but they also seem to greatly enjoy and benefit from using it, as they reported on their course surveys and demonstrated by high homework averages (78%).

Pedagogical Implications and Recommendations for Web 2.0 Reading Programs

Recent proposals for a standardized grading scheme for web-based reading materials are timely and welcome. This overview has shown how online reading lab stories and articles linked to Loucky's (2008) site can be easily copied and pasted into Cobb's Vocab Profiler for quick reading level analysis. The text of any scanned story or webpage can be analyzed in the same way, giving results that are extremely helpful to teacher, researcher or students in terms of word levels or frequency bands. Others such as McGovern's EFL Reading site report using a rudimentary scheme combining the readability statistics available with Word (Flesch Reading Ease, Flesch Kincaid Grade Level) with his own personal judgment based on experience as a teacher and writer.

Besides using these two Word readability formulas, both Cobb's Vocab Profiler and the AWL (using CAVE formula) site were used to assess basic reading level of articles from Balsamo's online reading lab. Links to each of these are included at the author's website, under L. Language Learning and Reading Labs Online. Teachers, learners or web writers should learn to use these. These were the results for assessing just the first article on "Hurricanes" from Balsamo's online Reading Lab: 1) in less than 2 seconds, so much linguistic and lexical data can be generated for any text such as this, either inputted from any online text, text file or scanned text that one must summarize only the word data types, as it generated several pages of data. They included this information about word families, types, tokens and percentages; plus a color-coded text showing word frequency bands clearly.

Two conditions could help *reduce the cognitive load or learning burden* of more difficult texts, online or in print, to enable learners reading at lower grade (such as grade levels 5-7) to endure such texts without undue frustration: a) allowing and instructing them to use online or portable bilingual/bilingualized dictionaries, or b) if they have a very strong interest and background knowledge in the field of a particular text. Otherwise we should always avoid frustration level materials, and employ reading materials at appropriate independent levels for free/extensive reading outside of class, or at instructional levels (generally not more than ½-1 year beyond independent levels) for content or classroom learning.

Conclusions Considering Cultural Aspects of Technology Usage

In regard to our Research Questions we have shown various ways new types of Web 2.0 technology can be employed to enhance CALL. First this needs to be done by making online reading more accessible to readers of all language backgrounds by adding instant access glossing (both bilingual and monolingual) and translation engines to all sites, along with listening support and summarization tools. Secondly, we have suggested some ways that interactive, mobile educational and social networks can be used to more effectively and enjoyably bring about the transformation and improvement of language learning promised by Web 2.0 and 3.0 technologies. Finally, we have demonstrated at our large Virtual Language Education site how to more fully integrate and apply this new technology to enhance vocabulary, reading and language learning. Naturally, Web 2.0 technologies can be used in the many ways shown by Loucky (2008), to contribute to the development of both intensive and extensive reading, and all four communication skills in any language available online.

In previous studies Loucky (2008) found Engineering students in Japan were generally more open to the use of technology and more adept at using electronic dictionaries effectively than typical humanities students of English in Japan, including English majors (Loucky, 2003b). No distinctions between male and female participants were found, although few females tend to major in Engineering in Japan. What relationship could we find between Chinese and Japanese students and their English language learning in this study? What common problems do they have with reading or writing in English? How can technology enhance reading strategies in Asian contexts, such as Japan and Taiwan, where our collaborative writing exchanges were done? These students do seem to be more open to using technology in the classroom than other students because of their society's normalization of technology, and due to having higher computer and English literacy than average Japanese lower level learners.

Using such digital devices as e-readers and mobile phones with Internet connectivity can enable students to gain better access to reading materials in the classroom or for mobile online learning. Such Web 2.0 and emerging Web 3.0 technology promises to revolutionize reading, especially as language learning becomes more mobile/portable, user-generated and –controlled. Language learning sites that enable users to download content directly to their portable or desktop devices should enhance out of class, independent language learning and use. So far, though, Japanese students' use of mobile phones for reading and vocabulary learning has been too expensive, slow or hard to keep on task (Loucky, 2003c). iPods could greatly enhance extensive reading with listening support if text and sound files could be easily downloaded simultaneously by users at different speeds and levels of text difficulty (as our students experimented with using the learn.com site). In Japan, for example, the DoCoMo cell phone service already offers downloadable novels and Manga to mobile phones.

Among the aspects of online learning course design to take into consideration in future online course development are these:

1. How to ensure that the website's purposes and learning objectives are clear to both students and teachers using them.
2. What are implications for learners' workload (how can blended in-class use help increase actual communication, learning and motivation while decreasing time they must spend working alone).
3. What are implications for teachers' workloads? (How can CALL help to decrease teachers' "take-home work," or enable them to even communicate or give feedback from home or office between infrequent classes?).
4. How can we ensure that end-users' online learning experiences are "of a seamless whole that incorporates all aspects of the online experience (conferencing, library, student and tutor homepage, etc.)" (Shield & Kukulska-Hume, 2004, p. 32), and better blend these together with other aspects of in-class or take-home integrated four-skills communicative language learning?

This study and website suggest initial answers to these questions, and could serve as a useful model for EAP/ESP/ETP online courses, as well as for blended reading courses to consider. It helps to advance an integrated model of how language learning websites can be better designed for blended in-class and mobile use, so that more enjoyable and effective language learning can take place, helping students to improve their vocabulary and reading skills online, as well as other communication skills interactively, face-to-face off-line.

REFERENCES

Akbulut, Y. (2006). Foreign language reading through hypermedia: Predictors of vocabulary learning and reading comprehension. *6th International Educational Technology Conference*, Famagusta, Northern Cyprus, 19-21 April, 43-50.

Bamford, J., & Day, R. R. (1997). *Extensive Reading: What is it? Why bother?* Retrieved on June 6, 2008 from: http://www.jalt-publications.org/tlt/files/97/may/extensive.html

Bryant, L., & Wilcox, D. (n.d.). *Some Lessons from Web 2.0 for Participation and E-democracy*. Retrieved September 9, 2008, from http://www.headshift.com/archives/002787.cfm.

Burston, J. (2007). WordChamp. *CALICO Journal*, *24*(2), 473–486.

Chun, D. (2006). CALL technologies for L2 reading. In L. Ducate & N. Arnold (Eds.), *Calling on CALL: From theory and research to new directions in foreign language teaching. CALICO Monograph Series, 5*, 69-98.

Coll, J. F. (2002). Richness of semantic encoding in a hypermedia-assisted instructional environment for ESP: Effects on incidental vocabulary retention among learners with low ability in the target language. *ReCALL*, *14*(2), 263–282. doi:10.1017/S0958344002000526

Devo, J. (2008). *Web 2.0: Fortune favours the brave*. Retrieved September 9, 2008, from http://www.computingbusiness.co.uk/computing-business/analysis/2210094/fortune-favours-brave-3771107

Dignan, L., Perlow, J., & Steinert-Threlkeld, T. (2007). *Tim Berners-Lee: From World Wide Web to giant global graph*. Retrieved May 10, 2008, from http://blogs.zdnet.com/BTL/?p=7126

Ekwall, E. E. (1976). *Diagnosis and remediation of the disabled reader*. Boston: Allyn & Bacon.

Grabe, W. (2004). Research on teaching reading. *Annual Review of Applied Linguistics*, *24*, 44–69. doi:10.1017/S0267190504000030

Loucky, J. P. (1994). Teaching and testing English reading skills of Japanese college students. *KASELE Kiyo*, *22*, 29–34.

Loucky, J. P. (1996). *Developing and testing vocabulary training methods and materials for Japanese college students studying English as a foreign language*. Ed.D. Thesis, Pensacola Christian College, Pensacola, FL.

Loucky, J. P. (1997). Summary of "Developing and testing vocabulary training methods and materials for Japanese college students studying English as a foreign language." *Annual Review of English Learning and Teaching 2*, JACET Kyushu-Okinawa Chapter, 15-36.

Loucky, J. P. (2002). Assessing the potential of computerized bilingual dictionaries for enhancing English vocabulary learning. In P. N. D. Lewis (Ed.), *The changing face of CALL: A Japanese perspective* (pp. 123-137). Lisse: Swets & Zeitlinger.

Loucky, J. P. (2003a). Testing vocabulary levels in Japan, Part II. *The Japanese Learner, 29*, (pp. 15-20). Oxford: Oxford University.

Loucky, J. P. (2003b). Improving access to target vocabulary using computerized bilingual dictionaries. *ReCALL*, *14*(2), 293–312.

Loucky, J. P. (2003c). Using computerized bilingual dictionaries to help maximize English vocabulary learning at Japanese colleges. *CALICO Journal*, *21*(1), 105–129.

Loucky, J. P. (2005). Combining the benefits of electronic and online dictionaries with CALL Web sites to produce effective and enjoyable vocabulary and language learning lessons. *Computer Assisted Language Learning*, *18*(5), 389–416. doi:10.1080/09588220500442764

Loucky, J. P. (2006a). Maximizing vocabulary development by systematically using a depth of lexical processing taxonomy, CALL resources, and effective strategies. *CALICO Journal*, *23*(2), 363–399.

Loucky, J. P. (2006b). Developing integrated online English courses for enjoyable reading and effective vocabulary learning. In *The Proceedings of JALT CALL 2005, Glocalization: Bringing people together.* Ritsumeikan University, Japan, June 3-5, 2005, 165-169.

Loucky, J. P. (2008). *CALL4 Internet Site*. Retrieved April 10, 2008, from http://www.call4all.us

Miller, P. (2005a). *Web 2.0: Building the new library*. Retrieved August 24, 2008, from http://www.ariadne.ac.uk/issue45/miller/

Nielsen, J., Schemenaur, P. J., & Fox, J. (1994). *Writing for the Web.* Retrieved June 7, 2008, from http://www.sun.com/980713/webwriting/

O'Reilly, T. (2004). *An architecture of participation*. Retrieved May 10, 2008, from http://www.oreillynet.com/pub/a/oreilly/tim/articles/architecture_of_participation.html

O'Reilly, T. (2005). Web 2.0: Compact Definition? *O'Reilly Radar blog.* Retrieved June 7, 2008, from http://radar.oreilly.com/archives/2005/10/web-20-compact-definition.html

O'Reilly, T. (2006). *Web 2.0 Compact Definition: Trying Again*. Retrieved January 20, 2007, from http://radar.oreilly.com/archives/2006/12/web-20-compact-definition-tryi.html

OLPC. (2008). *One Laptop per Child Foundation*. Retrieved April 10, 2008, from: http://laptop.org

Palmer, H. E. (1968). *The scientific study and teaching of languages*. Oxford: Oxford University Press.

Raygor, A. L., & Raygor, R. D. (1985). *Effective reading: Improving reading rates and comprehension*. McGraw-Hill Communication Skills Series: New York.

Reading Renaissance. (2005). Retrieved October 15, 2007, from http://www.renlearn.com

Robb, T. (2008). *Reading Site*. Retrieved May 10, 2008, from http://www.lclark.edu/~krauss/toppicks/reading.html

Shield, L., & Kukulska-Hume, A. (2004). Language learning websites: Designing for usability. *TEL & CAL: Zeitschrift für Neue Lernkulturen*, Januar, 27-32.

Spivack, N. (2006). *The third-generation web is coming*. Retrieved December 15, 2007, from http://www.kurzweilai.net/meme/frame.html?main=/articles/art0689.html?m%3D3

Taylor, D. (2005). *Writing for the Web: A comprehensive overview*. Retrieved October 15, 2007, from http://www.peakwriting.com/article.php?articleid=28

Walker, S., Schloss, P., Fletcher, C. R., Vogel, C. A., & Walker, R. C. (2007). *Visual-syntactic text formatting: A new method to enhance online reading*. Retrieved October 15, 2007, from http://www.readingonline.org/articles/art_index.asp?HREF=/articles/r_walker/

Wikipedia (2008a). *Web 3.0*. Retrieved May 10, 2008, from http://en.wikipedia.org/wiki/Web_3

Wikipedia (2008b). *Web 2.0*. Retrieved May 10, 2008, from http://en.wikipedia.org/wiki/Web_2

Willis, J. (1996). *A framework for task-based learning*. Harlow, UK: London.

KEY TERMS

Degrees of Reading Power (DRP): Touchtone Applied Science Associates' (TASA) Depth of Reading Power (DRP) program. TASA Literacy Online uses a scale of 0-100 in their measure of text and student reading level. They call these levels Degrees of Reading Power (DRP). Primary and Standard DRP tests assess learners' ability to comprehend surface meaning of prose, whereas Advanced tests assess the inferential and global reading skills of more proficient readers. DRP technology relies on the close link between text difficulty or readability level and comprehension test results. As such, they can be interpreted as criterion-referenced tests, indicating what a particular student can actually do.

Degrees of Word Meaning (DWM): TASA has designed and used tests of vocabulary in context called Degrees of Word Meaning (DWM). This vocabulary level testing scheme provides a brief Conversion Table, which helps teachers convert these DWM vocabulary level scores into an estimated size of students' reading vocabularies. Degrees of Word Meaning scores range from 850 (the equivalent to knowing over 157,000 words), to less than 300 (indicating that such a test taker knows 100 or fewer English words). Their products for educational assessment are include tests as well as online programs and steps for estimating both reading levels and readability of any text or book.

Digital Rights Management (DDR): An umbrella term that refers to access control technologies used by publishers and copyright holders to limit usage of digital media or devices. It may also refer to restrictions associated with specific instances of digital works or devices. DRM overlaps with software copy protection to some extent, however the term DRM is usually applied to creative media (music, films, etc.) whereas the term "copy protection" tends to refer to copy protection mechanisms in computer software.

Extensive Reading: This approach to reading is used when encouraging students to read widely, especially outside of class, at their Independent or Free Reading Level. Extensive reading is also known as pleasure reading, since its purpose is free, independent reading that is not overly dependent upon either teacher or dictionary.

Frustration Level: Learner recognizes less than 90% of running text. Comprehends under 50% of text. Such texts should either be totally avoided, unless working online with bilingual glossing available. Ideally language learners should also have fully bilingualized lexicons, concordancer and listening support available for any texts at less than Independent Level.

Intensive Reading: This approach to reading is used when intentionally teaching and practicing reading skills in classes or doing assignments out of class that require reading at one's Instructional Level, which may be from ½ to 2 years above free or Independent Level.

Independent Reading Level: Learners recognize 98-100% of words in text. Comprehend at better than 90%, so they can read such texts freely on their own.

Instructional Reading Level: Learners recognize 95-97% of words in text. Comprehends ideally at least 75%.

Online Language-Supported Manageable Text (OLSM Text): This refers to text not yet at a language learner's Independent Level, but made manageable via online tools such as fully bilingualized lexicons, concordancer and listening support. Levels might range as follows: learners may recognize 90-95% of words in such texts and comprehend ideally at about 75-89%, although with harder texts comprehension levels may fall between 51-74%.

Readability: Readability is an assessment of how easy a text is to understand for a given population. Online text readability includes four distinct constructs: 1) the reading ability or level of the user, 2) the readability level of a text, 3) its vocabulary level, and finally 4) readability assessment tests, instrument scales or indices themselves.

This work was previously published in Handbook of Research on Web 2.0 and Second Language Learning, edited by J. Wang, pp. 385-410, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.14
Querying Web Accessibility Knowledge from Web Graphs

Rui Lopes
LaSIGE, University of Lisbon, Portugal

Luís Carriço
LaSIGE, University of Lisbon, Portugal

ABSTRACT

Web Accessibility is a hot topic today. Striving for social inclusion has resulted in the requirement of providing accessible content to all users. However, since each user is unique, and the Web evolves in a decentralized way, little or none is known about the shape of the Web's accessibility on its own at a large scale, as well as from the point-of-view of each user. In this chapter the authors present the Web Accessibility Knowledge Framework as the foundation for specifying the relevant information about the accessibility of a Web page. This framework leverages Semantic Web technologies, side by side with audience modeling and accessibility metrics, as a way to study the Web as an entity with unique accessibility properties dependent from each user's point of view. Through this framework, the authors envision a set of queries that can help harnessing and inferring this kind of knowledge from Web graphs.

INTRODUCTION

Since its inception, the Web has become more and more prolific in people's lives. It is used as an information source, both one-way (e.g., newspapers) and two-way (e.g., blogging, forums, or even instant messaging). New Web sites and new content are produced and published each second by both professionals and amateurs, each one with different usability and accessibility quality marks. This fact, in conjunction with the Web's decentralized, yet highly connected architecture, puts challenges on the user experience when interacting and navigating between Web sites.

At the same time, the attractiveness of the Web brings more users to use it on a regular basis. This means that user diversity will be closer to real life where both unimpaired and impaired users coexist. Since each user has its own specific requirements, (dis)abilities, and preferences, their experience is different for each one, resulting in different satisfaction levels. In the same line of user diversity, device prolificacy and Internet connection ubiquity also contribute to the range of possible user experiences on interacting with the Web and, consequently, also have a stake in accessibility issues.

For all these reasons, the shape of the Web itself deeply influences each user's interactive experience in different ways. Users tend to navigate through the Web by avoiding Web sites that cannot be rendered correctly, which provide poor interactive capabilities for the specificities of the user or the device she/he is using to access the Web, reflecting negatively on users' experience. Therefore, it is required to understand the Web's graph of Web pages at a large scale from the point-of-view of each individual's requirements, constraints and preferences, and grasp this information to devise future advancements on Web standards and accessibility-related best practices. The inability to adapt the Web, its standards, technologies, and best practices will pose severe problems on the society in general, by leaving untouched the barriers towards a proper e-inclusion level that can actually cope with everyone, independently of impairments and related needs.

The main contributions of this Chapter are: (1) the establishment of a Web accessibility framework that can be used to create complex knowledge bases of large scale accessibility assessments; and (2) a set of query patterns to infer critical aspects of the accessibility of Web graphs with a fine-grained control (based on users' requirements and constraints). The proposed framework and the set of query patterns will form a core tool that helps analyzing the semantics of the accessibility of Web graphs. Next, we describe the relevant background work on Web accessibility and knowledge extraction from Web graphs.

BACKGROUND

Two main research topics have influence and contribute to the study of Web accessibility on large scale: the analysis of accessibility compliance of a Web page (or Web site), and the analysis of the Web's graph structure.

The Web Accessibility Initiative (WAI, n.d.) of the World Wide Web Consortium (W3C, n.d.) has strived for setting up the pace of Web Accessibility guidelines and standards, as a way to increase accessibility awareness to Web developers, designers, and usability experts.

The main forces of WAI are the Web Content Accessibility Guidelines, WCAG (Chisholm et al., 1999). WCAG defines a set of checkpoints to verify Web pages for specific issues that have impact on accessibility of contents, such as finding if images have equivalent textual captions. These guidelines have been updated to their second version (Caldwell et al., 2008) to better handle the automation of accessibility assessment procedures, thus dismissing the requirement of manual verification of checkpoint compliance.

Until recently, the results of accessibility assessment were presented in a human-readable format (i.e., Web page). While this is useful for developers and designers in general, this is of limited use for comparison and exchange of assessment results. Therefore, WAI has defined EARL, Evaluation and Report Language (Abou-Zahra, 2007), a standardized way to express evaluation results, including Web accessibility evaluations, in an OWL-based format (Dean & Schreiber, 2004).

EARL affords the full description of Web accessibility assessment scenarios, including the specification of *who* (or *what*) is performing the evaluation, the *resource* that is being evaluated, the *result*, and the *criteria* used in the evaluation.

However, EARL does not provide constructs to support the scenarios envisioned in macro scale Web accessibility assessments. It cannot cope with metrics (thus dismissing quantification of Web accessibility) and with the Web's graph structure. This way, EARL becomes limited to single Web page qualitative evaluations.

Lopes & Carriço (2008a) have shown that current Web accessibility practices are insufficient to cope with the whole spectrum of audiences (both disabled and unimpaired users), and that any user can influence everyone's interactive experience on the Web (especially regarding accessibility issues). As Kelly et al. (2007) have predicted, to cope with every user, holistic approaches to Web accessibility have to be taken into account. This includes tailoring of accessibility assessment procedures to each individual's characteristics, as thoroughly discussed by Vigo et al. (2007b).

Generalizing the concept of accessibility to all users (and not just to those that deeply depend on it – i.e., people with disabilities), the adequacy of user interfaces to each user's requirements, limitations, and preferences is the ultimate goal of *Universal Usability*, as defined by Shneiderman (2000). As detailed by Obrenovic et al. (2007), one has to take into account users, devices, and environmental settings when studying accessibility in a universal way. However, to our knowledge, there is no work on how to measure the universal usability quality of a single Web page, from the perspective of a unique user (*per definition* of universal usability).

When scaling up to the size of the Web, other aspects of analysis have to be taken into account. The characterization of the Web (e.g., its size, analysis metrics, statistics, etc.) is a hot topic today. *Web Science* is emerging as a discipline that studies the Web as a dynamic entity, as described by Berners-Lee et al. (2006). It is centered on how infrastructural requirements, application needs, and social interactions depend and feed each other in the Web ecology (Hendler et al., 2008).

At a more fundamental level, one of the core aspects of studying the Web concerns on how it is universally usable, as hypothesized and defended by Shneiderman (2007). However, since this discipline is fairly new, little is know about the Web from a universal usability point-of-view. It is known that the evolution of Web standards has influence on the way users navigate and interact with the Web (Weinreich et al., 2006), but not to what extent and what is the impact on each individual's characteristics. By having a proper characterization of the Web's graph from each individual's point of view (i.e., requirements, needs, constraints, preferences), more complex studies can be preformed at higher abstraction levels, such as in-depth Social Network Analysis (cf. Berger-Wolf & Saia, 2006) and other types of social studies.

In Lopes & Carriço (2008b) the authors presented a mathematical model to study universal usability on the Web. It supports the analysis of the Web from the point-of-view of each user's characteristics, and explains how the Web's structure influences user experience. While the authors have hypothesized how this model can be used to observe the evolution of the Web, it just provides a theoretical framework for the analysis of accessibility. Nevertheless, this model provides interesting contributions on how the query patterns presented in this Chapter should be formulated.

WEB ACCESSIBILITY KNOWLEDGE FRAMEWORK

In order to open the way to querying different Web accessibility properties from Web graphs, we have defined a supportive knowledge framework. This framework groups four different components, as depicted in Figure 1: *Web Graphs*, *Web Accessibility Assessment*, *Audiences*, and *Metrics*. The framework has been design according to the following requirements:

- *Universal*. The framework should not be limited to "traditional" accessibility audiences (such as people with visual impairments), but cope with different kinds of accessibility-prone issues, such as limited interaction devices (e.g., mobile phones), or adversary environment settings (e.g., poor lighting settings). The universality concept can (and, in fact, should) be also extended to all users and usage situations, thus allowing knowing the impact of Web accessibility and similar universal usability issues on any user.

- *Generalized*. The framework must not impose *a priori* any limitation or bias towards particular accessibility assessment concepts. It should define them at a meta-level, in order to be possible to define query patterns that are independent from particular instances (e.g., a query pattern depends on user characteristics, not on *a* user characteristic).
- *Extensible*. Since the accessibility assessment procedures change (mostly to enforce better analyses), the framework should support the application of different procedures.
- *Fine-grained*. As discussed earlier, current accessibility evaluation practices are black-boxed, leading to having just a general view of evaluation results. The framework should support fine-grained analyses, to support studying accessibility from the perspective of different audiences.
- *Scalable*. The framework should not impose limits to the size and complexity of encoded information (i.e., knowledge base).

Each component is defined through a specific OWL-based vocabulary, as the inclusion of already existing ontologies (mostly specified in OWL) lowers the burden of defining each component of the framework. Accordingly, we have developed this framework by extending the EARL ontology to support the elicited requirements. Next, each component of the framework is described in more detail. For details about the namespace prefixes used in the next Sections and their corresponding URI mappings, please consult the Appendix. Throughout this Section we will provide examples on how to describe accessibility knowledge based on the Notation 3 (N3) syntax (Berners-Lee, 2006).

Web Graphs

The first component in the framework relates to the specification of Web graphs. The goal of this component is to represent each Web page as a single resource, as well as its corresponding hyperlinking structure. Figure 2 presents the concepts that support the specification of Web graphs.

The main subject of constructing Web graphs is the Web page. Since the EARL specification only supports the specification of subjects that are available on the Web (*earl:Content* class), we have further refined the concept to limit its scope just to Web pages (the core subject of accessibility assessment procedures), through the *ev:Webpage* class.

Figure 1. Web accessibility knowledge framework

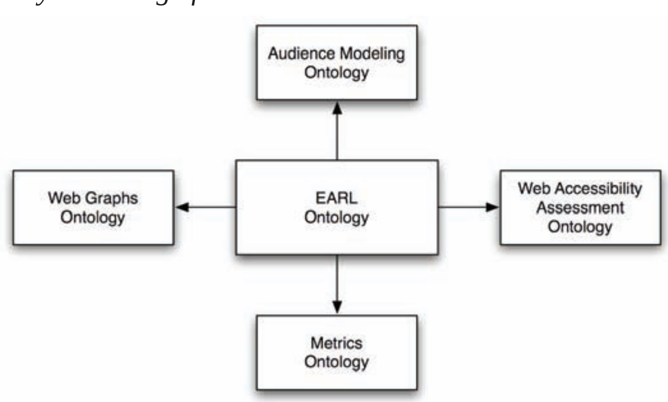
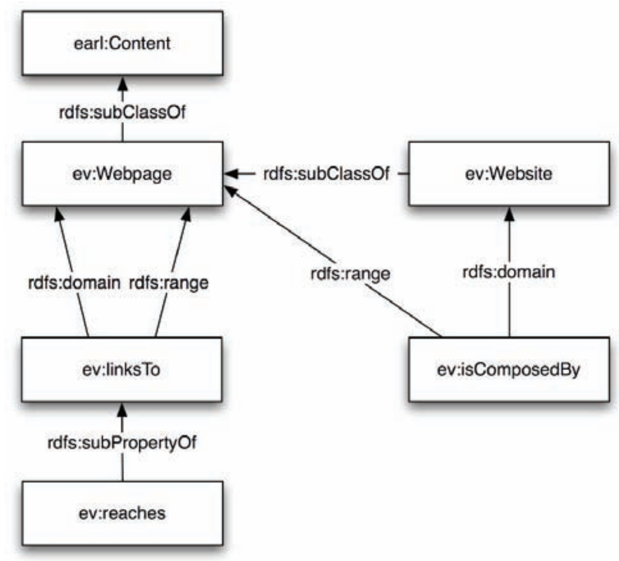


Figure 2. Web graphs ontology



Other types of content, such as images and CSS stylesheets (Bos, Çelik, Hickson, & Lie, 2007), were considered inherent of each Web page, from the perspective of evaluation procedures.

Two main properties (and their inverse) were defined to specify hyperlinks. The first, *ev:linksTo* (and its corresponding inverse property, *ev:islinkedBy*) establishes the direct relationship between two Web pages. The second property, *ev:reaches* (and its inverse, *ev:isReachedBy*), extends *ev:linksTo* with a transitive characteristic. This way, it becomes possible to query Web graphs from the perspective of reachability between two (or more) Web pages, not just on direct linking properties. This property will only afford knowing whether two Web pages are indirectly connected, leaving outside of the scope the number of links in between them. We have opted to explicitly define inverse properties, to afford the specification of queries that are more expressive and closer to natural language. To complement these constructs, we have specified the *ev:Website* class that, in conjunction with the *ev:isComposedBy* property (and its inverse, *ev:composes*), affords the direct specification of which Web pages belong to the same Website. To support out-of-the-box the specification of hyperlinking structure for Web sites, we have defined that *ev:Website* extends the *ev:Webpage* concept. However, the ontology cannot enforce the semantics that if two Web pages are linked, then their corresponding Web sites are also linked. Hence, we have devised two rules in SWRL (Horrocks et al., 2004) to afford linking scenarios, as presented next:

```
ev:linksTo(?website1, ?website2) =>
    ev:isComposedBy(?website1, ?webpage1) &
    ev:isComposedBy(?website2, ?webpage2) &
    ev:linksTo(?webpage1, ?webpage2)

ev:reaches(?website1, ?website2) =>
    ev:isComposedBy(?website1, ?webpage1) &
    ev:isComposedBy(?website2, ?webpage2) &
    ev:reaches(?webpage1, ?webpage2)
```

Next, we present a small example of how to define Web graphs, formally expressed (in the N3 format):

```
@base <http://example.com/>.

<b.html> a ev:Webpage.
<c.html> a ev:Webpage.
<a.html> a ev:Webpage;
    ev:linksTo <b.html>;
    ev:linksTo <c.html>.

<> a ev:Website.
<> ev:isComposedBy <a.html>;
 ev:isComposedBy <b.html>;
 ev:isComposedBy <c.html>.
```

Web Accessibility Assessment

The essential aspects for accessibility assessment results concern the description of the tests and their resulting outcome of applying them to a Web page. Consequently, the EARL ontology affords an extensible way of describing Web accessibility assessment results, in the form of *earl:Assertion* predicates. This includes, amongst other predicates, the specification of which test is being applied (i.e., *earl:TestCase*) and what is the result of its application to the Web page that its being evaluated (i.e., *earl:TestResult*).

In the second component of our framework, we have extended the EARL predicates for accessibility assessment by refining test cases (i.e., *earl:TestCase*) with appropriate semantics about the nature of the tests, regarding the different technologies used in Web pages. This will afford the fine-grained analysis of Web pages according to technological criteria, as depicted by the concepts in Figure 3.

The main predicates for describing the nature of the tests are: *ev:TestContent*, for the specification of tests applied to the actual contents (in different media) of Web pages; *ev:Structure*, for tests applied directly on the HTML structure itself; *ev:Style*, when testing styling properties (such as analyzing CSS); and *ev:Behavior*, to represent tests over scripts (e.g., Javascript).

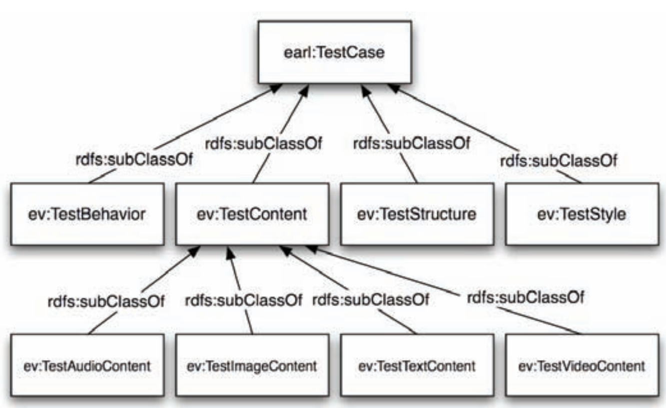
To better illustrate the usage of this ontology, we present next a classification of some WCAG 1.0 guidelines:

```
@prefix wcag10: <http://www.w3.org/TR/
WCAG10/#>.

wcag10:gl-color a ev:TestStyle.
wcag10:gl-structure-presentation a evTest-
Structure.
wcag10:gl-structure-presentation a evTest-
Style.
```

Audiences

We have defined a third component in our ontological framework to support the specification of audiences. This will ensure that different queries can be performed to a knowledge base of Web

Figure 3. Web accessibility assessment ontology



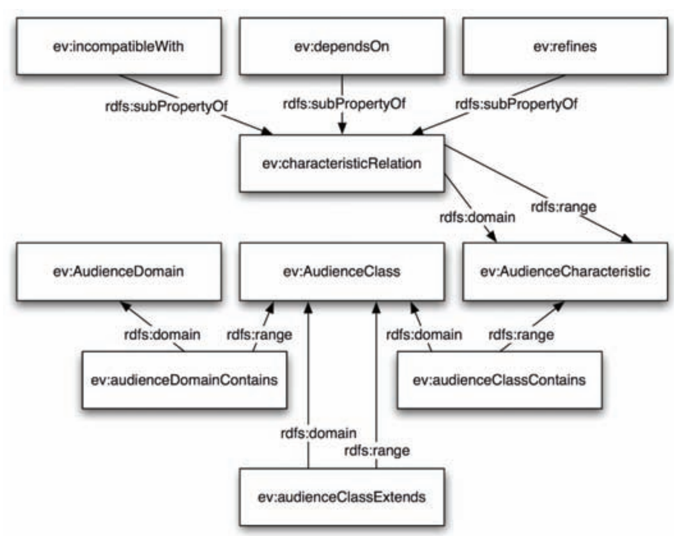
accessibility assessment according to the necessities and characteristics of different audiences. We based this support on earlier works on audience modeling, such as those described in Lopes & Carriço (2008a). Figure 4 depicts the complete ontological vocabulary to describe audiences (for simplicity, inverse properties are omitted).

The atomic concept in this vocabulary is *ev:AudienceCharacteristic*. Its purpose is to represent a single concept of an audience (e.g., a specific disability, a device characteristic, etc.). Since characteristics may represent concepts at different abstraction levels, they should be structured taxonomically. We introduce the *ev:refines* property (and its inverse, *ev:isRefinedBy*) to afford this expressivity.

The inherent nature of audiences raises the fact that they are often defined by several characteristics. Accordingly, this vocabulary introduces *ev:AudienceClass* as a way to represent them, along the side of the *ev:audienceClassContains* property (and its inverse, *ev:audienceCharacteristicIsContainedBy*) to map characteristic inclusion by an audience. However, since this association is merely syntactic, incoherent audiences might be described. To mitigate such issues we have introduced two additional concepts in the vocabulary. The first, *ev:dependsOn*, affords mapping dependencies between characteristics (such as total blindness depends on screen reader). The second, *ev:incompatibleWith*, allows the specification of incompatibilities between characteristics (e.g., total blindness is incompatible with screen). With these two properties, the semantics of audiences can be verified automatically. These properties, in conjunction with *ev:refines*, form the set of semantic relations that can be established between characteristics. Therefore, we introduce a generalization concept, *ev:characteristicRelation*, as an abstraction for the three concepts. This term affords inferring, e.g., if two characteristics have any kind of dependency between them.

While analyzing Web graphs from the perspective of a single audience can provide interesting results, the scope of such results is limited. It is often to perform comparative analyses of the results for a set of audiences. To support such scenarios, we have defined an audience aggregation concept, *ev:AudienceDomain*, to represent the domain of audiences that will be analyzed. The inclusion of an audience by a domain is represented through the *ev:audienceDomainContains* property (and corresponding inverse, *ev:audienceClassIsContainedBy*).

Figure 4. Audiences ontology



Lastly, we introduce in this vocabulary another concept to explore the synergies and differences between audiences, through the *ev:audienceClassExtends* property (and its *ev:audienceClassIsExtendedBy* counterpart). This extension mechanism is based on traditional object oriented modeling practices, i.e., an audience that extends another audience inherits its characteristics, thus creating parent-child relationships between audiences within a domain. Moreover, due to the fact that characteristics are taxonomically organized (through the *ev:refines* property), the characteristics of child audiences can be inferred and generalized to their common parent audience. A simple example follows, where a small taxonomy of characteristics is defined, and used in the definition of an audience domain.

```
@prefix tx: <http://taxonomy.com/>.
@prefix au: <http://audiences.com/>.

tx:characteristic a
ev:AudienceCharacteristic.
tx:disability a ev:AudienceCharacteristic.
tx:blind a ev:AudienceCharacteristic.
tx:totallyBlind a ev:AudienceCharacteristic.
tx:colorBlind a ev:AudienceCharacteristic.
tx:device a ev:AudienceCharacteristic.
tx:screen a ev:AudienceCharacteristic.

tx:disability ev:refines tx:characteristic.
tx:blind ev:refines tx:disability.
tx:totallyBlind ev:refines tx:blind.
tx:colorBlind ev:refines tx:blind.
tx:device ev:refines tx:characteristic.
tx:screen ev:refines tx:device.

tx:totallyBlind ev:incompatibleWith
tx:colorBlind.
tx:totallyBlind ev:incompatibleWith
tx:screen.
tx:colorBlind ev:dependsOn tx:screen.

au:domain1 a ev:AudienceDomain.

au:blind a ev:AudienceClass.
au:blind ev:audienceClassContains tx:blind.
au:domain1 ev:audienceDomainContains
au:blind.

au:totallyBlind a ev:AudienceClass.
au:totallyBlind ev:audienceClassContains
tx:totallyBlind.
au:domain1 ev:audienceDomainContains
au:totallyBlind.

au:colorBlind a ev:AudienceClass.
au:colorBlind a ev:audienceClassContains
tx:colorBlind.
au:colorBlind a ev:audienceClassContains
tx:screen.
au:domain1 ev:audienceDomainContains
au:colorBlind.

au:totallyBlind ev:audienceClassExtends
au:blind.
au:colorBlind ev:audienceClassExtends
au:blind.
```

However, affording the description of audience domains has limited applicability. To ensure that queries on Web graphs can be formulated based on the characteristics of audiences, there must be a mapping between audiences and accessibility assessment tests. Such vocabulary is synthesized in Figure 5.

In this vocabulary we have introduced a single property, *ev:requiresCharacteristic* (and its corresponding *ev:isRequiredByTest* counterpart), that maps a characteristic to a particular *earl:TestCase* instance. With this property, any audience or even entire domain can be mapped to the battery of tests that must be performed to a Web page, in order to obtain results tailored to these audiences. Dually, if the entire set of tests is performed over each Web page, their results can be queried from the perspective of different audiences or entire domains. An example follows, where the two previous examples are bound together. More spe-

Figure 5. Audience/test mapping sub-ontology

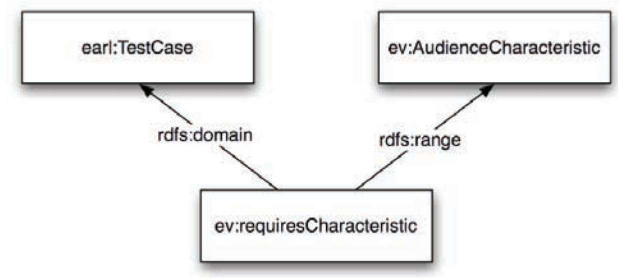
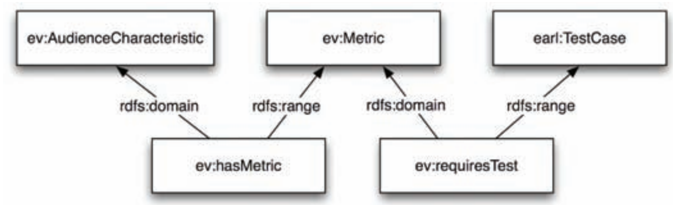


Figure 6. Metrics ontology



cifically, we map test cases to concrete audience characteristics that have been defined.

```
@prefix wcag10: <http://www.w3.org/TR/WCAG10/#>.
@prefix tx: <http://taxonomy.com/>.

wcag10:gl-color ev:requiresCharacteristic tx:colorBlind.
wcag10:gl-structurePresentation ev:requiresCharacteristic tx:totallyBlind.
```

Metrics

The last component in the framework concerns the specification of Web accessibility metrics, i.e., providing quantitative information about the accessibility of a Web page. Since different metrics can be applied to evaluation results, a supportive vocabulary for the specification of metrics must be extensible. This way, Web graphs can also be analyzed from the perspective of different metrics, thus allowing exploring which metric is better suited to different accessibility scenarios. Figure 6 depicts the vocabulary to support the specification of metrics.

The main concept in the metrics vocabulary is *ev:Metric*. Its purpose is to afford the specification of metrics that are applied to each Web page, based on the results of corresponding tests. While some metrics might be independent from specific tests, more concrete metrics can depend on the application of them. Therefore, we introduce the *ev:requiresTest* property to define dependency binds between metrics and tests (and its counterpart, *ev:isRequiredByMetric*). This property can be used, e.g., to specify consistency verification rules on the application of metrics, based on their semantics. Furthermore, by crossing this property with the *ev:requiresCharacteristic*, metrics can be mapped indirectly to audience characteristics. However, metrics can also be directly related to audience characteristics. This affords tying up specific quantification procedures to characteristics. Hence, we introduce an extra property in the vocabulary, *ev:hasMetric*, in order to support this type of scenarios. Next, we present a simple example on how to bind metrics with tests and characteristics.

```
@prefix wcag10: <http://www.w3.org/TR/WCAG10/#>.
@prefix m: <http://example.com/metrics#>.
@prefix tx: <http://taxonomy.com/>.

m:simpleMetric a ev:Metric;
    ev:requiresTest wcag10:gl-color;
    ev:requiresTest wcag10:gl-structurePresentation.

m:charMetric a ev:Metric.

tx:colorBlind ev:hasMetric m:charMetric.
```

We have introduced another concept on the vocabulary that is crucial to the specification of metrics. Each metric is supposed to have a concrete value, when applied to a Web page. Therefore, the vocabulary provides support to this feature through the *ev:hasMetricValue* datatype property, where metric values can be setup in the [0, 1] range (i.e., percentage). This way, since each metric does not yield an absolute value, Web graphs can be compared in the perspective of different metrics. With these constructs, the framework provides the support for specifying the resulting application of a given metric, in the context of an accessibility evaluation procedure. However, it is out of the scope of this Chapter to describe how these metrics are calculated (cf. Vigo et al., 2007a).

Consequently, since this property is abstract, concrete metrics properties must be derived from *ev:hasMetricValue* through subclassing. This extension to the metrics ontology is depicted in Figure 7.

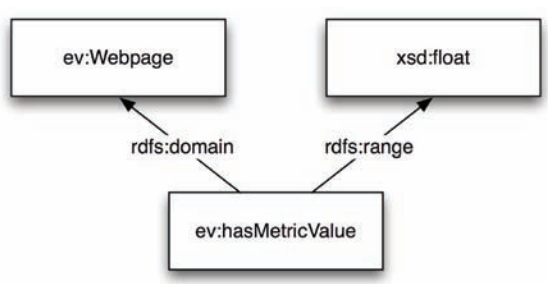
As a simple example, we present how to use this extension to the metrics ontology, by specifying a new datatype property, as well as its application in a concrete set of Web pages.

```
@prefix m: <http://example.com/metrics#>.
@base <http://example.com/>.
m:hasSimpleMetricValue rdfs:subPropertyOf ev:hasMetricValue.
<a.html> m:hasSimpleMetricValue 0.2.
<b.html> m:hasSimpleMetricValue 0.9.
<c.html> m:hasSimpleMetricValue 0.45.
```

However, with these constructs it is impossible to know what are the metric values associated with a specific characteristic or test case, for a given Web page. This happens due to *ev:Metric* instances are not automatically bound to datatype properties derived from *ev:hasMetricValue*. Consequently, query patterns cannot be created to explore complex mining scenarios, as each binding between metrics and datatype properties have to be artificially created on each query, which poses sever limitations on the generalization requirement for querying Web accessibility. In order to mitigate this situation, we defined another property, *ev:relatesToMetric* (and its counterpart, *ev:isRelatedToDatatypeProperty*), to draw both concepts together, as depicted in Figure 8.

Since we wanted to bind a datatype property directly to an *ev:Metric* instance, we had to import the OWL schema into our own ontology. This is due to the fact that, per definition, object properties bind class instances. Because *ev:hasMetricValue* (and subclassed datatype properties) are *owl:DatatypeProperty* instances, we circumvented this to afford the specification of richer and more complex query patterns that can remain agnostic to particular concepts or instances.

Figure 7. Metrics ontology extension



To exemplify the usage of this property, we have bound a metric instance to a particular datatype based on the previous examples, as shown next.

```
@prefix m: <http://example.com/metrics#>.

m:hasSimpleMetricValue  ev:relatesToMetric
m:simpleMetric.
```

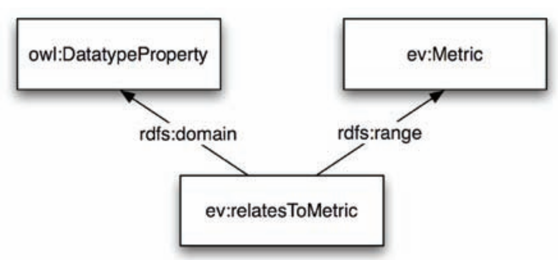
However, by setting *ev:relatesToMetric*'s domain to a generic OWL construct, one can bind metrics to any datatype property as there is no formal way to restrict the domain just to datatype properties derived from *ev:hasMetricValue*. To mitigate this issue, there must be an appropriate semantic enforcement through rules. The following SWRL rule affords this scenario:

```
ev:relatesToMetric(?datatypeProperty, ?metric) =>
    owl:subPropertyOf(?datatypeProperty,
ev:hasMetricValue)
```

QUERY PATTERNS

The extensions to the EARL ontology that we presented in the previous Section provide a comprehensive set of concepts that afford the full description of Web graphs from the perspective of Web accessibility and audience richness. This framework serves as the base ground for setting up Web graph knowledge bases that can be semantically queried in different forms. From the vast range of Semantic Web querying technologies, we opted to specify queries in the SPARQL language (Prud'hommeaux & Seaborne, 2008), as it is the *de facto* querying standard in the Semantic Web stack.

Figure 8. Metrics binding scheme



All examples in this Section will be based on the following SPARQL prefixes mapping:

```
PREFIX earl: < http://www.w3.org/ns/earl#>
PREFIX ev: <http://hcim.di.fc.ul.pt/ontologies/evaluation#>
PREFIX m: <http://example.com/metrics#>
PREFIX tx: <http://taxonomy.com/>
PREFIX au: <http://audiences.com/>
PREFIX wcag10: <http://www.w3.org/TR/WCAG10/#>
```

In this chapter, we have distinguished two different types of query patterns than can be applied to Web graphs: mining properties, and partitions extraction. Next, we describe each one of these pattern types.

Mining Web Site Properties

Web sites on their own can be analyzed from several perspectives. In this Section, we present some query patterns that can extract relevant information about a single Web page, as well as a set of Web pages perceived as one single entity (i.e., Web site). For practical purposes, all SPARQL patterns are applied to a dummy Web page (*http://example.com/a.html*) or Web site (*http://example.com*), whose semantics are marked as *ev:Webpage* and *ev:Website* instances, correspondingly; other instances that appear on queries are based on the examples presented in the previous Sections.

Metric thresholds. One of the simplest ways of verifying the accessibility of a single Web page relates to setting up quality thresholds. We have devised several query patterns for this purpose. The simplest query concerns a strict threshold that yields whether a Web page has a minimum quality level for a specific metric:

```
ASK {
    <http://example.com/a.html>
    m:hasSimpleMetricValue ?v.
    FILTER (?v >= 0.5)
}
```

Based on this pattern, one can generalize it for minimum and maximum boundaries, thus allowing checking if a Web page belongs to a particular quality cluster:

```
ASK {
    <http://example.com/a.html>
    m:hasSimpleMetricValue ?v.
    FILTER(?v >= 0.5 && ?v <= 0.75)
}
```

On the other hand, thresholds can be used to understand what metrics are above a certain value (or between two boundaries). Along these lines, the previous query pattern can be rewritten as:

```
SELECT ?metric
WHERE {
    <http://example.com/a.html> ?metricValue
    ?v.
    ?metricValue ev:relatesToMetric ?metric.
    FILTER (?v >= 0.5 && ?v <= 0.75)
}
```

Content quality. Another aspect that can be explored in a Web page deals with the semantics of the tests applied to it. In conjunction with metric value filtering, one can grasp the quality of a Web page based on the semantic categorization of test cases:

```
ASK {
    <http://example.com/a.html> ?metricValue
    ?v.
    ?metricValue ev:relatesToMetric ?metric.
    ?metric ev:requiresTest ?test.
    ?test rdfs:subClassOf
    ev:TestImageContent.
    FILTER (?v >= 0.5)
}
```

This query pattern can be extended to find out which types of test cases have an inherent quality above a given threshold (the DISTINCT query modifier has been used to remove duplicates):

```
SELECT DISTINCT ?testType
WHERE {
    <http://example.com/a.html> ?metricValue
    ?v.
    ?metricValue ev:relatesToMetric ?metric.
    ?metric ev:requiresTest ?test.
    ?test rdfs:subClassOf ?testType.
    FILTER (?v >= 0.5)
}
```

Characteristic quality. As explained earlier, characteristics can be bound to metrics. This feature of the framework allows the exploration of quality metrics similar to *metric thresholds*, but taking into account characteristics as the main feature to be analyzed:

```
ASK {
    <http://example.com/a.html> ?prop ?v.
    tx:colorBlind ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?v >= 0.5)
}
```

This pattern can be extended in order to leverage which characteristics have a quality level above a certain threshold:

```
SELECT ?char
WHERE {
    <http://example.com/a.html> ?prop ?v.
    ?char ev:hasMetric ?metric.
```

```
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?v >= 0.5)
}
```

By having a quality mark associated to characteristics, these can also be compared, to verify which one is better supported in a Web page. This can be directly achieved with the following query pattern:

```
ASK {
    <http://example.com/a.html> ?prop1 ?v1.
    <http://example.com/a.html> ?prop2 ?v2.
    tx:colorBlind ev:hasMetric ?metric1.
    tx:totallyBlind ev:hasMetric ?metric2.
    ?metric1 ev:isRelatedToDatatypeProperty
    ?prop1.
    ?metric2 ev:isRelatedToDatatypeProperty
    ?prop2.
    FILTER (?v1 > ?v2)
}
```

Furthermore, both patterns can be combined to extract which characteristics have a better quality than a predetermined one:

```
SELECT ?char
WHERE {
    <http://example.com/a.html> ?prop1 ?v1.
    <http://example.com/a.html> ?prop2 ?v2.
    tx:colorBlind ev:hasMetric ?metric1.
    ?char ev:hasMetric ?metric2.
    ?metric1 ev:isrelatedToDatatypeProperty
    ?prop1.
    ?metric2 ev:isrelatedToDatatypeProperty
    ?prop2.
    FILTER (?v1 > ?v2)
}
```

Audience quality. One of the important aspects discussed earlier pertains to knowing if a Web page has a certain degree of quality in what respects to a particular audience. The previous query pattern can be adapted to support this feature:

```
ASK {
    <http://example.com/a.html> ?prop ?v.
    au:totallyBlind ev:audienceClassContains
    ?char.
    ?char ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?v >= 0.5)
}
```

While this query pattern affords the explicit verification of the quality of a given audience, it is also relevant to explore and infer which audiences are supported in a Web page, with a given quality level. This pattern can be translated into SPARQL as:

```
SELECT ?audience
WHERE {
    <http://example.com/a.html> ?prop ?v.
    ?audience ev:audienceClassContains
    ?char.
    ?char ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?v >= 0.5)
}
```

Domain quality. In the same fashion as the previous patterns, one can obtain information about whether a domain is supported by a Web page or not, according to a specific threshold:

```
ASK {
    <http://example.com/a.html> ?prop ?v.
    au:domain1 ev:audienceDomainContains
    ?audience.
    ?audience ev:audienceClassContains
    ?char.
    ?char ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
```

```
    ?prop.
    FILTER (?v >= 0.5)
}
```

In the case where one wants to discover which domains are above a given threshold, the previous query pattern can be adapted in a simple way to cope with this requirement, as follows:

```
SELECT ?domain
WHERE {
    <http://example.com/a.html> ?prop ?v.
    ?domain  ev:audienceDomainContains
    ?audience.
    ?audience  ev:audienceClassContains
    ?char.
    ?char ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?v >= 0.5)
}
```

Website quality. While all of the previous patterns are targeted just to a single Web page, it is relevant to find out information about the set of Web pages from a unique entity point of view (e.g., a Web site). By exploring the Web graph ontology provided in the framework, Web sites can be analyzed as a single entity:

```
SELECT ?page
WHERE {
    ?page  ev:composes  <http://example.
    com>.
}
```

More complex patterns can be devised for Web sites, based on this pattern and the set of patterns presented above for Web pages. For instance, combining this pattern with characteristic quality analysis, Web sites can be analyzed to find out which ones are entirely accessible for any audience characteristic above a certain threshold:

```
SELECT ?site
WHERE {
    ?site ev:isComposedBy ?page.
    ?page ?prop ?v.
    ?char ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?v >= 0.5)
}
```

A variant on this query pattern can be defined as verifying if the average metric value is above the threshold, for a given characteristic. This would unify Web pages, thus analyzing a Web site as a single entity. However, SPARQL does not provide aggregation functions out of the box. Therefore, some implementations have circumvented this issue through, e.g., the AVG function. Without this function each metric value should be aggregated outside the query pattern and an average value calculation should be performed, which influences its scalability. Hence, this pattern uses the AVG function accordingly:

```
SELECT AVG(?v)
WHERE {
    <http://example.com>  ev:isComposedBy
    ?page.
    ?page ?prop ?v.
    tx:totallyBlind ev:hasMetric ?metric.
    ?metric:isRelatedToDatatypeProperty
    ?prop.
}
```

Semantically Extracting Web Graph Partitions

While capturing information about the accessibility of single Web pages or Web sites has value, it is more interesting to analyze Web graphs as a whole. The set of query patterns presented in the previous Section can be adapted to grasp new knowledge about entire Web graphs. In this

Section we present query patterns that afford the extraction of Web graph partitions according to accessibility criteria. Along the lines of the previous Section, all SPARQL patterns are applied to a set of dummy Web pages (e.g., *http://example.com/a.html*) or Web site (*http://example.com*), with the semantics of *ev:Webpage* and *ev:Website*, correspondingly; other instances that appear on queries are based on the examples presented in the previous Sections.

Reachability. The simplest information that can be obtained about a Web graph concerns its edges, i.e., the link structure. Edges are described through *ev:linksTo* property instances. The transitiveness of the *ev:reaches* property, based on *ev:linksTo*, allows the exploration of connectivity between Web pages (and between Web sites, as well). This query pattern will be used as the base support for extracting Web graph portions according to different accessibility semantics. Reachability can be a property queried between Web pages, e.g.:

```
ASK {
    <http://example.com/a.html>  ev:reaches
    <http://example.com/b.html>.
}
```

This notion can be extended to explore which Web pages can be reached from a starting point:

```
SELECT ?page
WHERE {
    <http://example.com/a.html>  ev:reaches
    ?page.
}
```

The opposite pattern, knowing which Web pages reach a specific ending point, can also be explored similarly:

```
SELECT ?page
WHERE {
    ?page ev:reaches <http://example.com/a.
    html>.
}
```

Lastly, based on these queries, Web graph portions can be extracted according to their linking structures. For these patterns, we use the CONSTRUCT query form provided in SPARQL. The simplest graph portion extraction concerns finding out the linking structure reached from a specific starting Web page:

```
CONSTRUCT {
    ?page ev:linksTo ?otherpage
}
WHERE {
    <http://example.com/a.html>  ev:reaches
    ?page.
    ?page ev:linksTo ?otherPage.
}
```

By generalizing this query pattern, the entire information about a particular Web graph portion can be extracted. While we could use the DESCRIBE query form, we opted to use CONSTRUCT since it is required to be supported in every SPARQL implementation. The query pattern is as follows:

```
CONSTRUCT {
    ?page ?prop ?value
}
WHERE {
    <http://example.com/a.html>  ev:reaches
    ?page.
    ?page ?prop ?value.
}
```

Lastly, all of these patterns can be further extended towards a macroscopic level, i.e., not centered on Web pages per se, but on Web sites. It is important to understand graph connectivity at this level, e.g. whether a Web site directly links to another one:

```
ASK {
    <http://example.com>  ev:isComposedBy
    ?page.
    <http://example2.com>  ev:isComposedBy
    ?page2.
    ?page ev:linksTo ?page2.
}
```

Based on this pattern, it might be relevant to understand what are the linking sources in such cases:

```
SELECT ?page
WHERE {
    <http://example.com>  ev:isComposedBy
    ?page.
    <http://example2.com>  ev:isComposedBy
    ?page2.
    <page ev:linksTo ?page2.
   }
```

Expanding further, one is able to find out which Web sites link directly to a given Web site:

```
SELECT ?site
WHERE {
    ?site ev:isComposedBy ?page.
    <http://example.com>  ev:isComposedBy
    ?page2.
    ?page ev:linksTo ?page2.
}
```

This type of pattern can be applied to all of the subsequent query patterns accordingly. For simplicity purposes, each one of the next query patterns is applied to Web pages.

Common characteristics. Based on the characteristics quality pattern for Web pages and Web sites, the same type of information can be acquired from entire Web graphs. Here, a quality threshold dictates which characteristics are above it in the entire Web graph:

```
SELECT ?char
WHERE {
    ?page ?prop ?v.
    ?char ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?v >= 0.5)
}
```

Based on this query pattern, Web graphs can be partitioned according to characteristic-oriented quality thresholds, following the same rules presented above:

```
CONSTRUCT {
    ?page ?prop ?value
}
WHERE {
    ?page rdf:type ev:Webpage.
    ?page ?prop ?value.
    ?char ev:hasMetric ?metric.
    ?metric ev:isrelatedToDatatypeProperty
    ?prop.
    FILTER (?value >= 0.5)
}
```

While this last pattern extracts the entire RDF graph, there are cases where just the corresponding Web graph structure (i.e., just the Web pages and linking structure) is extracted. In these cases the pattern can be easily adjusted as follows:

```
CONSTRUCT {
    ?page ev:linksTo ?otherPage.
}
WHERE {
    ?page ev:linksTo ?otherPage.
    ?page ?prop ?value.
    ?char ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?value >= 0.5)
}
```

Common audiences. The same type of query pattern can be applied to find out if a given Web graph is tailored to a specific audience:

```
ASK {
    ?page ?prop ?value.
    au:totallyBlind ev:audienceClassContains
    ?char.
    ?char ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?value >= 0.5)
}
```

Based on this query pattern, the Web graph itself can be partitioned according to this specific semantics:

```
CONSTRUCT {
    ?page ?prop ?value
}
WHERE {
    ?page rdf:type ev:Webpage.
    ?page ?prop ?value.
    au:totallyBlind ev:audienceClassContains
    ?char.
    ?char ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?value >= 0.5)
}
```

Characteristic reachability. This query pattern has been devised to find out which Web pages can be reached from a starting point, while maintaining a quality level above a specific threshold for a given characteristic:

```
SELECT ?page
WHERE {
    <http://example.com/a.html> ev:reaches
    ?page.
    ?page ?prop ?value.
    tx:totallyBlind ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?value >= 0.5)
}
```

However, the way this query pattern has been devised misses the intermediate Web pages that might not have the desired quality level for the selected characteristic. To mitigate this issue, all intermediate Web pages have to be verified accordingly:

```
SELECT ?otherPage
WHERE {
    ?page ev:linksTo ?otherPage.
    <http://example.com/a.html> ev:reaches
    ?page.
    <http://example.com/a.html> ev:reaches
    ?otherPage.
    ?page ?prop ?value.
    ?otherPage ?prop ?value2.
    tx:totallyBlind ev:hasMetric ?metric.
    ?metric ev:relatesToDatatypeProperty
    ?prop.
    FILTER (?value >= 0.5 && ?value2 >=
    0.5)
}
```

Accordingly, this pattern can be adapted to extract the corresponding Web graph portion. This is done by creating an RDF graph consisting of *ev:linksTo* derived triples, where both end-points have to be reached from the starting point, as follows:

```
CONSTRUCT {
    ?page ev:linksTo ?otherPage
}
WHERE {
    ?page ev:linksTo ?otherPage.
    <http://example.com/a.html> ev:reaches
    ?page.
    <http://example.com/a.html> ev:reaches
    ?otherPage.
```

```
    ?page ?prop ?value.
    ?otherPage ?prop ?value2.
    tx:totallyBlind ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?value >= 0.5 && ?value2 >=
    0.5)
}
```

This last version of the query pattern can be further adapted to find out just whether there are any Web pages that cannot be reached according to the devised semantics:

```
ASK {
    ?page ev:linksTo ?otherPage.
    <http://example.com/a.html> ev:reaches
    ?page.
    <http://example.com/a.html> ev:reaches
    ?otherPage.
    ?page ?prop ?value.
    ?otherPage ?prop ?value2.
    tx:totallyBlind ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypePropery
    ?prop.
    FILTER (?value < 0.5 && ?value2 < 0.5)
}
```

If one wants to know what Web pages are not reached through this method, the previous version of the query pattern can be further adapted. Please notice that this version of the pattern simply inverts the filter, in comparison with the second version of this query pattern:

```
SELECT DISTINCT ?page
WHERE {
    ?page ev:linksTo ?otherPage.
    <http://example.com/a.html> ev:reaches
    ?page.
    <http://example.com/a.html> ev:reaches
    ?otherPage.
    ?page ?prop ?value.
    ?otherPage ?prop ?value2.
    tx:totallyBlind ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypePropery
    ?prop.
    FILTER (?value < 0.5 && ?value2 < 0.5)
}
```

Audience reachability. As audiences are more closely representative of users (by aggregating characteristics), it is also important to study the graph reachability from this point of view. The simplest query pattern for audience reachability concerns finding out what Web pages are appropriate for a specific audience:

```
SELECT ?page
WHERE {
    <http://example.com/a.html> ev:reaches
    ?page.
    ?page ?prop ?value.
    au:totallyBlind ev:audienceClassContains
    ?char.
    ?char ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?value >= 0.5)
}
```

Like in characteristics reachability, one has to take into account that all Web pages in between must also have a quality level above the threshold that has been set. Accordingly, this query pattern must cope with this issue:

```
SELECT ?otherPage
WHERE {
    ?page ev:linksTo ?otherPage.
    <http://example.com/a.html> ev:reaches
    ?page.
    <http://example.com/a.html> ev:reaches
    ?otherPage.
    ?page ?prop ?value.
    ?otherPage ?prop ?value2.
    au:totallyBlind ev:audienceClassContains
    ?char.
```

```
    ?char ev:hasMetric ?metric.
    ?metric ev:relatesToDatatypeProperty ?prop.
    FILTER (?value >= 0.5 && ?value2 >= 0.5)
}
```

This pattern version can be easily adapted towards extracting the corresponding Web graph partition:

```
CONSTRUCT {
    ?page ev:linksTo ?otherPage
}
WHERE {
    ?page ev:linksTo ?otherPage.
    <http://example.com/a.html> ev:reaches ?page.
    <http://example.com/a.html> ev:reaches ?otherPage.
    ?page ?prop ?value.
    ?otherPage ?prop ?value2.
    au:totallyBlind ev:audienceClassContains ?char.
    ?char ev:hasMetric ?metric.
    ?metric ev:relatesToDatatypeProperty ?prop.
    FILTER (?value >= 0.5 && ?value2 >= 0.5)
}
```

It is also possible to build on this query pattern version to find out if there is any Web page that cannot be reached with at least the same quality level:

```
ASK {
    ?page ev:linksTo ?otherPage.
    <http://example.com/a.html> ev:reaches ?page.
    <http://example.com/a.html> ev:reaches ?otherPage.
    ?page ?prop ?value.
    ?otherPage ?prop ?value2.
    au:totallyBlind ev:audienceClassContains ?char.
    ?char ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypePropery ?prop.
    FILTER (?value < 0.5 && ?value2 < 0.5)
}
```

Likewise, we can also extract from the Web graph the set of Web pages that cannot be reached according to this semantics:

```
SELECT DISTINCT ?page
WHERE {
    ?page ev:linksTo ?otherPage.
    <http://example.com/a.html> ev:reaches ?page.
    <http://example.com/a.html> ev:reaches ?otherPage.
    ?page ?prop ?value.
    ?otherPage ?prop ?value2.
    au:totallyBlind ev:audienceClassContains ?char.
    ?char ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypePropery ?prop.
    FILTER (?value < 0.5 && ?value2 < 0.5)
}
```

Domain reachability. Along the lines of the previous two patterns, it is important to find out what partitions of a Web graph are reached from a starting point for all audiences within an audience domain, according to a previously set quality level threshold. The patterns for domain reachability follow closely the ones for characteristic and audience reachability. Therefore, we present a query pattern representative of the specific details for domain reachability. The following pattern affords the extraction of a Web graph partition for all the Web pages that are reachable from a starting point, based on a quality threshold:

```
CONSTRUCT {
    ?page ev:linksTo ?otherPage.
}
WHERE {
    ?page ev:linksTo ?otherPage.
    <http://example.com/a.html> ev:reaches
    ?page.
    <http://example.com/a.html> ev:reaches
    ?otherPage.
    ?page ?prop ?value.
    ?otherPage ?prop ?value2.
    au:domain1 ev:audienceDomainContains
    ?audience.
    ?audience ev:audienceClassContains
    ?char.
    ?char ev:hasMetric ?metric.
    ?metric ev:relatesToDatatypeProperty
    ?prop.
    FILTER (?value >= 0.5 && ?value2 >=
    0.5)
}
```

Another interesting pattern for domain reachability concerns finding out whether an audience domain has any audience that limits the reachability property:

```
ASK {
    ?page ev:linksTo ?otherPage.
    <http://example.com/a.html> ev:reaches
    ?page.
    <http://example.com/a.html> ev:reaches
    ?otherPage.
    ?page ?prop ?value.
    ?otherPage ?prop ?value2.
    au:domain1 ev:audienceDomainContains
    ?audience.
    ?audience ev:audienceClassContains
    ?char.
    ?char ev:hasMetric ?metric.
    ?metric ev:relatesToDatatypeProperty
    ?prop.
    FILTER (?value < 0.5 && ?value2 < 0.5)
}
```

This query pattern can be further converted to find out what are these audiences. This way, researchers can ask what are the specific audiences that limit reachability. This pattern is as follows:

```
SELECT ?audience
WHERE {
    ?page ev:linksTo ?otherPage.
    <http://example.com/a.html> ev:reaches
    ?page.
    <http://example.com/a.html> ev:reaches
    ?otherPage.
    ?page ?prop ?value.
    ?otherPage ?prop ?value2.
    au:domain1 ev:audienceDomainContains
    ?audience.
    ?audience ev:audienceClassContains
    ?char.
    ?char ev:hasMetric ?metric.
    ?metric ev:relatesToDatatypeProperty
    ?prop.
    FILTER (?value < 0.5 && ?value2 < 0.5)
}
```

Inward linking quality. As explained before, one of the great powers of the Web resides on how its linking structure is perceived and navigated by users. One important aspect of this property concerns the quality of the Web graph from the perspective of how Web sites are linked to each other. This query pattern explores linking to a specific ending point, i.e., all Web pages that link to a target Web page. First, it is important to extract the graph partition composed by the Web pages that point to it:

```
CONSTRUCT {
    ?page ev:linksTo <http://example.com/a.
    html>
}
WHERE {
    ?page ev:linksTo <http://example.com/a.
    html>.
}
```

Based on this simple query, quality thresholds can be set according to one of the query patterns presented in the previous Section (i.e., patterns for Web pages and Web sites), e.g., for characteristics:

```
CONSTRUCT {
    ?page ev:linksTo <http://example.com/a.
    html>
}
WHERE {
    ?page ev:linksTo <http://example.com/a.
    html>.
    ?page ?prop ?v.
    tx:colorBlind ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?v >= 0.5)
}
```

While this pattern query is interesting for extracting the Web graph based on a predetermined threshold, it is more important to extract it based on the quality of the target Web page. This query pattern can be further extended accordingly:

```
CONSTRUCT {
    ?page ev:linksTo <http://example.com/a.
    html>
}
WHERE {
    ?page ev:linksTo <http://example.com/a.
    html>.
    ?page ?prop ?v.
    <http://example.com/a.html> ?prop ?v2.
    tx:colorBlind ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?v >= ?v2)
}
```

Another aspect that can be explored based on this last version of the query pattern concerns knowing whether the target Web page has better quality than the Web pages that point to it. This allows us to understand if the target Web page can be perceived as an accessibility haven on navigation tasks:

```
ASK {
    ?page ev:linksTo <http://example.com/a.
    html>.
    ?page ?prop ?v.
    <http://example.com/a.html> ?prop ?v2.
    tx:colorBlind ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (?v < ?v2)
}
```

Outward linking quality. Dually to the previous query pattern, it is also important to understand the linking quality by setting up an initial starting Web page and explore the Web pages that it links to. The type of queries in this pattern follows closely the previous set of patterns with small changes. For instance, the following query leverages the Web graph partition of the Web pages that are safe to navigate:

```
CONSTRUCT {
    <http://example.com/a.html>  ev:linksTo
    ?page.
}
WHERE {
    <http://example.com/a.html>  ev:linksTo
    ?page.
    ?page ?prop ?v.
    <http://example.com/a.html> ?prop ?v2.
    tx:colorBlind ev:hasMetric ?metric.
    ?metric ev:isRelatedToDatatypeProperty
    ?prop.
    FILTER (v? >= ?v2)
}
```

Verticality. It is a fact that the Web is partially tailored to specific accessibility situations, e.g., “accessible versions” of a Web site. This property can be explored by studying the verticality of Web graphs. For example, given two different characteristics and a quality threshold, there might be an overlap between which Web pages are accessible to both. The amount of Web pages in this situation is directly related to the verticality of their corresponding partitions. This is done through the following query pattern:

```
SELECT ?page
WHERE {
    ?page ?prop1 ?value1.
    ?page ?prop2 ?value2.
    tx:colorBlind ev:hasMetric ?metric1.
    tx:totallyBlind ev:hasMetric ?metric2.
    ?metric1 ev:isRelatedToDatatypeProperty
    ?prop1.
    ?metric2 ev:isRelatedToDatatypeProperty
    ?prop2.
    FILTER (?value1 >= 0.5 && ?value2 >=
    0.5)
}
```

FUTURE TRENDS

The framework presented in this chapter is just one of the initial steps that can help understanding the impact of Web accessibility and Web standards on users, in a large scale (i.e., the *whole* Web) and with a fine-grained control over what aspects of Web accessibility and users are to be studied. We envision that semantic technologies can disrupt the way Web developers and designers think of accessibility and its social impact in the way users feed and consume information of the Web.

To grasp this knowledge, the framework we presented must be supported by its implementation and use in the analysis of large portions of the Web. Hence, we foresee that the following trends will help in this complex task:

- *Scalable architectures.* Building large scale Web accessibility observatories require scale-free approaches to crawl, store, process, and query the Web. We expect that with ongoing and future developments of scalable architectures that can cope with these type of tasks will help providing further insights on the influence that the Web's structure poses on Web accessibility issues.
- *Graph visualization algorithms.* There is a need for visualize large quantities of data (e.g., billions of metadata of Web pages), to grasp Web accessibility knowledge from semantic queries over Web graphs. Even when intelligent ways of extracting information from Web graph accessibility data, coping with billions of Web pages is not trivial. New graph visualization techniques can help lowering the burden of *finding the needle in the haystack*, i.e., the relevant information about the impact of Web accessibility at a large scale.
- *Automated verification.* Experts verify usability and accessibility problems in a manual/guided fashion. Since this approach is scale-bounded, there is the need for new automated verification procedures. With the advance in this research field (most probably with the aid of semantic technologies), more information can be obtained about usability and accessibility problems of the Web at a large scale. Significant advances to this challenge include understanding better how humans interact with computers, new models and theories for human psychology, as well as more pragmatic approaches such as statistical content analysis.
- *Metrics.* Accurate metrics provide better answers for finding the impact of Web accessibility implementation for all users. Having a base framework such as the one we presented in this chapter will help com-

paring metrics (and their corresponding application to Web graphs) and improve their accuracy.

- *Predictive and evolutionary models.* By having available smart models, the Web can be studied from predictive and evolutionary perspectives, opening the way to improving Web standards and Web accessibility assessment tools.

With advancements on these fronts, we foresee that the work described in this chapter can be put together within existing Web crawling, indexing and searching facilities with minor tweaks, forming an architecture for large scale Web accessibility assessments, as presented in Figure 9.

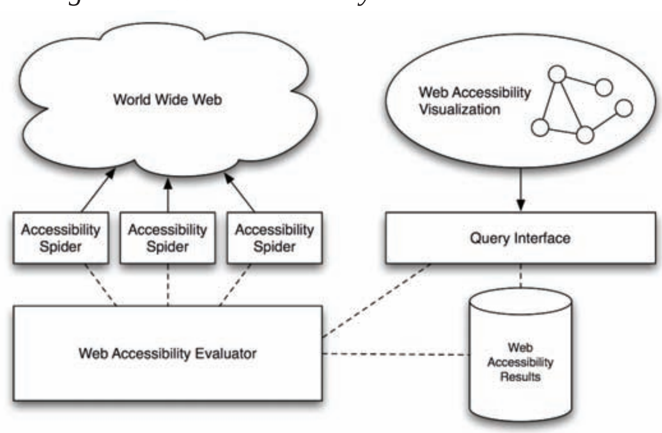
In this architecture the central aspect resides on the *Web accessibility results* repository, which should follow the metadata structures defined in this chapter. This repository holds all information about the accessibility semantics of the Web graph, as grasped by *Accessiblity Spiders* (similar to Web crawler's spiders) and an aggregating *Web accessiblity evaluator* module. Through the *Query Interface*, and the query patterns described in this chapter, we envision that this architecture will facilitate on visualizing Web accessibility at a large scale. We believe that this will provide clues on how Web standards and accessibility recommendations should evolve in the future towards a universally accessible and usable Web.

CONCLUSION

In this chapter we have presented a semantic knowledge framework for Web accessibility. This framework supports the definition of Web graphs and their accessibility properties. Through a set of query patterns, we have described a way to mine Web graphs in order to understand how the Web can cope with end users' intrinsic and transient characteristics, such as disabilities, interactive devices, etc.

We are currently developing ongoing work to implement this framework within the context of the architecture proposed in the previous Section in cooperation with the Portuguese Web Archive (PWA, n.d.) and apply it to study the entire Portuguese Web (around 40 million Web pages). We believe that the set of query patterns presented in this chapter will help us to understand the shape of the Web in what respects to its Web accessibility properties. More specifically, it will allow us discovering which Web sites are more accessible, and to verify if Web sites created by non-experts have significant accessibility problems, in comparison to those created by experts.

Figure 9. Architecture for large scale Web accessibility assessments



REFERENCES

Abou-Zahra, S. (2007). *Evaluation and report language (EARL) 1.0 schema*. Retrieved May 12, 2008, from http://www.w3.org/TR/EARL10-Schema/

Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N., & Weitzner, D. J. (2006). A framework for Web science. *Found. Trends Web Sci., 1*(1), 1-130.

Berners-Lee, T. (2006). *Notation 3*. Retrieved June 11, 2008, from http://www.w3.org/DesignIssues/Notation3

Berger-Wolf, T., & Saia, J. (2006). *A framework for analysis of dynamic social networks*. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, USA.

Bos, B., Çelik, T., Hickson, I., & Lie, H. (2007). *Cascading style sheets level 2 revision 1 (CSS 2.1) specification*. Retrieved May 12, 2008, from http://www.w3.org/TR/CSS21/

Caldwell, B., Cooper, M., Reid, L., & Vanderheiden, G. (2008). *Web content accessibility guidelines 2.0*. Retrieved June 10, 2008, from http://www.w3.org/TR/WCAG20/

Chisholm, W., Vanderheiden, G., & Jacobs, I. (1999). *Web content accessibility guidelines 1.0*. Retrieved June 10, 2008, from http://www.w3.org/TR/WAI-WEBCONTENT

Dean, M., & Schreiber, G. (Eds.). (2004). *OWL Web ontology language reference*. Retrieved June 11, 2008, from http://www.w3.org/TR/owl-ref/

Kelly, B., Sloan, D., Brown, S., Seale, J., Petrie, H., Lauke, P., & Ball, S. (2007). Accessibility 2.0: People, policies and processes. In *Proceedings of the 4th ACM International Cross-Disciplinary Conference on Web Accessibility*, Banff, Canada.

Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., & Weitzner, D. (2008). Web science: An interdisciplinary approach to understanding the World Wide Web. *Communications of ACM, 51*(7), 60-69. Advance online publication. Retrieved June 11, 2008, from http://webscience.org/documents/CACM-WebScience-Preprint.pdf

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004). *SWRL: A Semantic Web rule language combining OWL and RuleML*. Retrieved June 16, 2008, from http://www.w3.org/Submission/SWRL/

Lopes, R., & Carriço, L. (2008a). The impact of accessibility assessment in macro scale universal usability studies of the Web. In *Proceedings of the 5th ACM International Cross-Disciplinary Conference on Web Accessibility*, Beijing, China.

Lopes, R., & Carriço, L. (2008b). A model for universal usability on the Web. In *Proceedings of the WSW2008: Web Science Workshop*, Beijing, China.

Obrenovic, Z., Abascal, J., & Starcevic, D. (2007). Universal accessibility as a multimodal design issue. *Communications of the ACM, 50*(5), 83-88.

Prud'hommeaux, E., & Seaborne, A. (2008). *SPARQL query language for RDF*. Retrieved June 12, 2008, from http://www.w3.org/TR/rdf-sparql-query/

PWA. (n.d.). *Portuguese Web archive*. Retrieved June 11, 2008, from http://arquivo-web.fccn.pt/portuguese-web-archive-2?set_language=en

Shneiderman, B. (2000). Universal usability. *Communications of the ACM, 43*(5), 84-91.

Shneiderman, B. (2007). A provocative invitation to computer science. *Communications of the ACM, 50*(6), 25-27.

Vigo, M., Kobsa, A., Arrue, M., & Abascal, J. (2007a). Quantitative metrics for measuring Web accessibility. In *Proceedings of the 4th Interna-*

tional Cross-Disciplinary Conference on Web Accessibility, Banff, Canada.

Vigo, M., Kobsa, A., Arrue, M., & Abascal, J. (2007b). User-tailored Web accessibility evaluations. In *Proceedings of the 18th ACM Conference on Hypertext and Hypermedia*, Manchester, UK.

W3C. (n.d.). *World Wide Web Consortium*. Retrieved May 6, 2008, from http://www.w3.org

WAI. (n.d.). *Web accessibility initiative*. Retrieved June 10, 2008, from http://www.w3.org/WAI

Weinreich, H., Obendorf, H., Herder, E., & Mayer, M. (2006). Off the beaten tracks: Exploring three aspects of Web navigation. In *Proceedings of the 15th ACM International Conference on World Wide Web*, Edinburgh, Scotland.

KEY TERMS AND DEFINITIONS

Accessibility: The ability to access. Often tied to people with disabilities (e.g., total blindness), accessibility thrives to break the barriers to information access. We follow the strict sense of accessibility by embracing any situation where the ability to access information can be disrupted by device or even surrounding environment constraints.

Accessibility Guidelines: A set of best practices that must be followed by designers and developers when implementing software solutions (e.g., Web site) that will help on providing accessible information. By being guidelines, it should not be assumed that content is accessible just by following them.

Checkpoint: A concrete verification task that materializes a (part of a) guideline. Checkpoints can be fully automated if application technology provides corresponding support (e.g., verifying if all images have associated textual captions).

Metric: A quantification procedure based on several criteria. In the context of this Chapter, metrics quantify accessibility based on different accessibility checkpoints.

Universal Usability: A research field that studies the adequacy of user interfaces and information to all users, regardless of their characteristics, knowledge, or mean of interaction (Shneiderman, 2000).

Web Accessibility: The subfield of accessibility that is targeted to the specific technologies and architecture that compose the World Wide Web. This includes technologies such as HTML, CSS and JavaScript, as well as the HTTP protocol.

Web Graph: A formal representation of the Web's structure. Web pages are represented as the graph's nodes, whereas hyperlinks are represented as its arcs. By representing the Web as a graph, traditional graph analysis algorithms can be applied.

LIST OF NAMESPACE PREFIX/URI MAPPING

1. **earl:** http://www.w3.org/ns/earl#
2. **ev:** http://hcim.di.fc.ul.pt/ontologies/evaluation#

This work was previously published in Handbook of Research on Social Dimensions of Semantic Technologies and Web Services, edited by M. M. Cruz-Cunha; E. F. Oliveira; A. J. Tavares; L. G. Ferreira, pp. 88-112, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.15
Feature Selection for Web Page Classification

K. Selvakuberan
Tata Consultancy Services, India

M. Indra Devi
Thiagarajar College of Engineering, India

R. Rajaram
Thiagarajar College of Engineering, India

ABSTRACT

The World Wide Web serves as a huge, widely distributed, global information service center for news, advertisements, customer information, financial management, education, government, e-commerce and many others. The Web contains a rich and dynamic collection of hyperlink information. The Web page access and usage information provide rich sources for data mining. Web pages are classified based on the content and/or contextual information embedded in them. As the Web pages contain many irrelevant, infrequent, and stop words that reduce the performance of the classifier, selecting relevant representative features from the Web page is the essential preprocessing step. This provides secured accessing of the required information. The Web access and usage information can be mined to predict the authentication of the user accessing the Web page. This information may be used to personalize the information needed for the users and to preserve the privacy of the users by hiding the personal details. The issue lies in selecting the features which represent the Web pages and processing the details of the user needed the details. In this article we focus on the feature selection, issues in feature selections, and the most important feature selection techniques described and used by researchers.

DOI: 10.4018/978-1-60566-196-4.ch012

INTRODUCTION

There are an estimated 15 to 30 billion pages available in the World Wide Web with millions of pages being added daily. Describing and organizing this vast amount of content is essential for realizing the web's full potential as an information resource. Automatic classification of web pages is needed for the following reasons. (a) Large amount of information available in the internet makes it difficult for the human experts to classify them manually (b) The amount of Expertise needed is high (c) Web pages are dynamic and volatile in nature (e) More time and effort are required for classification. (f) Same type of classification scheme may not be applied to all pages (g) More experts needed for classification. Web page classification techniques use concepts from many fields like Information filtering and retrieval, Artificial Intelligence, Text mining, Machine learning techniques and so on. Information filtering and retrieval techniques usually build either a thesauri or indices by analyzing a corpus of already classified texts with specific algorithms. When new text is to be classified, thesaurus and index are used to find the similarity with already existing classification scheme to be associated with this new text.

Until the late 1980s, the most effective approach to web page classification seemed to be that of manually by building classification systems by means of knowledge-engineering techniques, i.e. manually defining a set of logical rules that encode expert knowledge on how to classify web page documents under the given set of categories. In the 1990s this perspective has been overturn, and the machine learning paradigm to automated web page classification has emerged and definitely superseded the knowledge-engineering approach. Within the machine learning paradigm, a general inductive process automatically builds an automatic text classifier by "learning", from a set of previously classified web documents, the characteristics of the categories of interest. The advantages of this approach are accuracy comparable to human performance and a considerable savings in terms of expert manpower, since no intervention from either knowledge engineers or domain experts is needed. Currently web page categorization may be seen as the meeting point of machine learning and information retrieval. As Machine Learning aims to address larger, more complex tasks, the problem of focusing on the most relevant information in a potentially overwhelming quantity of data has become increasingly important. For instance, data mining of corporate or scientific records often involves dealing with both many features and many examples, and the internet and World Wide Web have put a huge volume of low-quality information at the easy access of a learning system. Similar issues arise in the personalization of filtering systems for information retrieval, electronic mail, net news.

The main objective of this chapter is to focus on the feature selection techniques, need for feature selection, their issues in web page classification, feature selection for privacy preserving data mining and the future trends in feature selection.

LITERATURE SURVEY

Rudy Setiono and Huan Liu (1997) proposed that Discretization can turn numeric attributes into discrete ones. χ^2 is a simple algorithm. Principal Component Analysis-compose a small number of new features. It is improved from simple methods such as equi-width and equal frequency intervals. For each and every attributes calculate the χ^2 value for each and every interval. Combine the lowest interval values while approximation.

Shounak Roychowdhury (2001) proposed a technique called granular computing for processing and expressing chunks of information called granules. It reduces hypothesis search space, to reduce storage. Fuzzy set based feature elimination techniques in which subset generation and subset evaluation are employed. For optimal feature selection brute force technique is employed.

Catherine Blake and Wander Pratt (2001) suggested the relationship between the features used to represent the text and the quality model. A comparison of association rules based on three different concepts: words, manually assigned keywords, automatically assigned concepts are made. Bidirectional association rules on concepts or keywords are useful than the words used. Each individual feature should be informative. The quality of features should be meaningful. The concepts and keywords also represent fewer 90% of the words used in the medical diagnosis.

Martin, Mario and Anil (2004) discuss the various algorithms of clustering and the issues of feature selection such as what attributes and data should be selected. Feature saliency should be maintained. EM algorithm and mixture based clustering are employed. Minimum message length-saliency of irrelevant features is reduced to zero. Methods based on variance (PCA) need not produce best features. The Filter and Wrapper approaches are also employed for feature classification.

Christoph, Nidal Zeidat, and Zhenghong Zhao (2004) proposed an algorithm called supervised clustering. The goal is to identify the class uniform clusters that have high probability densities.

Four algorithms have been suggested.

1. A greedy algorithm with random restart
2. SRIDHCR, that seeks for solutions by inserting and removing single objects from the current solution
3. SPAM (a variation of the clustering algorithm PAM),
4. An evolutionary computing algorithm named SCEC, and a fast medoid-based top-down splitting algorithm, named TDS. The four algorithms were evaluated using a benchmark consisting of four UCI machine learning data sets. Fitness function and Impurity and number of clusters are also taken into account.

Huang, McCullagh, Black (2004) used ReliefF as a feature mining technique that is sensitive to the definition of relevance. It is computationally expensive in handling large data sets. They proposed an optimization algorithm (Feature Selection via Supervised Model Construction) for data transformation and starter selection, and evaluate its effectiveness with C4.5. Frequency based encoding scheme is employed for transforming categorical data into numerical data. The number of instances sampled from the data set determines the selection of features. Experiments are performed on UCI Repository data set and concluded that their proposed supervised model outperforms the other models

WEB PAGE CLASSIFICATION

Web page classification, also known as web page categorization, is the process of assigning a web page to one or more predefined category labels. Classification is often posed as a supervised learning problem in which a set of labeled data is used to train a classifier which can be applied to label future examples. The general problem of web page classification can be divided into multiple sub-problems: subject classification, functional classification, sentiment classification, and other types of classification. Subject classification is concerned about the subject or topic of a web page. For example, Classifying whether a page is about “arts”, “business” or “sports” is an instance of subject classification. Functional classification cares about the role that the web page plays. For example, deciding a page to be a “personal homepage”, “course page” or “admission page” is an instance of functional classification. Sentiment classification focuses on the opinion that is presented in a web page, i.e., the author’s attitude about some particular topic. Based on the number of classes in the problem, classification can be divided into binary classification and multi-class classification, where binary classification categorizes instances into exactly one of two classes positive or negative) ; multi-class classification deals with more than two classes. Based on the number of classes that can be assigned to an instance, classification can be divided into single-label classification and multi-label classification. In single-label classification, one and only one class label is to be assigned to each instance, while in multi-label classification, more than one class can be assigned to an instance.

If a problem is multi-class, say four-class classification, it means four classes are involved.

Classification plays a vital role in many information management and retrieval tasks. In case of the Web, classification of page content is essential to focused crawling, to the assisted development of web directories, to topic-specific web link analysis, and to analysis of the topical structure of the Web. Web page classification can also help improve the quality of web search. Earlier surveys in web page classification typically lack a detailed discussion of the utilization of web specific features. In this article, we carefully examine the web-specific features and algorithms that have been explored and found to be useful for web page classification. The contributions of this article are:

- a detailed review of useful web-specific features for classification;
- various feature selection techniques used in classification; and,
- a discussion of future research directions.

WHAT IS FEATURE SELECTION?

Feature extraction or selection is one of the most important steps in pattern recognition or pattern classification, data mining, machine learning and so on. Generally speaking, only classification information is included sufficiently in the eigenvector, classifier can classify the classification rightly. However, it is difficult to measure classification information in all features. Data preprocessing is an indispensable step in effective data analysis. It prepares data for data mining and machine learning, which aim to turn data into business intelligence or knowledge. Feature selection is a preprocessing technique commonly used on high dimensional data. Feature selection studies how to select a subset or list of attributes or variables that are used to construct models describing data. Its purposes include reducing dimensionality, removing irrelevant and redundant features, reducing the amount of data needed for learning, improving algorithms' predictive accuracy, and increasing the constructed models' comprehensibility. Feature-selection methods are particularly welcome in interdisciplinary collaborations because the selected features retain the original meanings domain experts are familiar with. The rapid developments in computer science and engineering allow for data collection at an unprecedented speed and present new challenges to feature selection. Wide data sets, which have a huge number of features but relatively few instances, introduce a novel challenge to feature selection.

Need for Feature Selection

The web pages need 80% of the preprocessing work since the web pages have large amount of useless information. Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user for example, in a neural network. There are a number of different tools and methods used for preprocessing, including: sampling, which selects a representative subset from a large population of data; transformation, which manipulates raw data to produce a single input; denoising, which removes noise from data; normalization, which organizes data for more efficient access; and feature extraction, which pulls out specified data that is significant in some particular context.

In a customer relationship management (CRM) context, data preprocessing is a component of Web mining. Web usage logs may be preprocessed to extract meaningful sets of data called user transactions, which consist of groups of URL references. User sessions may be tracked to identify the user, the Web sites requested and their order, and the length of time spent on each one. Once these have been pulled out of the raw data, they yield more useful information that can be put to the user's purposes, such as consumer research, marketing, or personalization.

Preprocessing makes it possible for complex homepages to be delivered lightning fast, and lets you significantly increase the number of pages served in an extremely cost-effective manner. The idea of preprocessing content for web pages grew out of necessity. Preprocessing can be achieved by the following ways:

A . Removing HTML tags:

HTML tags indicate the formats of web pages. For instance, the content within <title> and </title> pair is the title of a web page; the content enclosed by <table> and </table> pair is a table. These HTML tags may indicate the importance of their enclosed content and they can thus help weight their enclosed content. The tags themselves are removed after weighting their enclosed content.

B . Removing stop words:

Stop words are frequent words that carry little information, such as prepositions, pronouns, and conjunctions. They are removed by comparing the input text with a "stop list" of words.

C . Removing rare words:

Removing words whose number of occurrences in the text is less than a predefined threshold.

D . Performing word stemming:

Word stemming is done by grouping words that have the same stem or root, such as computer, compute, and computing. The Porter stemmer is a well-known algorithm for performing this task.

Issues in Feature selection

Feature extraction or selection is one of the most important steps in pattern recognition or pattern classification, data mining, machine learning and so on. But the increasing feature brings disadvantages for classification problem. On one hand, feature increased gives difficulties to calculate, because the more data occupy amount of memory space and computerization time, on the other hand, a lot of features include certainly many correlation factors respectively, which results to information repeat and waste. Therefore, we must take measures to decrease the feature dimension under not decreasing recognition effect; this is called the problems of feature optimum extraction or selection. On the other hand the number of features needs to be constrained to reduce noise and to limit the burden on system resources. The number of features needs to be constrained to reduce noise and to limit the burden on system resources.

Characteristics of Selected Features

The purposes of automated text categorization, features should be:

1. Relatively few in number
2. Moderate in frequency of assignment
3. Low in redundancy
4. Low in noise
5. Related in semantic scope to the classes to be assigned
6. Relatively unambiguous in meaning

DIMENSIONALITY REDUCTION BY FEATURE SELECTION

In statistics, dimension reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction. Feature selection, also known as variable selection, feature reduction, attribute selection or variable subset selection, is the technique, commonly used in machine learning, of selecting a subset of relevant features for building robust learning models. . By removing most irrelevant and redundant features from the data, feature selection helps improve the performance of learning models by:

- Alleviating the effect of the curse of dimensionality.
- Enhancing generalization capability.
- Speeding up learning process.
- Improving model interpretability.

Feature selection also helps people to acquire better understanding about their data by telling them that which are the important features and how they are related with each other. Feature selection selects a subset of the original feature space based on some criteria. Two broad approaches for feature selection have been presented in the literature: the wrapper approach and the filter approach. The wrapper approach employs a search through the space of feature subsets. It uses an estimated accuracy for a learning algorithm as the measure of goodness for a particular feature subset. Thus the feature selection is being "wrapped around" a learning algorithm. For example, for a neural network algorithm the wrapper approach selects an initial subset of features and measures the performance of the network; then it generates an "improved set of features" and measures the performance of the network. This process is repeated until it reaches a termination condition (either a minimal value of error or a number of iterations). While some wrapper based methods have encountered some success for classification tasks, they are often prohibitively expensive to run and can break down when a very large number of features are present. For the filter approach, feature selection is performed as a preprocessing step before applying machine learning. Thus the method of feature selection is independent to the learning algorithm. The filter algorithm does not incur the high computational cost and is commonly used in classification systems even in a very high feature space.

Feature extraction is a special form of dimensionality reduction.

When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant (much data, but not much information) then the input data will be transformed into a reduced representation set of features (also named features vector). Transforming the input data into the set of features is called *features extraction*. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which overfits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy

Jun Yan, Benyu Zhang, Ning Liu, Shuicheng Yan, Qiansheng Cheng, Weiguo Fan, Qiang Yang, Wensi Xi, and Zheng Chen (2006) give an overview of the popularly used feature extraction and selection algorithms under a unified framework. They propose two novel dimensionality reduction algorithms based on the Orthogonal Centroid algorithm (OC). The first is an Incremental OC (IOC) algorithm for feature extraction. The second algorithm is an Orthogonal Centroid Feature Selection (OCFS) method which can provide optimal solutions according to the OC criterion. Both are designed under the same optimization criterion. Experiments on Reuters Corpus Volume-1 data set and some public large-scale text data sets indicate that the two algorithms are favorable in terms of their effectiveness and efficiency when compared with other state-of-the-art algorithms.

FEATURE SELECTION STEPS

Preprocessing of web pages is the first step for the web page classification problem. Web pages cannot be processed as such because of the size, content and nature of the web pages. Dimensionality reduction is an essential data preprocessing technique for large-scale and streaming data classification tasks. It can be used to improve both the efficiency and the effectiveness of classifiers. Traditional dimensionality reduction approaches fall into two categories: Feature Extraction and Feature Selection. Techniques in the feature extraction category are typically more effective than those in feature selection category. Feature extraction refers to the extraction from the various features of the web page such as Title, Meta and URL of the web page.

Feature selection is a process that selects a subset of original features. The optimality of a feature subset is measured by an evaluation criterion. A typical feature selection process consists of four basic steps namely, subset generation, subset evaluation, stopping criterion, and result validation. Subset generation produces candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation criterion. If the new subset turns out to be better, it replaces the previous best subset. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied.

FEATURE SELECTION TECHNIQUES

PCA

Principal Component Analysis (PCA) involves a mathematical procedure that transforms a large number of correlated variables into a smaller number of uncorrelated variables called principal components. The objectives of principal component analysis are to discover (or reduce) the dimensionality of the data set and identifies new meaningful underlying variables. The mathematical technique used in PCA is called eigen values. PCA is a classical statistical method that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA is theoretically the optimum transform for a given data in least square terms.

Principal Component Analysis is a feature selection technique that searches for c k-dimensional orthogonal vectors that can be used to represent the data, where c<=k. The original data are those projected into a smaller space, resulting in data compression. PCA can be used as a form of dimensionality reduction. The input data are normalized, so that each attribute falls within the same range. It helps to ensure that attributes with large domains do not dominate the smaller domains. It uses a special coordinate system that depends upon the cloud of points. Place the first axis in the direction of greater variance of the points to maximize the variance along that axis. The second axis is perpendicular to it. It is also a technique used to reduce multi dimensional data sets to lower dimensions for analysis.

PCA is a non-parametric analysis and independent of any hypothesis about data probability distribution. PCA compression and decompression are easy operations to perform given the model parameters. However, the latter two properties are regarded as weakness as well as strength, in that being non-parametric, no a-priori assumptions can be incorporated and that PCA compressions often incur loss of information. When used for the applications of clustering, the main limitation of PCA is that it does not consider class separability since it does not take into account the class label of the feature vector. PCA simply performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance. There is no guarantee that the directions of maximum variance will contain good features for discrimination

Fei Wu,Yonglei Zhou and Changshui Zhang proposed a general iterative framework for relevant linear feature extraction that efficiently utilizes both the side information and unlabeled data to enhance gradually algorithms performance and robustness. Both good relevant feature extraction and reasonable similarity matrix estimation can be realized. They adopt Relevant Component Analysis (RCA) under this framework and get the derived Iterative Self-Enhanced Relevant Component Analysis (ISERCA) algorithm. The experimental results on several data sets show that ISERCA outperforms RCA. Side-information represents some equivalence constraint between pair of samples, indicating whether the two samples originate from the same but unknown category (positive constraint) or from two different categories (negative constraint). They use “labeled data” to denote the samples involved in the given side-information

Information Gain

Information Gain is a method that removes the less informative attributes, collecting the more informative ones for use in concept description analysis. It is a dimension based data analysis method. An arbitrary sample belongs to class C_i with the probability s/S where S is the total number of samples. The expected information gain is given by:

$$I(s_1,s_2,\ldots,s_m) = -\sum s/S \log s/S$$

and the gain is given by

$$\text{Gain} = I(s_1,s_2,\ldots,s_m) - E(A)$$

Information theoretic methods are also used to evaluate features: the mutual information between a relevant feature and the class labels should be high. Nonparametric methods can be used to compute mutual information involving continuous features.Although information gain is usually a good measure for deciding the relevance of an attribute, it is not perfect. A notable problem occurs when information gain is applied to attributes that can take on a large number of distinct values.

Mark Last and Oded, Maimon (2004) proposed an algorithm called Information theoretic algorithm that is based on minimal subset features. Information Network, is a tree like structure is formed by input features and the targeted classification. Unlike other decision-tree models, the information network uses the same input attribute across the nodes of a given layer (level). The input attributes are selected incrementally by the algorithm to maximize a global decrease in the conditional entropy of the target attribute. They employ the pre pruning approach: When no attribute causes a statistically significant decrease in the entropy, the network construction is stopped. More information implies higher accuracy of classification.

TF-IDF

Each document is represented in the term space, such that $d = \{w_1, w_2, \ldots, w_n\}$, where$w_i$, $i = 1, \ldots, n$, is the weight of term i in the document. The weight of a term could be simply calculated as the frequency of the term in that document ($w_i = tf_i$); i.e. how many times it appeared in the document. A more popular term weighting scheme is TF×IDF (Term Frequency ×Inverse Document Frequency), which takes into account the document frequency of a term (df $_i$), the nuwmber of documents in which the term appears. A typical inverse document frequency (idf) factor of this type is given by log(N/ df $_i$). Thus the TF×IDF weight of a term is w $_i$ = tf$_i$ × log(N/ df$_i$. In other words, terms that appear more frequently in a certain document but less frequently in other documents are given higher weights in that document, since it has higher correlation with that document than others. On the other hand, terms that appear frequently in all documents are penalized in all documents since they have less discrimination power.

Daniele Riboni(2005) conducted various experiments on a corpus of 8000 documents belonging to 10 Yahoo! categories, using Kernel Perceptron and Naive Bayes classifiers. They introduce a new method for representing linked pages using local information that makes hypertext categorization feasible for real-time applications. Experimental results show that the local words with a hyper textual one can improve classification performance.They tested five different text sources for web page representation namely:

BODY, META, TITLE, MT, the union of META and TITLE content and BMT, the union of BODY, META and TITLE content.

Alexandros Kalousis, Julien Prados, Melanie Hilario (2005) suggested that study is an attempt to fill the gap by quantifying the sensitivity of feature selection algorithms to variations in the training set. The authors assess the stability of feature selection algorithms based on the stability of the feature preferences that they express in the form of weights scores, ranks or a selected feature subset. They examine a number of measures to quantify the stability of feature preferences and propose an empirical way to estimate them.

Chi-Square (CHI)

In probability theory and statistics, the chi-square distribution (also chi-squared or χ^2 distribution) is one of the most widely used theoretical probability distributions in inferential statistics, i.e. in statistical significance tests. It is useful because, under reasonable assumptions, easily calculated quantities can be proven to have distributions that approximate to the chi-square distribution if the null hypothesis is true. The chi-square distribution has one parameter: *k* - a positive integer that specifies the number of degrees of freedom (i.e. the number of X_i). The chi-square distribution is a special case of the gamma distribution. The best-known situations in which the chi-square distribution is used are the common chi-square tests for goodness of fit of an observed distribution to a theoretical one, and of the independence of two criteria of classification of qualitative data. However, many other statistical tests lead to a use of this distribution. Chi-square measures the lack of independence between a term t and a category c_i and can defined as

$$x^2\left(t, c_i\right) = \frac{N\left[P\left(t, c_i\right)P\left(\bar{t}, \overline{c_i}\right) - P\left(t, \overline{c_i}\right)P\left(\bar{t}, c_i\right)\right]^2}{P\left(t\right)P\left(\bar{t}\right)P\left(c_i\right)P\left(\overline{c_i}\right)}$$

Hongjun Lu, Sam Yuan Sung and Ying Lu (1996) propose Conflict analysis that is finding a set of attributes having perfect association with the class labels contingency table analysis is used with the nominal variables –the variables whose values are from an unordered set used chi square statistics.

Correlation coefficient (CC)

Correlation is a measure to identify the relationship between the attributes. For that purpose, correlation coefficients are introduced. The correlation coefficient is a number between -1 and 1 which measures the degree to which two variables are linearly related. If there is perfect linear relationship with positive slope between the two variables, we have a correlation coefficient of 1; if there is positive correlation, whenever one variable has a high (low) value, so does the other. If there is a perfect linear relationship with negative slope between the two variables, we have a correlation coefficient of -1; if there is negative correlation, whenever one variable has a high (low) value; the other has a low (high) value. A correlation coefficient of 0 means that there is no linear relationship between the variables.

Wen-Zhou Chen and Lei Li (2004) proposed a model of Correlation based Modified SVM, which ranks the features according to the correlation measures. Forward selection search with correlation based method to form a feature subset, labeled training examples each with a feature vector and a class are used.

Correlation coefficient of a word t with a category c_i and can be defined as

$$CC\left(t,c_i\right)=\frac{\sqrt{N}\left[P\left(t,c_i\right)P\left(\bar{t},\overline{c_i}\right)-P\left(t,\overline{c_i}\right)P\left(\bar{t},c_i\right)\right]}{\sqrt{P\left(t\right)P\left(\bar{t}\right)P\left(c_i\right)P\left(\overline{c_i}\right)}}$$

Odds ratio (OR)

The odds ratio is one of a range of statistics used to assess the risk of a particular outcome if a certain factor (or exposure) is present. The odds ratio is a relative measure of risk, telling us how much more likely it is that someone who is exposed to the factor under study will develop the outcome as compared to someone who is not exposed. Odds are a way of presenting probabilities. The odds of an event happening is the probability that the event will happen divided by the probability that the event will not happen. It is a measure used for selecting terms for relevance feedback. The basic idea is that the distribution of features on the relevant documents is different from the distribution of features on the non-relevant documents. It is defined as follows:

$$OR\left(t,c_i\right)=\log\frac{P\left(t,\middle|c_i\right)\left[1-P\left(t\middle|\overline{c_i}\right)\right]}{\left[1-PP\left(t,\middle|c_i\right)\right]P\left(t\middle|\overline{c_i}\right)}$$

Filters and Wrappers

The filter approaches evaluate the relevance of each feature (subset) using the data set alone, regardless of the subsequent learning algorithm. The filter model relies on general characteristics of the training dataset to select some features without involving any learning algorithm. The wrapper model requires one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are selected. It invokes the learning algorithm to evaluate the quality of each feature (subset). Specifically, a learning algorithm (e.g., a nearest neighbor classifier, a decision tree, a naive Bayes method) is run on a feature subset and the feature subset is assessed by some estimate of the classification accuracy. Wrappers are usually more computationally demanding, but they can be superior in accuracy when compared with filters, which ignore the properties of the learning task at hand. Wrapper models tend to find features better suited to the predetermined learning algorithm resulting in superior learning performance, but it is also computationally expensive compared to the filter model.

Huan Liu and Lei Yu proposed that feature selection algorithms for classification and clustering, groups and compares different algorithms with a categorizing framework based on search strategies, evaluation criteria, and data mining tasks, reveals unattempted combinations, and provides guidelines in selecting feature selection algorithms. With the categorizing framework, an integrated system for intelligent feature selection is built up. For employing feature selection, wrapper, filter and hybrid model are adopted.

Elı´as, Elena Montane´ s, Irene Dı´az, Jose´ Ranilla, and Ricardo Mones (2005) suggested that to select the relevant features by a family of linear filtering approaches. The feature selection approaches are bag of words representation, filter and wrapper approaches, term frequency, document frequency, inverted document frequency and the Information Gain indicates the presence of word in the category or not. The distribution of documents over the categories is considered by introducing the concept of canonical or unconditional rule which says that any document belongs to the category. This rule is used as a reference for the rest of rules of the same category. SVM classifier is employed with Reuters 21578 collections as the experimental data.

Decision Tree Based Feature Selection

The basic decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide and conquer manner. Nodes in the decision tree involve testing a particular attribute. The test at a particular node compares the attribute values with a constant.

A decision tree can be constructed top-down using the information gain in the following way:

1. begin at the root node
2. determine the attribute with the highest information gain which is not used in an ancestor node
3. add a child node for each possible value of that attribute
4. attach all examples to the child node where the attribute values of the examples are identical to the attribute value attached to the node
5. if all examples attached to the child node can be classified uniquely add that classification to that node and mark it as leaf node
6. go back to step two if there is at least one more unused attribute left, otherwise add the classification of most of the examples attached to the child node

Philip Laird Ronald Saul (1994) proposed that each object is represented in the form of vector of attribute values. Genetic based algorithm includes decision trees, feed forward neural networks and Bayesian classifiers. Value-class pair(x,C) is generated. Based on the value-class pair decision tree is constructed. At each node in the tree, a feature is tested for inclusion in two or more sets or in ranges. Minimum message length (MML) is used to determine the mutual information gain. Fringe technique is employed to identify the identical sub trees and the determination of the root.

Heuristic Search Trees

The predominant state-space planning methods in artificial intelligence are collectively known as heuristic search. Heuristic search is not concerned with changing the approximate, or “heuristic,” value function, but only with making improved action selections given the current value function. In other words, heuristic search is planning as part of a policy computation. In heuristic search, for each state encountered, a large tree of possible continuations is considered. The approximate value function is applied to the leaf nodes and then backed up toward the current state at the root. The backing up stops at the state-action nodes for the current state. Once the backed-up values of these nodes are computed, the best of them is chosen as the current action, and then all backed-up values are discarded. In conventional heuristic search no effort is made to save the backed-up values by changing the approximate value function. In fact, the value function is generally designed by people and never changed as a result of search.

In conventional heuristic search, this process computes backed-up values of the possible actions, but does not attempt to save them. Thus, heuristic search can be viewed as an extension of the idea of a greedy policy beyond a single step. Search methods traverse the attribute space to find a good subset. Quality is measured by the chosen attribute subsets. Subsets that have been cached are evaluated.

Pat Langley (1994) proposes that the Heuristics search through a space of feature set. Search Space specifies a subset of original feature. First determine the starting point and next the direction in which they are applied .It includes forward selection & backward elimination greedy approach (BFS)

Clustering with Tree Representation

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Data clustering algorithms can be hierarchical or partitional. Hierarchical algorithms find successive clusters using previously established clusters, whereas partitional algorithms determine all clusters at once. Hierarchical algorithms can be agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. Two-way clustering, co-clustering or biclustering are clustering methods where not only the objects are clustered but also the features of the objects, i.e., if the data is represented in a data matrix, the rows and columns are clustered simultaneously

An algorithm for feature selection that clusters attributes using a special metric and, generates clusters that are placed in a cluster tree. Clustering is obtained by extracting those clusters that are situated at a given height in this tree.

Richard Butterworth, Gregory Piatetsky-Shapiro and Dan (2005) devised an algorithm for feature selection that clusters attributes using a special metric and then makes use of the dendrogram of the resulting cluster hierarchy to choose the most relevant attributes. The main interest of our technique resides in the improved understanding of the structure of the analyzed data and of the relative importance of the attributes for the selection process. Hierarchical algorithms generate clusters that are placed in a cluster tree, which is commonly known as a dendrogram. Clusters are obtained by extracting those clusters that are situated at a given height in this tree.

Hierarchical Trees

A hierarchical method creates a hierarchical decomposition of the given set of objects. A hierarchical method can be classified as agglomerative or divisive approach which is based on how the hierarchical decomposition is formed. The agglomerative approach is a bottom-up approach. The divisive approach is a top-down approach.

Shou-Bin Dong (2004) proposed a technique of hierarchical classification of web content based on the combination of both textual and visual features. Images can be ignored during classification. Combination of multiple classifiers is employed. The most widely used approach typically combines the classifier outputs directly by means of simple combining rules or functions. It relates to techniques like majority vote, threshold voting, averaged Bayes classifier, different linear combinations of a posteriori probabilities, maximum and minimum rules, product rule. Page, summary, title, image related text are also taken into account.

Suresh, Jitender, Vijay and Heyeri Sever (1996) use four feature selection algorithms-all the algorithms start with the same feature space but the heuristic used for pruning is different. The objective is to find a small amount of features that are sufficient and necessary for describing the feature space

1. **Best fit SBS:** Start with all the features and the features are removed one at a time
2. **Hybrid Heuristic SBS:** The current node is considered as the root node. The root nodes are expanded in such a way that the successors have one less condition attributes than its predecessors
3. **Alternating heuristic SBS:** The current node is assigned as the root node. Alternately BFS and first fit can be employed to find the next node of detail.
4. **k-level best SBS:** It divides the search space into k-groups starting from the root node to the leaf node. The last stage contains less than k-levels.

FUTURE TRENDS IN FEATURE SELECTION APPROACH

Hybrid Feature Selection

Hybrid feature selection combines different feature selection algorithms. It took the best characteristics of each and every algorithm and combines them effectively. Now –a-days the researchers used to combine the multiple feature selection algorithms for classification.

Yi Lu Murphey and Hong Guo (2000) proposed a model called hybrid feature selection algorithm using three different statistical measurements such as class pair-wised distance, linear separability, and overlapped feature histogram. They applied Bayes EM algorithm to select a sub-optimal set. The hybrid feature selection algorithm can be used as a preprocessing in a classification system and it is independent of the classifier to be used in the subsequence stage.

Hwanjo Yu, Kevin Chen-Chuan Chang and Jiawei Han (2002) propose a model that eliminates the need for negative training data and they concluded that the classification accuracy is increased by using two learners.

Sampling Technique and Correlation

Sampling is a reduction technique as it allows a large data set to be represented as small samples of data. The samples are thus correlated and the results are validated. Sampling is that part of statistical practice concerned with the selection of individual observations intended to yield some knowledge about a population of concern, especially for the purposes of statistical inference. Each observation measures one or more properties (weight, location, etc.) of an observable entity enumerated to distinguish objects or individuals.

Sriharsha Veeramachaneni Paolo Avesani (2004) proposed a technique of sampling the feature values with the ultimate goal of choosing between alternative candidate features with minimum sampling cost. Their heuristic algorithm is based on extracting candidate features in a region of the instance space. An experimental evaluation on a standard database shows that it is possible outperform a random sub sampling policy in terms of the accuracy in feature selection. The basic idea is to prescribe an iterative policy that chooses the next instance on which the candidate features are to be probed.

PRIVACY PRESERVING DATA MINING USING FEATURE SELECTION TECHNIQUES

Data mining is usually performed for applications not having personal details like weather forecasting, GPSS, etc. But data mining for banking, e-commerce, mobile and the like applications involve databases with personal details. Here the security aspects have to be taken into account while retrieving the features used to find the required web page by the user. One way to achieve this is by using a multilevel authentication system. The user giving the query must be an authenticated user to pose that query. Another way is to distort the data by inserting noise so that no one can access the personal details only the results or prediction can be shown. Yet another way is encrypt the information stored in the database so that if more than one person/company involved in the data mining, no one can get the other person/company details. Data mining can also be used to detect susceptible queries that intend to retrieve the personal information and also can be used to detect the person/program who did it. All the above feature selection techniques can be used to deal the security issues and to find the best features.

CONCLUSION

In this chapter we focus on various feature selection methods and how they are used by many researchers for selecting optimal feature set for web page classification problem. We mainly concentrate on the various ways a particular method is used for the purpose rather than giving references that solve the problem in the same way. This article may be incomplete in the aspect that it does not refer all the papers in the feature selection area but this is a useful research work for researchers who do research in the web page representation and web page classification areas.

REFERENCES

Blake, C., & Pratt, W. (2001). Better rules,f ewer features: A semantic approach to selecting features from text. IEEE-2001, 59-66.

Butterworth, R., Piatetsky-Shapiro, G., & Simovici, D. A. (2005). On Feature Selection through Clustering. *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05).*

Chen, W.-Z., & Li, L. (2004). Correlation and MSVM-based Feature Selection. *Proceedings of the Third International Conference on Machine Learning and Cybernetics*, Shanghai, 26-29 August 2004.

Combarro, E. F., Montanes, E., Dı'az, I., Ranilla, J., & Mones, R. (2005). Introducing a family of linear measures for feature selection in text categorization. *IEEE Transactions on Knowledge and Data Engineering*, *17*(9), 1223–1232. doi:10.1109/TKDE.2005.149

Eick, C. F., Zeidat, N., & Zhao, Z. (2004). Supervised Clustering – Algorithms and Benefits. *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004).*

Huang, Y., McCullagh, P. J., & Black, N. D. (2004). Feature Selection via Supervised Model Construction. Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04).

Kalousis, A., Prados, J., & Hilario, M. (2005). Stability of Feature Selection Algorithms. *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05).*

Langley, P. (1994). Selection of Relevant Features on Machine Learning. *Proceedings of AAAI conference on Relevance-1994.*

Last, M., & Maimon, O. (2004, February). A compact and accurate model for classification. *IEEE Transactions on Knowledge and Data Engineering*, *16*(2), 203–215. doi:10.1109/TKDE.2004.1269598

Law, M. H. C., Figueiredo, M. A. T., & Jain, A. K. (2004, September). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(9), 1154–1166. doi:10.1109/TPAMI.2004.71

Liu, H., & Yu, L. (2005). Towards integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, *17*(4), 491–502. doi:10.1109/TKDE.2005.66

Lu, H., Sung, S. Y., & Lu, Y. (1996). On Preprocessing Data for Effective Classification. *Workshop on Research Issues on Data Mining, Proceedings of ACM.*

Murphey, Y.-L., & Guo, H. (2000). Automatic Feature Selection - A hybrid statistical approach. *IEEE-2000*, 382-385.

Riboni, D. (2005). Feature Selection for Web Page Classification. *Proceedings of International Conference on Data Mining, ICDM'05.*

Roychowdhury, S. (2001). Feature Subset Selection using Granular Information. *IEEE-2001*, 2041-2045

Saul, P. L. R. (1994). Automated Feature Extraction for supervised Learning, 674-679.

Scherf, M., & Brauer, W. (1997). Feature Selection by Means of a Feature Weighting Approach. Technical Report No. FKI-221-97, Forschungsberichte kunstliche Intelligenz, Institut fur Informatik, Technische Universitat Munchen,http://citeseer.ist.psu.edu/scherf97feature.html

Setino, R., & Liu, H. (1997). Feature selection via discretization. *IEEE Transactions on Knowledge and Data Engineering*, *9*(4), 642–645. doi:10.1109/69.617056

Shou-Bin, D. (2004). The Hierarchical Classification of Web content by the combination of Textual and Visual Features. *Proceedings of the Third International Conference on Machine Learning and Cybernetics*, Shanghai, 26-29 August 2004,1524-1529.

Suresh, K., Jitender, C., Deogun, S., & Vijay, V. Raghavan, & Sever, H. (1996). A comparison of Feature Selection Algorithms in the context of Rough Classifiers. *Proceedings of Fifth IEEE conference on Fuzzy Systems, 2*(8-11), 1122-1128.

Wu, F., Zhou, Y., & Zhang, C. (2004). Relevant Linear Feature Extraction Using Side-information and Unlabeled Data. *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04).*

Yan, J., Zhang, B., Liu, N., Yan, S., Cheng, Q., & Fan, W. (2006). Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing. *IEEE Transactions on Knowledge and Data Engineering*, *18*(3), 320–333. doi:10.1109/TKDE.2006.45

Yu, H., Chen-Chuan Chang, K., & Han, J. (2002). Heterogeneous Learner for Web Page Classification. *Proceedings of International Conference on Data Mining 2002.*

KEY TERMS

Feature Extraction: Feature extraction is the preceding step to feature selection. It includes the collection of features from the Web pages. It can be done using any of the feature extraction techniques. All the collected features during the feature extraction phase need not be relevant.

Feature Selection: Feature selection is a process of selecting the relevant features needed for classification purposes. Feature selection is the most important step because the selected features determine the classification accuracy. Feature selection is a preprocessing technique commonly used on high dimensional data. Feature selection studies how to select a subset or list of attributes or variables that are used to construct models describing data. Its purposes include reducing dimensionality, removing irrelevant and redundant features, reducing the amount of data needed for learning, improving algorithms' predictive accuracy, and increasing the constructed models' comprehensibility.

HTML Structure: HTML refers to Hypertext Markup Language. It is used for the design of Web pages. Most of the Web pages contain html tags enclosed.

Machine Learning: Machine learning usually refers to the changes in systems that perform tasks associated with artificial intelligence. Such tasks involve recognition, diagnosis, planning, robot control and prediction etc.It is a technique commonly used for automatic classification of Web pages. It includes both the training phase and the testing phase. The training phase is used to learn from examples. Based on the learning the classifier classifies during the testing phase. Machine learning has a wide spectrum of applications including natural language processing, syntactic pattern recognition, search engines, medical diagnosis, bioinformatics, detecting credit card fraud, stock market analysis, classifying DNA sequences, speech and handwriting recognition, object recognition in computer vision, game playing and robot locomotion

Preprocessing: Preprocessing is the process of removing unnecessary information before the classification of Web pages. It includes removing common words, rare words, removing HTML tags and stemming.

Web Page Classification: Web page classification, also known as Web page categorization, may be defined as the task of determining whether a Web page belongs to a category or categories. It can be done by considering the various features like URL, head, title and the body contents of the Web page.

World Wide Web: The World Wide Web (commonly shortened to the Web) is a system of interlinked hypertext documents accessed via the Internet. With a Web browser, a user views Web pages that may contain text, images, videos, and other multimedia and navigates between them using hyperlinks. The World Wide Web is also called as net of networks. It contains a large amount of information needed for the Internet browsers. Boundless amount of information can be downloaded at any point of time.

This work was previously published in Social Implications of Data Mining and Information Privacy: Interdisciplinary Frameworks and Solutions, edited by E. Eyob, pp. 213-228, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.16
Implementing Collaborative Problem-Based Learning with Web 2.0

Steven C. Mills
The University Center of Southern Oklahoma, USA

ABSTRACT

Educators face the challenge of keeping classroom learning relevant for a generation of students who have never known life without computers, cell phones, and email. With Web 2.0 technologies educators can easily mediate student-centered learning experiences that engage students collaboratively in problem-solving and critical thinking. This chapter describes how Web 2.0 technologies can supply communication tools and information resources that facilitate the application of a robust set of instructional methodologies in the K-12 classroom. When the pedagogical features of Web 2.0 technologies are used with problem-solving methodologies, teachers can create powerful student-centered learning experiences for educating students for the 21st century.

INTRODUCTION

An 8th grade science teacher, Ms. S, retrieves her MP3 player from the computer-connected cradle where it's spent the night scanning the 17 podcasts she subscribes to. Having detected three new programs, the computer downloaded the files and copied them to the handheld. En route to work, Ms. S inserts the device into her dash-mounted cradle and reviews the podcasts, selecting a colleague's classroom presentation on global warming and a NASA conference lecture about interstellar space travel…

Meanwhile, social studies teacher Ms. L scans through sites tagged genetics in the school's social bookmark service. Her students may need quick access to them as they discuss genetic engineering current events during class…All assignments in Ms. L's class are turned in via blogs because she finds that their conversational nature encourages students to think and write in more depth than traditional formal essays or short answer assignments. Another advantage of receiving assignments in blog format is that both she and her students can subscribe, which means all of the students' blogs appear in her aggregator, and students can reap the benefits of seeing each other's work.

DOI: 10.4018/978-1-60566-120-9.ch024

A few doors down the hall, veteran English teacher, Mr. P, is reviewing a new batch of student wikis. In an effort to help the students become better communicators, he never provides study guides for tests, instead relying on students to construct their own study resources using their team wikis. He rewards teams that create the most useful/popular study guides. Mr. P uses a wiki tool installed on the school's network…

From *A Day in the Life of Web 2.0* by David Warlick

Over the past decade and a half since the creation of the Internet and the World Wide Web the use of information technology has significantly increased in K-12 classrooms. As the Web continues to evolve, new Internet and Web technologies become facts of life for today's students. And as David Warlick (2006) indicates in *A Day in the Life of Web 2.0*, teachers are using Web 2.0 technologies in K-12 classrooms because it is compatible with the technology many students use on a daily basis through popular websites such as MySpace, Wikipedia and Flikr.

According to Lee Rainie, Director of the Pew Internet and American Life project, "an American teen is more likely than her parents to own a digital music player like an iPod, to have posted writing, pictures or video on the Internet, to have created a blog or profile on a social networking website like MySpace, to have downloaded digital content such as songs, games, movies, or software, to have shared a remix or 'mashup' creation with friends, and to have snapped a photo or video with a cell phone." (Rainie, 2006). Because of this daily high-level interaction that youth have with technology, educators face the challenge of keeping classroom learning relevant for a generation of students who have never known life without computers, cell phones, and email. Baird and Fisher (2005-2006) note that "neomillennial students expect interactive, engaging content and course material that motivates them to learn through challenging pedagogy." Thus, the bar has been raised seemingly beyond the technological expertise of many educators for providing learning experiences that keep today's students interested and engaged.

The good news for educators is that the latest expression of the World Wide Web, known as Web 2.0, provides online information resources and communication technologies that are easier to use and simpler to implement, requiring far less technological expertise than the preceding generation of Internet applications. With Web 2.0 it is far easier for educators to mediate student-centered learning experiences utilizing the pedagogical features of these new technologies. For example, the social networking capabilities of Web 2.0 familiar to most students can promote student engagement in learning because students actively participate in constructing a learning landscape based on social interactions and information exchanges with peers (Baird & Fisher, 2005-2006).

Web 2.0 is a term used to describe Web technologies used to harness collective intelligence, provide interfaces and services across multiple devices, and enhance collaboration. Although the term, Web 2.0, suggests a new version or generation of the World Wide Web, in reality it refers to a re-visioning of the Web—what Downes (2005) characterizes not as a technological revolution but as a social revolution. Thus, Web 2.0 relates to new ways to use the Web rather than an update to the technical specifications of the Web. According to Downes, Web 2.0 represents a shift from an information-consumption medium to an information-creation platform: "In a nutshell, what was happening was that the Web was shifting from being a medium, in which information was transmitted and consumed, into being a platform, in which content was created, shared, remixed, repurposed, and passed along. And what people were doing with the Web was not merely reading books, listening to the radio or watching TV, but having a conversation, with a vocabulary consisting not just of words but of images, video, multimedia and whatever they could get their hands on."

Perhaps the most well-known example of Web 2.0 is a blog. Blogs permit writers, known as a bloggers, to post commentary and visitors to leave responses. Other examples of Web 2.0 applications are wikis, podcasts, Web feeds, social bookmarking, and social networking. Interactive online games and virtual worlds could also be considered Web 2.0 applications. While the label "Web 2.0" may imply a more advanced use of technology to deploy the Internet, what is important to this discussion is the pedagogical features of Web 2.0 that can be exploited for relevant and effective classroom instruction. Web 2.0 provides the affordance of collaborative information discovery that can be converted into instructional methodologies (Alexander, 2006) that can scaffold student learning in K-12 classrooms.

The International Society for Technology in Education's (2007) release of the National Educational Technology Standards for Students (NETS-S) indicates that to learn effectively and live productively in an increasingly digital world, students should know and be able to use technology for creativity and innovation, communication and collaboration, research and information fluency, and critical thinking, problem-solving, and decision-making. For K-12 classrooms to address these standards, instructional approaches must facilitate active, resource-rich, student-centered learning environments that help students learn to think critically, analyze and synthesize information to solve technical, social, economic, political, and scientific problems, and work productively in groups (Mills, 2006). Classroom learning activities should focus on rich, multidisciplinary learning tasks that address complex sets of learning objectives.

One way for K-12 schools and classrooms to address the NETS-S is by deploying Web 2.0 technology in the classroom. Web 2.0 is useful for implementing complex learning objectives because of its ability to combine technology with collaboration to facilitate functional communication with professionals, subject experts and other students who have Internet access. Using Web 2.0 technologies, information can be manipulated and meshed to generate knowledge for creating new solutions, solving problems or making decisions because Web 2.0 permits substantive information exchanges and collaborations among peers and with subject-matter experts outside of the classroom.

Some Web 2.0 technologies that are commonly used by today's students that have broad subject-matter applications include blogs, wikis, Web feeds, podcasts, and Web conferencing. For example, blogs and wikis can enhance writing experiences. With a school or classroom blog students can receive editing advice and content corrections from peers. Wiki software can be used to enhance research and writing objectives by enabling students to add a section of research to a document being created by a group or class. The live editing of the document permitted by the wiki allows for writing and research revisions through peer editing. Thus, students strengthen writing skills and sharpen research skills because they are engaged in learning through a collaborative communication process.

With RSS or Web feeds, teachers and administrators can communicate news, events, and activities to students and parents. For example, school administrators can send announcements about school events to parents and teachers can send homework assignments and other classroom reminders to students. Podcasts can be used in the same way as RSS feeds to mass communicate with students and parents. With podcasts, teachers and administrators use pre-recorded audio or video messages. Teachers can also use podcasts to record classroom instruction to create an archive for students to review for exams or for absent students to access and use for make-up work.

Web 2.0 has applied the complex technologies of videoconferencing to the personal computer. Communication at a distance that used to require expensive decoding equipment can now be performed with webconferencing software via most personal computer connections to the Internet. With webconferencing guest speakers can be brought into the local classroom from anywhere in the world there is an Internet connection. Students can speak with scientists and other content experts or even students from classes in different countries.

The vast resources of data and information on the Internet supply tools and resources that permit application of a broader and more powerful set of instructional methodologies in the classroom. Using Web 2.0 technologies to support teaching and learning makes it possible to use powerful methodologies such as cases, projects, and problems that are relevant and representative of real-world tasks. However, using Web 2.0 to implement these more powerful instructional methodologies requires teachers to create learning experiences that align with the classroom curriculum.

Because Web 2.0 technologies best support instructional methodologies that employ collaborative and authentic learning experiences, implementing Web 2.0 technologies in the K-12 classroom will generally take the form of project- or problem-based learning (PBL). Therefore, the following sections of this chapter will first establish a framework for PBL that demonstrates its effective implementation when embedded with Web 2.0 capabilities and then will describe some Web 2.0 technologies that are effective for developing PBL experiences.

LEARNING THEORY, PROBLEM-BASED LEARNING, AND WEB 2.0

Cognitive learning theories influence most modern pedagogy. Cognitive learning theories place more emphasis on factors that are internal to the learner than behavioral theories, which place more emphasis on factors within the environment (Smith & Ragan, 1999). Information processing theory is a cognitive learning theory that has provided an important contribution to the field of instructional design and development (Smith & Ragan) as well as to learning and developmental psychology (Bransford, Brown, & Cocking, 2000).

Information processing theory hypothesizes a set of structures in the brain that work much like a computer. For learning to occur, a series of transformations of information takes place in or through these structures. Like a computer, the human brain receives information into working memory, performs operations on the information to change its form and content, and stores it in long-term memory, then locates it and generates responses to it.

Information processing theory suggests that there is a two-way flow of information as we try to make sense of the world around us (Huitt, 2003). We use information that we gather through the senses and information we have stored in memory in a dynamic process to construct meaning about our environment and our relations to it. Information processing theory defines the learning process as based on integrating or assimilating information into long-term memory in a meaningful way that includes gathering and representing information, called *encoding*, storing information, called *retention*, and then getting at the information when needed, called *retrieval*.

Transfer is an important process for acquiring a deep or meaningful understanding of learned information. Transfer of learning occurs when a learner unconsciously or deliberately applies knowledge or skills associated with one task to the completion of another task. In one sense, the purpose of learning is to transfer knowledge and skills from one situation, context, or domain to another. Transfer of learning can occur along a continuum of varying degrees of difficulty where transfer of certain knowledge or skills can be developed to a high-level of automaticity while transfer of other knowledge or skills requires conscious, deliberate efforts.

PBL is an instructional methodology that is fully compatible with information processing theory. With PBL students construct an individual understanding of the problem and then develop and present a solution. PBL is well-suited to encoding and transfer because it situates learning in real-world problems and develops solutions through collaborative processes.

Web 2.0 technologies can be used to support problem-based learning in the K-12 classroom. For example, Web 2.0 applications can provide information tools and resources that can be deployed to simulate real-world problems by providing both content and context for a problem. Using Web 2.0 technologies, the Web can be explored to identify and acquire the information necessary to understand the problem and students can communicate with outside experts or peers to develop solutions to the problem.

With PBL the teacher guides students through a problem-solving process. Students may first reason through the problem and apply knowledge they already have to the problem. This elaboration of prior knowledge helps students understand what information they need to acquire to better understand and resolve the problem. As students begin to research and acquire information about the problem and reach possible solutions, they develop the information literacy skills they need to become self-directed learners.

Learning Strategies and Problem-Based Learning

Many learning theorists believe that working memory has limited capabilities; therefore, it is necessary for learners to employ strategies to regulate their learning when performing learning tasks. For example, information processing theory assumes that that a control mechanism is required to oversee the encoding, transformation, processing, storage, retrieval and utilization of information because of mental constraints on the amount of information that can be processed and the automaticity of the task. Therefore, not all of the processing capacity of the system is available because an executive function that oversees this process will use up some of this capability when one is learning a new task or is confronted with a new environment.

Learning, then, is facilitated by executive functions called cognitive and metacognitive strategies. Cognitive strategies include mental activities such as acquiring, selecting and organizing information, rehearsing material to be learned, relating new material to information in memory, and retrieving and retaining different kinds of knowledge. Metacognitive strategies deal with strategic learning or "learning to learn." Some common metacognitive strategies include connecting new information to former knowledge, deliberate selection of cognitive strategies, and planning, monitoring, and evaluating cognitive processes.

Problem-solving and related research activities can provide opportunities for developing the cognitive and metacognitive strategies that facilitate learning. Students can learn how to learn by developing a repertoire of cognitive or thinking processes that can be applied to solve problems. As students perform problem-solving activities, teachers can focus student attention not only on solutions (products) but also on how tasks are accomplished (processes).

When using problem-solving activities in the K-12 classroom, learning strategies can be embedded in the instructional content and procedures so that these process and product goals are accomplished and evaluated. Web 2.0 can provide technology tools teachers can use to mediate instruction and instructional procedures effectively. Web 2.0 technologies such as social bookmarks, blogs, and wiki can be embedded into PBL experiences to scaffold the cognitive and metacognitive strategies that result in potentially powerful, high-impact learning experiences.

Collaboration and Problem-Based Learning

PBL focuses on challenging problems or tasks that can improve higher-order thinking skills. Because PBL is often used to perform complex tasks or solve ill-structured or ill-defined problems, a student working independently may not possess the knowledge, skills, or time to accomplish the task. Therefore, PBL can be used to help students learn to work together collaboratively and cooperatively.

Collaborative learning refers to a variety of educational approaches that involve shared intellectual efforts by peers or experts. Collaborative learning generates dialog and interaction among peers or communities of peers and experts for the purpose of constructing collective knowledge or shared understanding about a concept, case, or problem. Schrage (1990) defined collaboration as "two or more individuals with complementary skills interacting to create a shared understanding" (p. 40). With collaborative learning, peers are responsible for one another's learning as well as their own learning. Peers work in groups of two or more to search for a mutual understanding, solution, or meaning and to create a product of their shared learning experience.

Collaborative learning activities can range from classroom discussions that may include short lectures to participation on research teams. Generally, collaborative learning activities are designed to encourage interaction among students and promote consensus-building in reaching a shared solution. Through collaboration and cooperation, peers investigate subject matter at varying levels. Collaborative learning builds student awareness of different perspectives. By justifying and defending their ideas to peers, students build deeper knowledge and understanding of a topic. Collaborative learning is especially appropriate for complex and ill-defined problems because the construction of complex knowledge seems to be facilitated by collaborative processes (Feltovitch, Spiro, Coulson, & Feltovich, 1996). Collaborative problem-solving can be supported with Web 2.0 communication tools that facilitate the exchange of ideas among participants.

Group learning can build collective knowledge based on shared problem-solving, interpersonal feedback, and social support and encouragement. Small groups can work together; collaboration and teamwork can be facilitative and can provide scaffolding for the construction of knowledge as well as enhance student satisfaction and learning (Doran, 2001). To form and use groups effectively for collaborative learning, teachers should employ a number of techniques including incorporating team-building activities at the beginning of the year, establishing a feedback process, requiring groups to report progress to the teacher online or face-to-face, evaluating group experiences and providing evaluation or assessment information to the teacher, and employing multiple instructional strategies with group work (Doran).

Collaborative problem-solving using Web 2.0 technologies may utilize teams of students within the same classroom or in classrooms in multiple locations. Therefore, forming effective groups is an important aspect of using online communication to increase student participation in learning and to develop a learning community in the classroom. Online communication tools can be used effectively for collaborative problem-solving, even among students in the same classroom working in small groups because students can access information resources, experts, or peers anywhere in the world. Wikis or blogs may work best for teams of students formed within a classroom while webconferencing, podcasts, and social networking may work best with distant participants.

Collaborative problem-solving in the K-12 classroom can maximize learning by organizing instruction according to the abilities and learning needs of students. When students work in groups, they bring different abilities and expertise to the learning dynamic. Group learning can promote academic achievement by allowing students to encourage each other, ask questions of one another, require each other to justify opinions and reasoning, and reflect upon their collective knowledge, (Brown, A. & Palincsar, A., 1989; Cohen, 1994; Johnson & Johnson, 1994). Collaborative learning is easy and inexpensive to implement and can promote improved behavior, attendance and positive attitudes about school (Slavin, 1987).

Having students work in groups encourages discussion and develops social skills useful for a professional working environment for which they are training. Group collaboration takes advantage of learner-learner interactions rather than learner-content interactions for learning. Small groups working on a project learn the team-building skills and goal-setting strategies needed to be productive in the workplace. Collaborative learning activities can increase students' satisfaction with the learning process and can decrease the time required by the teacher for administering and structuring a course, program, or other unit of instruction.

Collaborative learning is about building learning communities in the classroom, and learning communities can provide an organizing structure and delivery system for the practice of collaborative learning. Learning communities frequently provide more time and space for collaborative learning and other more complicated educational approaches. K-12 classrooms can provide a sense of community by promoting and supporting collaborative learning.

From its beginning the Internet and its associated applications and services have been substantially social and Web 2.0 technologies have served to increase the extent of these social capabilities to make the Web especially connective (Alexander, 2006). According to Alexander it is the affordance of collaborative information discovery that makes Web 2.0 technologies useful for education and instruction. Web 2.0 technologies such as social networking, media sharing, podcasting, and webcasting make it possible to transform classrooms into virtual learning communities by connecting students with peers, experts, and resources beyond the classroom. K-12 classrooms can create collective knowledge through the discoveries of students sharing information, interests and learning objectives via the Internet. This collective knowledge can even transcend a single classroom and a single school year through collaborative, problem-solving activities that allow students to build upon the knowledge and discoveries of predecessors.

WEB 2.0 TECHNOLGIES FOR PROBLEM-BASED LEARNING

Problem-solving consists of moving from an initial, undesired situation to a desired goal and so problem-solving is a process of planning and executing a set or series of steps to reach the goal (Moursund, 1999). There are numerous Web 2.0 applications and services that are available to support collaborative information problem-solving that can engage students in meaningful, challenging, and motivating inquiry and critical thinking.

In this section we will consider several Web 2.0 technologies that have particular application to instruction, especially in developing collaborative PBL experiences for the K-12 classroom. This discussion is not meant to be a comprehensive list of Web 2.0 technologies but rather a substantive discussion of Web 2.0 technologies that are most familiar to today's students and have features and capabilities that are relatively simple to use and to embed in PBL experiences.

RSS and Aggregators

RSS stands for Rich Site Summary or Really Simple Syndication. RSS is a standardized Web format and application used for the distribution of content from Web pages. In the recent history of the Web, RSS technology was used primarily for content feeds from blogs and other online content sites like newspapers. Now RSS technology plays a critical role in several Web 2.0 technologies such as podcasts and social networks and the Web 2.0 applications that use RSS technology continue to increase. Therefore, it is important to have a basic operational knowledge of RSS in order to effectively use many Web 2.0 applications.

Feeds are the actual content items that are posted on a website. RSS syndicates all of the content and will send everything that has been requested. Any site with one of the following icons will generate RSS feeds: - Feed Icon; XML - XML Icon; and RSS - RSS Icon.

Another type of feed is the Atom feed. Atom was developed as an alternative to RSS because of some dissatisfaction with RSS. Atom is widely-adopted but incompatible with RSS. The Atom feed can be downloaded by websites that syndicate content from aggregators that subscribe to Atom feeds. The intention of Atom was to ease the difficulty of developing applications with web feeds.

When feeds are received, there must be some way of viewing the content associated with the feed. An application that collects RSS feeds for later viewing is called an aggregator, feed reader, or news reader. An aggregator is a program or website that collects RSS feeds for viewing or reading. Aggregators can be either Web-based, an extension of a browser or email program, or a stand-alone program that is installed on the user's computer or desktop.

Desktop aggregators are designed to maintain RSS subscriptions and then collect Web feeds and group them together using a user-friendly interface, usually resembling the interface of popular e-mail clients. Web-based aggregators are hosted on remote web servers and use a web service to maintain subscriptions and receive the Web feeds into an account the user sets up on the Web-based aggregator website. The advantage of a Web-based aggregator over a client aggregator is that it is available through the Web and so it can be accessed anywhere by a user with an Internet connection. Some Web-based aggregators that can be used in schools and classrooms are Bloglines (http://www.bloglines.com), Netvibes (http://www.netvibes.com/) and Google Reader (http://www.google.com/reader). Microsoft Outlook, Firefox, Internet Explorer, and Safari all have reader extensions that can be embedded into the applications.

An example of how RSS feeds and aggregators can be used in the classroom can be demonstrated by an English teacher who is teaching a writing unit on current events. The teacher makes the assignment for students to read articles on current event topics from several major newspapers and blogs and then write a theme about the issue. The teacher can subscribe to the RSS feeds from several relevant and appropriate blogs and newspapers. Students can then open the aggregator and read the stories without having to find and retrieve the content from its original source (either paper or digital). Additionally, the teacher can delete any content from the aggregator that is not relevant to the assignment or appropriate to the age-level of the student.

RSS is becoming a valuable tool for collecting and viewing digitally archived knowledge. RSS feeds can provide teachers and students with the opportunity to evaluate and filter content on the Web. For example, RSS can help teachers manage and review student postings on a school blog. When combined with other Web 2.0 technologies, applications, and services, RSS feeds can help focus and refine the process of conducting research and generating content based on new information.

Social Bookmarking and Folksonomies

According to Wikipedia, social bookmarking is a method for Internet users to store, organize, search, and manage bookmarks of Web pages on the Internet with the help of metadata ("Social Bookmarking," 2008). Social bookmarking is particularly useful for collecting Internet resources (links to resources) that are to be shared with others. Social bookmarking allows users to save bookmarks (links to Web pages known as URLs or Website addresses) that may be accessed in the future to a public website where it is annotated with descriptive information, tagged with keywords or descriptors, and designated as public or private. Registered users of the website can browse or search the website and view the public bookmarks, tags, and classification schemes that other registered users have created.

The tags used by social bookmarking services are folksonomic. Folksonomic tagging is the practice of collaboratively creating and managing classifications or categorizations of content or data by the creators and consumers of the content as opposed to the traditional indexing schemes that are created by subject-matter experts. Thus, a folksonomy is a user generated taxonomy—a taxonomy created by the "folks" who use it.

Folksonomic tagging is intended to make information easy to search and navigate because it uses a shared vocabulary that is originated by its primary users. Folksonomies arise where user-generated content such as pictures or videos is shared or where existing content such as websites, books, scholarly works, and blog entries are collaboratively tagged.

A shift from the use of formal taxonomies or indexes to folksonomic taxonomies has important implications for teaching and learning because it changes the way information is stored and retrieved. In other words, it may become less important for students to know where information is found and more important to know how to retrieve it using a framework created by their peers.

Social bookmarking was launched by the advent of Joshua Schacter's del.icio.us (Alexander, 2006). Del.icio.us (http://del.icio.us/) is a service for storing, describing, and sharing bookmarked web pages online, which allows access to the bookmarks and additions to the bookmarks from any computer. Users register with del.icio.us and create their own del.icio.us account (my del.icio.us). Users can annotate each URL with a line of text describing the bookmarked URL and tag it with one or more keywords or descriptors to help organize and remember the bookmark. A user can be a group; thus, a group of students or a whole class can have a shared del.icio.us account to archive Internet-based research pertaining to a specific project or multiple projects.

With del.icio.us bookmarks can be shared publicly (or not) so that peers and other people can view them for reference and collaboration. Additionally, a user can browse and search del.icio.us for interesting and useful bookmarks of other users, which is made easy with tags, and then add other's bookmarks to one's own collection. RSS feeds can be found on the bottom of almost every page within del.icio.us and the index can be used to subscribe to other people's feeds.

Educators are finding numerous ways to use social bookmarking in the classroom. For example, the teacher can research existing sites to find appropriate and relevant articles, documents, or multimedia resources. The teacher can then direct students to conduct research for a project using the designated bookmarks. Another approach is for individual students, student teams, or even the whole class to use social bookmarking services for collaboration and sharing common information in regard to a specific project or multiple projects. Students select and archive their own resources on a given topic and then share those links with their peers. A group account can be established or students can join groups with similar interests.

Some social bookmarking sites allow teachers to review and comment on resources the students have bookmarked. Other sites allow users to post notes directly to the account's Web page and then teachers can verify if a student has linked to online resource and if they have understood what they have read or viewed based on the notes they have posted. Schools or classes can then place links in their home page to a social bookmarking service. A school may want to share social bookmarking accounts among the teachers in a certain department or even between different departments for interdisciplinary projects. Some social bookmarking sites provide citation services that will create a bibliography on a new Web page so bibliographic information can be cut and pasted into a document.

Some other social bookmarking sites that can be used in schools and classrooms include Edtags (http://www.edtags.org), Diigo (http://www.diigo.com/), and Furl (http://www.furl.net/).

Social Writing: Blogs and Wikis

A blog (short for Web log) is a tool that allows authors to quickly and easily publish (or post) content similar to that of a diary or journal on the Web. According to Wikipedia, blogs are usually maintained by an individual with regular entries of text commentary or other material such as graphics or video that are displayed in reverse chronological order ("Blog," 2008). While many blogs provide commentary or news on a particular subject, others function as more personal online diaries.

Blog sites are organized like conventional websites. A blog page may include text, graphics, and navigation links much like a standard Web page. Blog sites, however, are more dynamic than standard websites because new content is posted to a blog on a daily (or more often) basis. Each new blog entry or posting starts a thread for subsequent comments (responses) made by persons reading the blog entry. Postings are often short and frequently updated and may contain text, images, and links to other blogs, Web pages, or media related to its topic. Postings appear in reverse chronological order and can include archived entries.

Edublogs (http://www.edublogs.org/) is an example of a blog service that provides online hosting for educational blogging. Edublogs is a free service without any advertising, ample uploading space, and numerous features and currently hosts hundreds of thousands of blogs for teachers, and students and other educators. With Edublogs students and teachers can upload or copy and paste information into the blog page. The user can manage who gets access to the blogs through password and plugin safety measures. Edublogs can be used to connect multiple student and teacher blogs and the teacher can manage and edit blog posts and responses through an administrative panel. Edublogs permits the embedding of online video, multimedia presentations, and slideshows into blogs. Multiple blogs are easily created so it is easy to set up a dedicated blog for a specific project or event. Edublogs can even be used to create a multi-layered, multimedia-rich website that does not appear to resemble a blog.

Wikis take social writing and interaction a step further by allowing collaborative editing of a document on the Web. According to Wikipedia ("Wiki," 2008) a wiki is a collection of Web pages designed to enable anyone who accesses it to contribute or modify content. Wiki platforms can be used to create collaborative websites and power community websites. One of the best known wikis is Wikipedia, a collaborative, online encyclopedia. A wiki can be edited by its readers while a blog is written by one person and everyone else reads it and makes comments.

Wikis permit groups to work collaboratively on the content of the site using nothing but a standard Web browser. The wiki platform tracks the history of a document as it is revised. When a revision to the content takes place, the revised version becomes the current version and an older version is archived.

Blogs and wikis work well in classrooms because they are amazingly simple to implement and easy to use, requiring minimal technical knowledge. Blogs can be effortlessly created, edited, and updated at any time from any place from a computer with Internet access. Thus, teachers can extend the boundaries of classroom learning to where students live and play. Blogs can be a powerful tool for enabling learning because they provide authentic and convenient opportunities for students to read, write, and discuss collaboratively.

Teachers can use blogs or wikis as a portal to create and cultivate the K-12 classroom as a community of learners. Blogs can be used not only to scaffold learning but to perform the management and coordination tasks of a learning community including the posting of handouts and assignments. Blogs can provide a space where teachers and students work together to accomplish learning goals. Peer review is a basic or embedded feature of a blog and teachers can provide online mentoring for student blogs. Blogs can facilitate substantive discussions about a topic or issue both within and outside of the classroom because they provide the opportunity and time for students to reflect on learning. Wikis can be used to allow students to collaborate on a group report, compile data, or share the results of their research. Teachers in a particular grade-level or department might use a wiki to develop curriculum, class assignments, or lesson plans.

Some other blog and wiki sites that can be used in schools and classrooms include Class Blogmeister (http://classblogmeister.com/), Bloglines (http://www.bloglines.com/), and Wikispaces for Educators (http://www.wikispaces.com/site/for/teachers).

Social Networking and Media Sharing

Boyd and Ellison (2007) define a social network site as a Web-based information-sharing service that allows individuals to construct a profile within a restricted system, delineate users with whom they share a connection, and view and navigate a list of connections and those made by others within the system. The primary feature of social network sites, according to Boyd and Ellison, is not so much that users are permitted to meet strangers, but that users are enabled to publicly declare their social networks. Myspace (http://www.myspace.com) is the most popular example of a social network.

Safe social networks, or smart social networks as defined by Yarmosh (2006) place certain limitations on the connectivity among users. A safe or smart social network permits users to manage who contacts them, define who can comment on posts and pictures, and remove their profile from the search index. Yarmosh says that smart social networks bring intelligence into the network by permitting users to define and manage their online relationships, which makes smart social networks more appealing. Facebook (http://www.facebook.com) is widely used in higher education circles and permits users to control how they share their information and who can see it. The safe or smart social network is the better application for K-12 classroom integration of social networking.

The use of social networking in education focuses on allowing students to construct a learning landscape. Users can create new social networks in very little time, with no technical skill required. Once a name and a URL are selected, the user indicates whether the network is private (only invited people can view or join) or public; writes a tag-line and description of the network; assigns keywords; chooses from a selection of features (such as photos or videos, a blog, events, groups, or gadgets) and uses drag-and-drop tools to place those features on the page; chooses a visual theme (colors, fonts, sizes) and can customize these choices; and decides what information users will be asked to provide to join the network. If membership is restricted, the creator can invite individuals to join. RSS feeds let users subscribe to updates from specific parts of the social network.

Today's students spend countless hours on Facebook and MySpace. Using social networking in the classroom provides an opportunity for teachers to take advantage of an experience that is familiar and comfortable to students in order to engage them in learning experiences. Social networks are a model for how a community of learners functions because they establish a social and collaborative environment that meets the needs of its members and they establish standards for interaction among its members. By creating social networks around academic topics or projects, a teacher can facilitate a strong sense of community that facilitates personal interactions with the goal of creating collective knowledge. Social networks in the K-12 classroom can provide an opportunity for students to cultivate and sustain a network of peers similar to a network of professional contacts and relationships and to view those relationships in the broad context of learning.

Elgg (http://www.elgg.org/) defines itself as a personal learning landscape—a social network platform with multiple capabilities including eportfolios, tagging, blogging, podcasting, and an RSS reader. Elgg combines capabilities for personal Web publishing with the capabilities of social networking. Elgg differs from a regular blog or a social network by giving users control over who can access their content. Each profile item, blog post, or uploaded file can be assigned its own access restrictions from fully public to only readable by a particular group or individual. Elgg users can register a free account on Elgg.net, have their own installation of the application hosted on Elgg.net, or download Elgg and host it on their own Web server. However, a certain amount of technical knowledge is required to maintain an Elgg server.

Elgg is designed to provide deep learning through conversational immersion over time. According to Campbell, Ammann, and Dieu (2005) learners are encouraged to write a weekly blog on topics that are relevant to their own interests. They can then tag their posts with keywords and search for other Elgg users who are writing about similar topics. Elgg users can be added to the contact lists, can join communities that are relevant to their interest, and can use the file repository to share audio messages, photos, or short videos. Teachers can take part in the activities as peer facilitators to help students make connections and to provide comments and feedback. Teachers can also monitor and participate in student activities.

As the phenomenon of social media has expanded, websites that were focused only on media sharing began deploying social networking features and becoming social network sites themselves (Boyd and Ellison, 2007). Some examples of media sharing sites with social networking features include Flickr (photo sharing), Last.FM (music listening), and YouTube (video sharing).

The photo-sharing site Flickr (http://www.flickr.com/) can be used in K-12 classrooms as a resource for images used in presentations, projects, and portfolios. Because many of the images uploaded to Flickr carry a Creative Commons license, they are suitable for educational use. The folksonomic tagging of images makes it much easier for students to find relevant content. Students can also use Flickr to publish their digital photography to a wider audience. Flickr includes the capability to add hot-spots to an image so that portions of images can be annotated.

VoiceThread (http://www.voicethread.com/) allows users to create an online media album that can hold multiple forms of media including images, text, and videos. Commentaries can be added to the media using a mixture of voice using a microphone or telephone, text, audio file, or video using a Webcam and shared or exported to an archival movie for offline use as a DVD or MP3 file. A VoiceThread allows group conversations to be collected and shared in one place, from anywhere in the world. VoiceThread has made its premium account available to K-12 educators for free.

Some other social networking and media sharing sites that can be used in schools and classrooms include Classroom 2.0 (http://www.classroom20.com/), Ning (http://www.ning.com/), and Imbee (http://www.imbee.com/).

Multimedia Broadcasting: Podcasting, Webcasting, Webconferencing, and Video Blogging

A podcast is an audio mini-program, in MP3 format, broadcast over the Internet. Podcasts can be downloaded and listened to on any MP3-compatible digital music player such as Apple's iPod. Users can either download a podcast once or subscribe to the RSS service for regular or periodic downloads. Podcasts can also be downloaded to a computer using podcasting applications such as iPodder. The producer creating the podcast, called a podcaster, can easily create podcasts with a microphone, a computer, videoediting software, a tool to generate RSS files, and ample server space to host large files. Podcasting is useful for the K-12 classroom because it easy and generally free for listeners (students) and minimal costs for producers (teachers).

Podcasting allows instruction to become portable. Podcasting is useful in the K-12 classroom because students are already familiar with the underlying technology. Instructional podcasting can promote the use of iPods and MP3 players not only for entertainment but also for educational experiences. For example, with podcasting teachers can create or record lesson content for students who miss class or provide access to subject-matter experts through interviews. Additionally, students can create their own podcasts as a record or log of classroom or project activities, a way to collect notes for a project or on a lesson, or a reflection on what they have learned.

A webcast is a video/audio file that is distributed over the Internet using streaming media technology. A webcast can be either be live or recorded, and can be distributed live (synchronous) or recorded (asynchronous). Essentially, webcasting is the transmission of linear audio or video content over the Internet to many simultaneous listeners or viewers.

With the emergence of Web 2.0 technologies webcasting provides a relatively simple and inexpensive option for synchronous and asynchronous mass communication anywhere in the world. For most online events that use webcasting, all that is needed is a computer with an Internet connection and speakers. While connecting to a webcast is a relatively simple process, transmitting live or recorded webcasts requires the capability to capture and produce video content—usually using a video camera, firewire, and videoediting software. The final cut of the video production must then be encoded to convert it to a streaming format using encoding software such as Windows Media Encoder for a Windows-based computer broadcasting a Windows Media stream or QuickTime Broadcaster for a Macintosh computer broadcasting a QuickTime stream.

Webconferencing is similar to webcasting but with functionality such as electronic whiteboards and chat. Webconferencing offers a way to engage students in fully interactive, online learning experiences; however webconferences can also be used for tutoring and online office hours.

Webconferencing applications use common browser plug-ins and connect through a local or remote hosting service. At the scheduled time, participants log on to a website to join class sessions, participate in online office hours, or take part in other scheduled events. The webconference window usually includes a pane that lists current participants, a chat pane for written interaction, an audio/video pane, and a content window. The audio/video pane can show only the instructor or presenter or it can include other users if they have webcams. Many applications use voice over Internet protocol (VoIP) for the audio segment, eliminating the need for a separate phone connection. The content pane shows applications from the presenter's desktop, which can include text or multimedia resources.

Presenters can be seen and heard in real time by session participants, who can communicate with one another and the instructor through the chat pane, the audio and video, or tools such as a shared whiteboard. The instructor can respond to questions from participants, demonstrate applications, and share access to them in the content pane, as well as manage the layout of the environment. Sessions can be recorded and archived for later access and can be converted for playback on portable devices.

Webconferences can be used in K-12 classrooms in a number of ways. Webconference technology allows distant groups of students to interact over the Web, work on shared topics, and build a learning community. Webconferences offer an easy way to bring subject-matter experts into a classroom. Webconferences provide an alternative to asynchronous online instruction because many students are more engaged when they can see and hear instructors, seek clarification, and communicate in real time.

A videoblog, or vlog, is a blog using video rather than text or audio as its primary media source. Cell phone cameras and digital cameras that record short video sequences or inexpensive video cameras usually provide the raw footage of a videoblog. Videoblogs are usually accompanied by text or still images, and some vlogs include folksonomic tagging. Digital videoediting software is often used to produce high quality video segments that include background music and special effects.

A videoblog is updated regularly like a text blog. Like a text blog, a videoblog offers a simple mechanism for subscription and delivery through RSS feeds. Videoblogging can offer a richer and more intense Web experience than text blogging because it combines movies, audio, images, still photos, and text. Because it is becoming easier to record and edit video segments and quickly post them to a website, videoblogs can be a useful tool for recording classroom demonstrations, lectures, and lab experiments. Videoblogs can also be used for personal expression and reflection and are useful for eportfolios, presentations, and digital storytelling.

Some multimedia broadcasting applications or services that can be used in schools and classrooms include Podomatic (http://www.podomatic.com/), Audacity (http://audacity.sourceforge.net/), the podcasting section of iTunes has a category dedicated to education, Microsoft Office Live Meeting (http://office.microsoft.com/en-us/livemeeting/), Wimba Classroom (http://www.wimba.com/products/wimbaclassroom/), and Youtube (http://www.youtube.com/);

CONCLUSION

Web 2.0 has significantly lowered the barriers to access and effective use of technology to supplement and enhance instruction and to impact teaching and learning in the K-12 classroom. It is relatively simple to embed a del.icio.us tag, a classroom blog, or a Web conference with a subject-matter expert into a PBL activity. While the future of Web 2.0 is emergent, the path from the previous generation of Internet technologies to Web 2.0 demonstrates that subsequent generations of Internet technologies may become a normal, if not requisite, part of mainstream pedagogy.

Many of the features of Web 2.0 have great potential for instructional use that makes it more than just another medium for the delivery of instruction. When Web 2.0 applications and services are embedded into a PBL approach to learning, teachers can implement powerful learning experiences that facilitate independent and collaborative student-centered learning experiences. Web 2.0 allows teachers to engage students in learning experiences with authentic and relevant learning contexts. Web 2.0 provides scaffolds that accommodate collaborative PBL.

With Web 2.0, teachers have tools to create high-impact learning experiences where students work in-depth with content to express their knowledge and understanding of the content. Thus, Web 2.0 can play a critical role in the K-12 classroom by establishing PBL environments and experiences that educate students for the 21st century.

REFERENCES

Alexander, B. (2006, March/April). Web 2.0: A new wave of innovation for teaching and learning? *EDUCAUSE Review*, *41*(2), 32–44.

Baird, D., & Fisher, M. (2005-2006). Neomillennial user experience design strategies: Utilizing social networking media to support 'Always On' learning styles. *Journal of Educational Technology Systems*, *34*(1), 5–32. doi:10.2190/6WMW-47L0-M81Q-12G1

Blog. (2008, July 22). In *Wikipedia, the free encyclopedia*. Retrieved from http://en.wikipedia.org/wiki/Blog/

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), article 11. Retrieved July 25, 2008 from http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How People Learn: Brain, Mind, Experience, and School*. Washington, D.C.: National Academy Press.

Brown, A., & Palincsar, A. (1989). Guided, cooperative learning and individual knowledge acquisition. In L. Resnick (Ed.), *Knowledge, learning and instruction* (pp. 307-336). Mahwah, NJ: Lawrence Erlbaum Associates.

Campbell, A., Ammann, R., & Dieu, B. (2005, September). Elgg—A personal learning landscape. *Teaching English as a Second Language eJournal 9*(2).

Cohen, E. G. (1994). *Designing groupwork: Strategies for the heterogeneous classroom*. New York: Teachers College Press.

Doran, C.L. (2001). The effective use of learning groups in online education. *New Horizons in Adult Education*, *15*(2).

Downes, S. (2005, October 17). E-learning 2.0. *eLearn Magazine*. Retrieved from, http://elearnmag.org/subpage.cfm?section=articles&article=29-1

Feltovitch, P. J., Spiro, R. J., Coulson, R. L., & Feltovich, J. (1996). Collaboration within and among minds: Mastering complexity, individually and in groups. In T. Koschmann (Ed.), *CSCL: Theory and practice of an emerging paradigm* (pp. 25-44). Mahwah, NJ: Erlbaum.

Huitt, W. (2003). The information processing approach to cognition. *Educational Psychology Interactive.* Valdosta, GA: Valdosta State University. Retrieved from, http://chiron.valdosta.edu/whuitt/col/cogsys/infoproc.html

International Society for Technology in Education. (2007). *National educational technology standards for students 2007*. Retrieved June 30, 2008 from http://cnets.iste.org/

Johnson, D. W., & Johnson, R. T. (1994). *Learning Together and Alone. Cooperative, Competitive, and Individualistic Learning*, (4th Ed.). Edina, MN: Interaction Book Company.

Mills, S. C. (2006). *Using the Internet for Active Teaching and Learning.* Upper Saddle River, NJ: Pearson Education, Inc.

Moursund, D. (1999). *Project-based learning using information technology.* Eugene, OR: International Society for Technology in Education.

Rainie, L. (2006, September 28). *Digital 'Natives' Invade the Workplace*. Retrieved from http://pewresearch.org/pubs/70/digital-natives-invade-the-workplace

Schrage, M. (1990). *Shared Minds*. New York: Random House.

Slavin, R. (1987). Cooperative learning: Can students help students learn? *Instructor*, (March): 74–78.

Smith, P. L., & Ragan, T. J. (1999). *Instructional Design*, (2nd Ed.). Upper Saddle River, NJ: Prentice Hall.

Warlick, D. (2006, October 15), A Day in the Life of Web 2.0. *Tech & Learning.* Retrieved from http://www.techlearning.com/showArticle.php?articleID=193200296

Wikipedia, the free encyclopedia. (2008, July 23). *Wiki*. Retrieved July 25, 2008, from http://en.wikipedia.org/wiki/Wiki

Wikipedia, the free encyclopedia. (2008, July 24). *Social Bookmarking.* Retrieved from http://en.wikipedia.org/wiki/Social_bookmark

Yarmosh, K. (2006, July 26). Smart Social Networks. *ReadWriteWeb*. Retrieved from http://www.readwriteweb.com/archives/smart_social_ne.php

KEY TERMS AND DEFINITIONS

Blog: Short for Web log, a blog is a Web 2.0 technology that allows authors to quickly and easily publish (or post) content similar to that of a diary or journal on the Web. Blogs consist of regular or periodic entries of text commentary or other material such as graphics or video that are displayed in reverse chronological order. Some blogs provide commentary or news on a particular subject while others function as more personal online diaries. Blog entries are often short and frequently updated. Blogs are organized much like conventional Web pages and may include text, graphics, and navigation links. Each new blog entry starts a thread for subsequent comments made by persons reading the blog entry.

Collaborative Learning: A variety of instructional approaches that involve shared intellectual efforts by peers and/or experts. Collaborative learning generates dialog and interaction among peers or communities of peers and experts for the purpose of constructing collective knowledge or shared understanding about a concept, case, or problem. Peers work in groups of two or more to search for a mutual understanding, solution, or meaning and to create a product based on their shared learning experience.

Folksonomic Tagging: The practice of collaboratively creating and managing classifications or categorizations of content or data by the creators and consumers of the content instead of using traditional indexing schemes that are created by subject-matter experts. Thus, a folksonomy is a user generated taxonomy. Folksonomies arise when existing content such as websites, books, scholarly works, blog entries, pictures or videos are collaboratively tagged. Folksonomic tagging is intended to make information easy to search and navigate because it uses a shared vocabulary that is originated by its primary users.

Podcast: An audio mini-program in MP3 format that is broadcast over the Internet. Podcasts can be downloaded and listened to on any MP3-compatible digital music player such as Apple's iPod. Users can either download a podcast once or subscribe to an RSS service for regular or periodic downloads. Podcasts can also be downloaded to a computer using podcasting applications.

Problem-Based Learning: Also known as project-based learning, PBL is an instructional methodology that helps students construct an individual understanding of a problem and then develop and present a solution. With PBL the teacher guides students through a problem-solving process. Students first reason through the problem and apply knowledge they already have to the problem and then students research and acquire information about the problem and reach possible solutions. PBL generally situates learning in real-world problems and allows students to develop solutions through collaborative processes.

RSS Feeds: Also known as news feeds or Web feeds, RSS feeds are the actual content items that are published on Web pages and generated by RSS. RSS stands for Rich Site Summary or Really Simple Syndication and is a Web 2.0 technology used for the syndication of content from Web pages. RSS allows users to subscribe and receive the RSS feeds and view the content without visiting the original website. With RSS, content from Web pages comes to the user rather than user going to get the content.

Social Bookmarking: A Web 2.0 service for storing, describing, and sharing bookmarked Web pages online, which allows access to the bookmarks and additions to the bookmarks from any computer connected to the Internet. Users register with the social bookmarking service and create their own account comprised of annotated and tagged navigation links (URLs). The social bookmarking service usually permits the bookmarks to be annotated with a line of text describing the link and then tagged with one or more keywords or descriptors to help organize and remember the bookmark. A widely used social bookmarking site is Del.icio.us (http://del.icio.us/)

Social Networking: A Web-based information-sharing service that allows individuals to construct a profile within a restricted system, delineate the users with whom they want to share a connection, and view and navigate a list of connections and those made by others within the system. The primary feature of social network sites is that users are enabled to publicly declare their social networks. Myspace (http://www.myspace.com) is the most popular example of a social network.

Web 2.0: A term used to identify Web technologies that harness collective intelligence, provide interfaces and services across multiple devices, and enhance collaboration. Examples of Web 2.0 applications, services, and technologies are blogs, podcasts, social bookmarking, social networking, Web feeds, and wikis. Although the designation, Web 2.0, suggests a new version or generation of the World Wide Web, in reality it refers to a re-visioning of the Web.

Wiki: A website that permits collaborative editing of a document on the Web. A wiki is a collection of Web pages designed to enable anyone who accesses it to contribute or modify content. When a revision to the content takes place, the revised version becomes the current version and an older version is archived. A wiki is different from a blog because the content of the Web page can be edited by its readers while blog content is written and posted by one person and then everyone else reads it and makes comments. One of the best known wikis is Wikipedia, a collaborative, online encyclopedia.

This work was previously published in Handbook of Research on New Media Literacy at the K-12 Level: Issues and Challenges , edited by L.T.W. Hin; R. Subramaniam, pp. 372-388 , copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.17
Machine Learning and Web Mining: Methods and Applications in Societal Benefit Areas

Georgios Lappas
Technological Educational Institution of Western Macedonia, Kastoria Campus, Greece

ABSTRACT

This chapter reviews research on machine learning and Web mining methods that are related to areas of social benefit. It shows that machine learning and Web mining methods may provide intelligent Web services of social interest. The chapter reveals a growing interest for using advanced computational methods, such as machine learning and Web mining, for better services to the public, as most research identified in the literature has been conducted during the last years. The chapter objective is to help researchers and academics from different disciplines to understand how Web mining and machine learning methods are applied to Web data. Furthermore it aims to provide the latest developments on research that is related to societal benefit areas.

DOI: 10.4018/978-1-59904-657-0.ch005

INTRODUCTION

The Web is constantly becoming a central part of social, cultural, political, educational, academic, and commercial life and contains a wide range of information and applications in areas that are of societal interest. Web mining is the field of data mining that is related to the discovery of knowledge from the Web. The Web can be considered as a tremendously large and rich in content knowledge base of heterogeneous entries without any well specified structure, which proportionally makes the Web at least as complex as any known complex database and perhaps the largest knowledge repository. The vast information that surrounds the Web does not come only from the content of Websites, but is also related to usage of Web pages, navigation paths and networking between the links of Web-pages. All these properties establish the Web as a very challenging area for the machine learning community to apply their methods usually for extracting new knowledge, discovering interesting patterns and enhancing the efficiency of Websites by providing user-demand content and design.

Web mining is a relatively new area, broadly interdisciplinary, attracting researchers from: computer science fields like artificial intelligence, machine learning, databases, and information retrieval specialists; from business studies fields like marketing, administrative and e-commerce specialists; and from social and communication studies fields such as social network analyzers, pedagogical scientists, and political science specialists. Herrera-Viedma and Pasi (2006) denote that due to the complexity of Web research there is a requirement for the use of interdisciplinary approaches like statistics, databases, information retrieval, decision theory, artificial intelligence, cognitive social theory and behavioral science. As a relatively new area there is a lot of confusion when comparing research efforts from different point of views (Kosala & Blockeel, 2000) and therefore there is a need for surveys that record and aggregate efforts done by independent researchers, provide definitions and explain structures and taxonomies of the field from various points of view.

The overall objective of this chapter is to provide a review of different machine learning approaches to Web mining and draw conclusions on their applicability in societal benefit areas. The novelty of this review is that it focuses on Web mining in societal benefit areas. There exist similar work related to Web mining in (Baldi, Frasconi, & Smyth, 2003; Chakrabarti, 2003; Chen & Chau, 2004; Pal, Talwar & Mitra, 2002). Baldi et al. (2003) cover research and theory on aspects of Internet and Web modeling at the information level based on mathematical, probabilistic, and graphical treatment. Chakrabarti focuses on studies that connect users to the information they seek from the Web providing lots of programs with pseudocode. Chen and Chau provide an extended review of how machine-learning techniques for traditional information retrieval systems have been improved and adapted for Web mining applications. Pal et al. (2002) present an overview of machine learning techniques with focusing on a specific Web mining category, the Web content mining that will be described in next section. This work is differentiated from the aforementioned related work as the chapter particularly focuses on Web mining and machine learning that may help and benefit societal areas in ways of extracting new knowledge, providing support for decision making and empowering valuable management of societal issues. This survey aims to help researchers and academics from different disciplines to understand Web mining and machine learning methods. Thus, it is aimed at a relatively broad audience and tries to provide them with a different and more open view on Web research. Therefore this work addresses researchers from both computer science and other than computer science disciplines with the intention: (a) for computer science researchers, to provide them with the latest developments on the theory and applications of Web mining, focusing also to the need for Web mining applications in societal beneficial areas, and (b) for researchers from other than computer science disciplines, to draw their attention to existing machine learning methods that may help them to seek for more effective results in their Web research.

Later in the chapter, some background to the different perspectives of Web mining has been provided with a short review on machine learning methods. Afterwards, a study on related machine learning methods applied to Web mining have been put forward, which is followed by applications related to societal benefit areas. Finally it discusses current trends and future challenges on machine learning and Web mining.

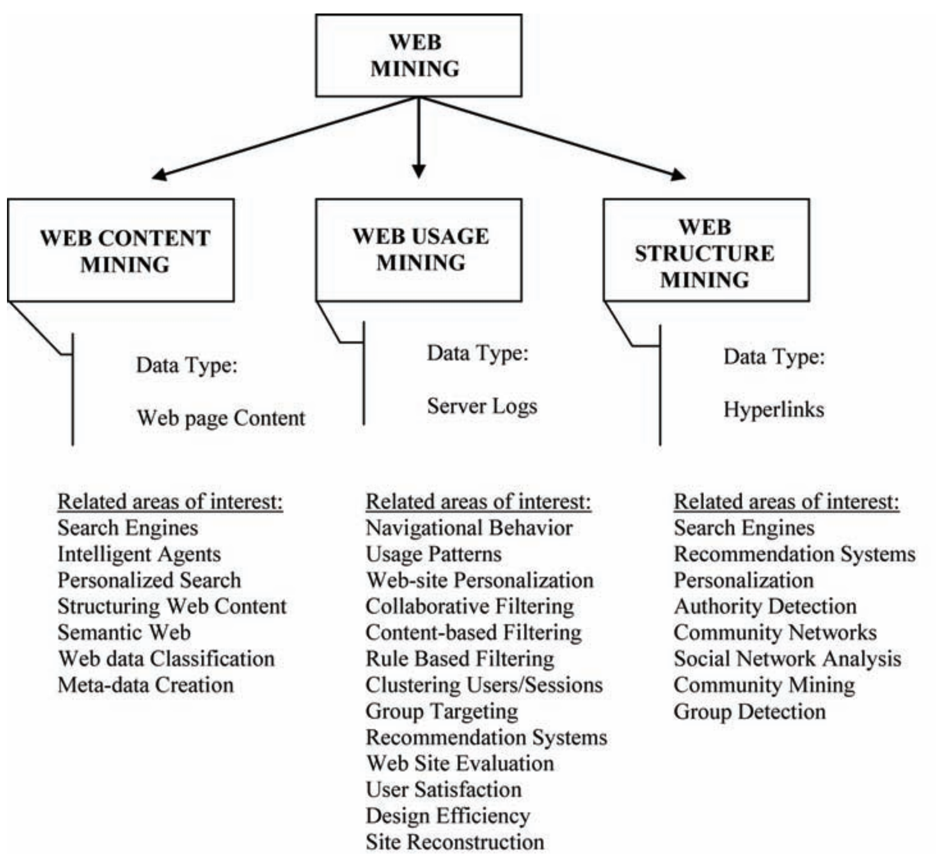
WEB MINING OVERVIEW

The word *mining* means extracting something useful or valuable, such as mining gold from the earth. The expectation of useful or valuable information discovery from the Web is enclosed in the term "Web mining." Definitional, Web mining refers to the application of data mining techniques to the World Wide Web (Cooley, Mobasher & Srivastava, 1997), or else is the area of data mining that refers to the use of algorithms for extracting patterns from resources distributed in the Web. Over the years, Web mining has been extended to denote the use of data mining and other similar techniques to discover resources, patterns and knowledge from the Web and Web-related data (Chen & Chau, 2004).

According to the different sources of data analysis, Web mining is divided into three mining categories. Figure 1 shows the Web mining taxonomy and the sources of data that are used in mining. Moreover, it displays Web mining categories and the related areas of research interest:

a. *Web content mining* focus on the discovery of knowledge from the content of Web pages and therefore the target data consist of multivariate type of data contained in a Web page as text, images, multimedia, and so forth.
b. *Web usage mining* focus on the discovery of knowledge from user navigation data when visiting a Website. The target data are requests from users recorded in special

Figure 1. Taxonomy of Web mining according to source of target data



files stored in the Website's servers called log files.

c. *Web structure mining* deals with the connectivity of Websites and the extraction of knowledge from hyperlinks of the Web.

The above taxonomy is now broadly used in Web mining (Scime, 2005) and has the origins from Coley et al. (1997) who introduced Web content mining and Web usage mining and Kosala and Blockeel (2000), who added Web structure mining.

A well-known problem related to Web content mining, is experienced by any Web user trying to find relevant Web pages that interests the user from the huge amount of available pages. Current search tools suffer from low precision due to irrelevant results (Chakrabarti, 2000). Lawrence and Giles (1999) raise issues related to search engine problems. Search engines are not able to index all pages resulting in imprecise and incomplete searches due to information overload. The overload problem is very difficult to cope as information on the Web is immensely and grows dynamically raising scalability issues.

Moreover, myriad of text and multimedia data are available on the Web prompting the need for intelligence techniques for developing automatic mining using artificial intelligence tools. Such automatic mining is performed by intelligent systems called "intelligent agents" or "agents" that search the Web for relevant information using domain characteristics and user profiles to organize and interpret the discovery information. Agents may be used for intelligent search, for classification of Web pages, and for personalized search by learning user preferences and discovering Web sources meeting these preferences.

Web content mining is more than selecting relevant documents on the Web. Web content mining is related to information extraction and knowledge discovery from analyzing a collection of Web documents. Related to Web content mining is the effort for organizing the semistructured Web data into structured collection of resources leading to more efficient querying mechanisms and more efficient information collection or extraction. This effort is the main characteristic of the "Semantic Web" (Berners-Lee, Hendler & Lassila, 2001), which is the next Web generation. Semantic Web is based on "ontologies," which are metadata related to the Web page content that make the site meaningful to search engines. Sebastiani's study (2002) may be used as a source for Web content mining.

Web usage mining tries to find patterns of navigational behavior from users visiting a Website. These patterns of navigational behavior can be valuable when searching answers to questions like: How efficient is our Website in delivering information? How the users perceive the structure of the Website? Can we predict users next visit? Can we make our site meeting user needs? Can we increase user satisfaction? Can we targeting specific groups of users and make Web content personalized to them?

Answer to these questions may come from the analysis of the data from log files stored in Web servers. A log file is usually a large file that contains all requests of all users to the Website as they arrive in time. Log files may have various formats according to the information stored. The most common format uses information about user IP, date and time of request, type of request (for example get a Web page), a code denoting whether the request has be successfully served or failed, and number of bytes transferred to user. However, Web usage mining should not be confused with tools that analyze log files in order to provide statistics about the site such us: page hits, times of visits, hits per hour or per day or per month, and so forth. While this information might be interesting or valuable for Website owners, they have low data analysis. Web usage mining is more sophisticated as it refers to find users access behavior (Levene & Loizou, 1999) and usage patterns (Buchner, Mulvenna, Anand & Hughes, 1999). It has become a necessity task to provide Web administrators with meaningful information about users and usage patterns for improving quality of Web information and service performance (Eirinaki & Vazirgiannis, 2003; Spiliopoulou & Pohle, 2001; Wang, Abraham & Smith, 2005). Successful Websites may be those that are customized to meet user preferences both in the presentation of information and in relevance of the content that best fits the user.

In this context, Website personalization is the process of customizing the content and structure of a Website to the specific needs of each user taking advantage of user's navigational behavior (Eirinaki & Vazirgiannis, 2003). Recommendation systems support Website personalization by tracking user's behavior and recommending similar items to those liked in the past (content-based learning), or by inviting users to rate objects and state their preferences and interests so that recommendations can be offered to them based on other users rates with similar preferences (collaborative filtering), or by asking questions to the user and providing services tailored to user needs according to the user's answers (rule-based filtering).

On the other hand, content-based filtering is the most common method for Web personalization from server log files and has attracted considerable attention from researchers (Mobasher, Jain, Han, & Srivastava, 1996; Mobasher, Cooley, & Srivastava, 1999; Ngu & Wu, 1997; Spiliopoulou, Pohle, & Faulstich, 1999; Srivastava, Cooley, Deshpande, & Tan, 2000; Wolfang & Lars, 2000) for constructing user models that represent the behavior of users. Such systems usually apply classification methods or clustering algorithms on Web usage data.

Along this perspective, a common methodology for discovering usage and user behavior patterns consists of the following steps: reconstructing user sessions, that is, the navigational sequence of Web-pages of a user in the site; comparing them with other user's sessions; and clustering or classifying the sessions to extract knowledge of navigational behavior. Extracted usage and user behavior patterns may be used in targeting specific groups of users; in various recommendation systems; and in evaluation and reconstruction of the Website to meet design efficiency issues and user satisfaction requirements. Detailed surveys on Web usage mining are presented by Faca and Lanzi (2005), and by Srivastava et al. (2000).

Subsequently, Web structure mining is closely related to analyzing hyperlinks and link structure on the Web for information retrieval and knowledge discovery. Web structure mining can be used by search engines to rank the relevancy between Websites and to classify them according to their similarity and relationship (Kosala & Blockeel, 2000). Google search engine, for instance, is based on PageRank algorithm (Brin & Page, 1998), which states that the relevance of a page increases with the number of hyperlinks to it from other pages, and in particular of other relevant pages. Personalization and recommendation systems based on hyperlinks are also studied in Web structure mining.

Web structure mining is used for identifying "authorities," which are Web pages that are pointed to by a large set of other Web pages (Desikan, Srivastava, Kumar & Tan, 2002) that make them candidates of good sources of information. Web structure mining is also used for discovering "social networks on the Web" by extracting knowledge from similarity links. The term is closely related to "link analysis" research, which has been developed in various fields over the last decade such as computer science and mathematics for graph-theory, and social and communication sciences for social network analysis (Foot, Schneider, Dougherty, Xenos & Larsen, 2003; Park, 2003; Wasserman & Faust, 1994).

This method is based on building a graph out of a set of related data (Badia & Kantardzic, 2005) and to apply social network theory (Wasserman & Faust, 1994) to discover similarities. Thus, a social network is modeled by a graph, where the nodes represent individuals whereas an edge between two nodes represents a direct relationship between the individuals. Recently Getoor and Diehl (2005) introduce the term "link mining" to put special emphasis on the links as the main data for analysis and provide an extended survey on the work that is related to link mining.

A new term, namely, community mining is a major research area on social networks that emphasizes on discovering groups of individuals, who by sharing the same properties form a specific community on the Web. Domain applications related to Web structure mining of social interest are: criminal investigations and security on the Web, digital libraries where authoring, citations and cross-references form the community of academics and their publications etc. Detailed survey on Web structure mining can be found in Desikan et al. (2002) and Getoor & Diehl (2005).

The taxonomy previously described is based on the characteristics of the source data. Usually when working with one of the three data sources (Web content, log files, hyperlinks), researchers might think the corresponding category. However, this is not strict and might combine source data and target application as for example they can use hyperlinks to predict Web content (Mladenic & Grobelnik, 1999). Another example is "Web community" (Zhang,Yu & Hou, 2005), a term closer to Web structure mining, however, is used for the analysis and construction of "Web communities" not only from hyperlinks, but also from Web document content and user access logs. Mobasher, Dai, Luo, Sung, and Zhu (2000) combine Web usage mining and Web content mining for creating user content profiles. Web usage data combined with ontologies and semantics for improving Web personalization are currently proposed in various systems (Berendt, 2002; Dai & Mobasher, 2003; Oberle, Berendt, Hotho & Gonzalez, 2003; Spiliopoulou & Pohle, 2001).

MACHINE LEARNING OVERVIEW

Machine learning is the basic method for most data mining approaches and therefore will be also an important method in Web mining research. It is a broad field of artificial intelligence investigating the use of algorithms acting as intelligent learning methods in computer systems to gain experience, so that this experience can be used when making decisions based on previous learned tasks. The machine learning methods cover a wide range of learning methods, where some of them have been inspired from nature. Neural networks are inspired from human brain and its neurons for the learning, information storing and information retrieval capability. Genetic algorithms and evolutionary algorithms are inspired from Darwin's theory for the surviving characteristics of the fittest in a population that evolves in time. Other machine learning methods are designed to reach to a decision by asking simple yes/no questions following a path from a tree based graph (decision trees) or to derive rules that find interesting associations and/or correlation relationships among large set of data items (association rules).

Representatives of machine learning methods are: artificial neural networks (ANN), self-organizing maps, Hopfield network, genetic algorithms, evolutionary algorithms, fuzzy systems, rough sets, rule-based systems, support vector machines, decision trees, Bayesian and probabilistic models. Describing in details each of these methods will overpass the chapter. The reader can find many textbooks that describe in details all the previously mentioned methods (Bishop, 2003; Duda, Hart & Stork, 2001; Michaklski & Tecuci, 1994; Mitchell, 1997).

At the same time machine learning systems are capable of solving a number of problems related to pattern classification, data clustering, predicting purposes, and information retrieval. In traditional data mining, one can identify that machine learning is used for tackling four types of data mining problems: classification, clustering, association rules and prediction problems.

The task in classification problems is to assign classes to objects according to their characteristics (features). The central aim in designing a classifier is to train the classifier with patterns of known labels drawn out of the total number of available data, which usually are labeled as "positive examples" for samples that belong to a known class and "negative examples" for all those samples that do not belong to the known class. The classifier success is evaluated by the ability to generalize, that is, the ability to predict correctly the label of novel (unseen to the classifier) patterns that have been left out from the training process.

The training process uses an adjustment mechanism that iteratively adjusts the parameters of the classifier in order to get closer to learning the class of the training data. Evaluating classifier generalization one may have an estimation of the performance of the classifier in classifying and predicting labels of newly collected data. Since the class of each data example during the training phase is provided to the classifier, the type of learning is called "supervised learning," where the supervision takes place in adjustments of the classifier parameters so that a misclassified data example in an iteratively learning process is classified correctly.

Classification problems also deal with prediction, as the task in classification is to minimize the error of misclassified test data and therefore classifiers according to the quality of collected data and to the accuracy rate of performance of the classifier may predict classes. In this aspect, prediction is harder when instead of a discrete class one try to find the next value in the range of hypothesis after training the model with historical data, as for example, predicting the closing price of a security in stock-market based on historical financial data.

On the contrary, clustering is a method that uses a machine learning approach called "unsupervised learning," where no predefined classes exist and the task is to find groups of similar objects creating a cluster for each group. Therefore, in a cluster belong data that have similar features between them and at they same time they have dissimilar features with the rest of data. Association rules aims to find relationships and interesting patterns from large data sets.

Although the overwhelming majority of machine learning research is based on supervised and unsupervised learning models, there exist two more types of learning: reinforcement learning and multi-instance learning. Reinforcement learning tries to learn behavior through trial-and-error interactions with a dynamic environment. The difference from supervised learning is that correct classes are never presented, nor suboptimal actions explicitly corrected. In multi-instance learning (Dietterich, Lathrop & Lozano-Perez, 1997) the training data consists of "bags" each containing many instances, while in supervised learning the data set for training consists of positive and negative examples. A bag is labeled positively if it contains at least one positive instance. The task is to learn some concept from the training set of bags for predicting the label of unseen bags. Training bags have known labels, however, the instances have unknown labels so the training process comprises labeled data that are composed of unlabeled instances and the task is to predict the labels of unseen data.

Normally decision trees, and rule-based models are used to solve supervised learning problems; self-organizing maps (SOM), and clustering models are typically used in unsupervised learning problems; and genetic and evolutionary algorithms are typically used in reinforcement learning problems. The rest of machine learning methods are used in both supervised, unsupervised, reinforcement, and multi-instance learning problems.

MACHINE LEARNING APPLIED TO WEB MINING

Machine learning techniques can be very helpful when applied to the process of Web mining. Although there is a close relation between machine learning and Web mining one should denote that Web mining is not actually the application of machine learning techniques on the Web (Kosala & Blockeel, 2000). Other methods studying interesting patterns on the Web may be methods of statistical analysis (Gibson & Ward, 2000; Sharma & Woodward, 2001; Yannas & Lappas, 2005; Yannas & Lappas, 2006), or heuristics (Sutcliffe, 2001).

Primitive Web mining attempts to find patterns to explain various Web "phenomena" can be also found in qualitative Web research methods (Demertzis, Diamantaki, Gazi, & Sartzetakis, 2005; Gillani, 1998; Margolis, Resnick, & Tu, 1997; Maule, 1998; Li, 1998; Reeves, and Dehoney, 1998) that usually rely on observations, annotation of online and archived Web objects, interviews or surveys of Web administrators and users, textual analysis, focus groups, social experiments (Schneider & Foot, 2004) to analyze and explain various Web phenomena.

This approach is usually originated from social science and communication researchers, where the ability to apply more advanced computerized methods like machine learning is limited, however, the interesting of such methods is the expressive power to interpret and explain such phenomena. To the best of our knowledge the author has not identified any combined machine learning and qualitative studies. It could be very interesting to see how such studies will be empowered from the intelligence and automations of machine learning and from the interpretation ability of social sciences.

In comparison to data mining, Web mining may have a few common characteristics similar to machine learning methods and approaches. However, working with Web data is more difficult due to fact that Web data are formed dynamically, change frequently, and their structure cannot be stored in a fixed length database with known features and characteristics. On the contrary, most data mining systems are well structured and remain static over time. Moreover, Web data have many different data type such as text, tables, links, sidebars, layouts, images, audio, video, pdf files, word files, postscript files, executable files, animation files, and so forth, to name a few. Detection of such data types can be a hard problem that needs considerable effort to solve it as with table detection in Websites, where support vector machines and decision trees can be used for attacking this problem (Wang & Hu, 2002). Lastly, the Web is considerably larger than traditional databases in terms of magnitude due to the billions of existing Websites. Before going to the next section, the author presents an indicative research that relates machine learning with the three Web mining categories.

Machine Learning and Web Content Mining

Intelligent indexing text on the Web is the primary goal of search engines building their databases. Machine learning techniques and Web content mining are widely used in this task. Neural networks are commonly used for Web document classification. They are trained by existing Web date for learning to correctly classify patterns of Web documents. They produce high classification accuracy and are very popular among researchers for learning and classifying Web documents (Cirasa, Pilato, Sorbello, & Vassallo, 2000; Fukuda, Passos, Pacheco, Neto, Valerio, & Roberto, 2000; Pilato, Vitabile, Vassallo, Conti, & Sorbello, 2003).

Apart from neural networks classifiers, systems based on support vector machines for Web document classification are presented in Sun, Lim, and Ng (2002), and Yu, Han, and Chang (2002), whereas Esposito, Malerba, Di Pace, and Leo (1999) use three different classification models (decision trees, centroids and k-nearest-neighbor) for automated classification of Web pages; whereas hybrid systems like in Kuo, Liao, and Tu (2005) combine neural networks with genetic algorithms to analyse Web browsing paths for a recommendation system based on intelligent agents.

Also, Bayesian classifiers for text categorization in Syskill and Webert (Pazzani & Billsus, 1997) are used in a recommendation system to recommend Web pages, and in Mooney and Roy (2000) to produce content-based book recommendations. Semeraro, Basile, Degemmis, and Lops (2006) train a Bayes classifier that infers user profiles as binary text classifiers (likes and dislikes) in an application that acts like a conference participant advisor that suggests conference papers to be read and talks to be attended by a conference participant.

Similarly, reinforcement learning and Bayes networks are used as intelligent agents in Rennie and McCallum(1999) for learning and classifying efficiently Web documents. Stamatakis, Karkaletsis, Paliouras, Horlock, Grover, and Curran (2003) compare various machine learning approaches (decision trees, support vector machines, nearest neighbour classifier, naïve baies) for identifying domain-specific Websites.

Machine Learning and Web Usage Mining

Classifiers and clustering algorithms are usually used for analyzing hyperlinks in Web usage mining. Pierrakos, Paliouras, Papatheodorou, Karkaletsis, and Dikaiakos (2003) use clustering applied to Web usage mining for creating community specific directories to offer users a more personalized view of the Web according to their preferences. Hu and Meng (2005) present a system that combines the intelligent agent approach with collaborative filtering using neural networks and Bayes network in order to retrieve relevant information. Zhou, Jiang, and Li (2005) apply multi-instance learning on Web mining by using the browsing history of the user in the Web index recommendation problem for recommending unseen Web pages.

Yao, Hamilton, and Wang (2002) combine three different machine learning techniques: association rules, clustering and decision trees to help users navigate a Website by analysing and learning from Web usage mining and user behavior. A hybrid approach that uses self-organizing maps (SOM), (Kohonen, 1990) and a neuro-fuzzy model is applied on log files by Wang et al. (2005) for Web traffic mining in order to predict Web server traffic. Genetic algorithms are used in (Tug, Sakiroglu & Arslan, 2006) for the discovery of user sequential accesses from log files.

Machine Learning and Web Structure Mining

The most famous application of Web structure mining is the Google search engine based on Brin and Page's (1998) PageRang algorithm for ranking pages relevance. Mladenic and Grobelnik (1999) use the k-nearest-neighbor algorithm to train a system for predicting Web content from hyperlinks. Wu, Gordon, DeMaagd, and Fan (2006) use principal cluster analysis to identify a small number of major topics from millions of navigational data. Lu and Getoor (2003) apply classifiers for link-based object classification. Probabilistic models are used in Matsuo, Ohsawa, and Ishizuka (2001) for Web search and identifying Web communities; in Lempel and Moran, (2001) for Web search; Cohn and Chang, (2000), Getoor, Segal, Tasker, and Koller (2001), and Richardson and Domingos, (2002) for Web page classification.

APPLICATIONS OF WEB MINING TO SOCIETAL BENEFIT AREAS

Web mining may benefit those organizations that want to utilize the Web as a knowledge base for supporting decision-making. Pattern discovery, analysis, and interpretation of mined patterns may lead to take better decisions for the organization and for the provided services. E-commerce and e-business are two fields that may be empowered by Web mining with lots of applications to increase sales, doing intelligence business or even used in crisis management as in Tango-Lowy and Lewis (2005), where Web mining and self organizing maps are used in crisis scenarios.

Lots of Web mining applications found in the literature describe the effectiveness of the application from the Web administration point of view. The target in these applications is taking advantage of the mined knowledge from the users to increase the benefits of the organization. In this chapter, the author focuses on social beneficial areas from Web mining, and hence the point of view is on Web mining applications that can help users or group of users. An obvious societal benefit is that Web mining research efforts lead to user (or group of users) satisfaction by providing accurate and relevant information retrieval; by providing customized information; by learning about user's demands so that services can target specific groups or even individual users; and by providing personalized services. The author identified research on the following areas, where Web mining offers societal benefits: Helpdesks and recommendation systems; digital libraries; security and crime investigation; e-learning; e-government services; and e-politics and e-democracy.

Helpdesks and Recommendation Systems

Recommendation systems are based on user modeling that are mainly derived from content-based learning or from collaborative filtering (Zukerman & Albrecht, 2001). Content–based learning uses a user's past usage behavior and acts as an indicator of his/her future behavior. Collaborative filtering is based on ratings of user favors, like rating music or movies, so that rating history of a user can be associated with similar preferences of other users. So a user is classified in a user model, where recommendations can be addressed to the user according to favors of other people from the specific classified user model. Hybrid recommendation systems that take benefits from both collaborative filtering and content-based learning have been also investigated in literature (Melville, Mooney, & Nagarajan, 2002; Sarwar, Karypis, Konstan, & Riedl, 2000).

Martin-Guerrero, Palomares, Balaguer-Ballester, Soria-Olivas, Gomez-Sanchis, and Soriano-Asensi (2006) propose a recommender model for predicting user preferences based on common clustering algorithms in a citizen Web portal. Clustering and collaboration filtering is used in Hayes, Avesani, and Veeramachaneni (2006) for a blog recommendation system. A blog is a journal-style Website usually written by a single user, where entries are presented in a reverse chronological order.

ReferralWeb (Kautz, Selman & Shah, 1997) is a project that mines social networks from the Web by using collaborative filtering for identifying experts that could answer questions asked by individuals. Nasraoui and Pavuluri (2004) using neural networks provide accurate Web recommendations based on a committee of predictors. Yao et al. (2002) created PagePrompter, an agent-based recommender that helps users navigate a Website by analysing and learning from Web usage mining and user behavior. The interesting part of Pageprompter is that it combines three different machine learning techniques: association rules, clustering, and decision trees for achieving its task.

Pierrakos et al. (2003) use clustering applied to Web usage mining for creating community specific directories to offer users a more personalized view of the Web according to their preferences and may be assisted by using these directories as starting points on their navigation. Garofalakis, Kappos, and Mourloukos (1999) studied Website optimization using Webpage popularity. Scheffer

(2004) created an e-mail answering assistant by semisupervised text classification.

Fast and accurate Web services are practical implications from improved helpdesks and recommendation systems.

Digital Libraries

Digital libraries provide precious information distributed all around the world without necessarily having the need to be physically present in a traditional library building. In this context, Web mining research aiming to offer better services on digital libraries have been identified in literature. Adafre and Rijke (2005) use clustering for discovering missing hypertext links in Wikipedia, the largest encyclopedia on the Web that is created and modified by many volunteer authors. Web mining on Wikipedia is also investigated by Gleim, Mehler, and Dehmer (2006). Bhattacharya and Getoor (2004) use clustering for detecting group of entities, like authors, from links and resolving the coreference problem of multiple references to the same paper in autonomous citation indexing engines, like CiteSeer (Giles, Bollacker, & Lawrence, 1998).

CiteSeer is an important resource for computer scientists for searching electronic versions of papers. Cai, Shao, He, Yan, and Han (2005) work also is based on another well-known digital library in computer science community, the DBLP library at http://dblp.uni-trier.de. Their work is related to machine learning feature extraction algorithms in order to discover hidden communities in heterogeneous social networks. Graph analysis for discovering Web communities can be modeled by using Bayesian networks as demonstrated by Goldenberg and Moore (2005) for identifying coauthorship networks.

A content-based learning book recommendation system is proposed by Mooney and Roy (2000) based on Web pages of the Amazon online digital store. Large portals with news updated frequently per day consist of rich information and may be considered as part of digital libraries in the way that newspaper articles are indexed and available to readers in traditional libraries.

At the same time, news sites are large portal sites that increase their content on a daily basis. For such sites, the interpretation of Web content to meaningful content can be classified into semantic categories in order to make both information retrieval and presentation easier for individuals and group of users is very important (Eirinaki & Vazirgiannis, 2003). Liu, Yu, and Le (2005) use fuzzy clustering to identify meaningful news patterns from Web news stream data. However, the wide distribution of knowledge on the one side and the easiness of access to this knowledge on the other side from various groups of people like researchers, academics, students, pupils, professionals or independents are the most valuable practical implications of Web mining to this societal interest area.

E-Learning

Web mining may be used for enhancing the learning process in e-learning applications. Bellaachia, Vommina, and Berrada (2006), introduce a framework, where they use log files to analyze the navigational behavior and the performance of e-learners so that to personalize the learning content of an adaptive learning environment in order to make the learner reach his learning objective. Zaiane (2001) studies the use of machine learning techniques and Web usage mining to enhance Web-based learning environments for the educator to better evaluate the learning process and for the learners to help them in their learning task. Students' Web logs are investigated and analysed in Cohen and Nachmias (2006) in a Web-supported instruction model. Improved e-learning services that accommodate user needs are practical implications from Web mining and machine learning to the e-learning area.

Security and Crime Investigation

Web mining techniques may be used for identifying cyber-crime actions like Internet fraud and fraudulent Websites, illegal online gambling, hacking, virus spreading, child pornography distribution, and cyberterrorism. Chen, Qin, Reid, Chung, Zhou, and Xi (2004) note that clustering and classification techniques can reveal identities of cybercriminals, whereas neural networks, decision tress, genetic algorithms, and support vector machines can be used to crime patterns and network visualization. Chen et al. (2004) provide a detailed study on methods against terrorist groups on the Web for predictive modeling, terrorist network analysis, visualization of terrorists' activities, linkages and relationships.

Similarly, Wu et al. (2006) based on user's online activities use principal cluster analysis to identify a small number of major topics from millions of navigational data in an approach that can be useful in security against terrorism. Do, Chang, and Hui (2004) implemented a system that can benefit safe Internet browsing in school, home and workplace. The system monitors and filters Web access by applying Web mining for performing Web data classification in order to classify Web data in a "white list" of allowed pages or blacklist of blocked Web pages. Social Networks extracted from instant messaging by using clustering is investigated by Resig, Dawara, Homan, and Teredesai (2004). Enjoying a more secure environment having better online and offline protection are implications of Web mining and machine learning to this societal interest area.

E-Government Services

The processes through which government organizations interact with citizens for satisfying user (or group of users) preferences leads to better social services. The major characteristics of e-government systems are related to the use of technology to deliver services electronically focusing on citizens needs by providing adequate information and enhanced services to citizens to support political conduct of government. Empowered by Web mining methods e-government systems may provide customized services to citizens resulting to user satisfaction, quality of services, support in citizens decision making, and finally leads to social benefits. However, such social benefits mainly rely on the organization's willingness, knowledge, and ability to move on the level of using Web mining.

The e-government dimension of an institution is usually implemented gradually. E-government maturity models (Irani, Al-Sebie & Elliman, 2006; Lappas & Yannas, 2006) describe the online stages an organization goes through time, becoming more mature in using the Web for providing better services to citizens. The maturity stages start from the organization's first attempt to be online aiming at publishing useful citizens' information and move to higher maturity stages of being interactive, making transactions and finally transforming the functionality of the organization to operate their business and services electronically through the Web. But, maturity stages described in literature do not have a Web mining dimension, which the author considers that should be the climax in maturity stages.

Riedl (2003) states that by using interviews and Web mining the actual access to information by citizens should be tracked, analyzed and used for the redesign of e-government information services. E-Government literature reveals that only recently Web mining has attracted researchers in e-government applications. Fang and Sheng (2005) present a Web mining approach for designing better Web portal for e-government. Hong and Lee (2005) propose an intelligent Web information system of government based on Web usage mining to help disadvantaged users make good decisions-making for their profit improvements. In the health sector, Mayer, Karkaletsis, Stamatakis, Leis, Villarroel, and Thomeczek (2006) investigate improvements of health services by quality labelling of medical

Web content in the recently announced MedIEQ project. Conclusively, e-government aim to improve government's services to citizens and any improvement to this direction lead to valuable implications of Web mining and machine learning to national and local societies.

E-Politics and E-Democracy

E-politics provides political information and politics "*on demand*" to the citizens by improving the political transparency and democracy, benefiting parties, candidates, citizens and the society. Election campaigners, parties, members of parliament, and members of local governments on the Web are part of e-politics. Despite the importance of e-politics in democracy there is limited Web mining methods to meet citizen needs. The author has identified in the literature research that only refers in mining political social networks on the Web. Link analysis has been used to estimate the size of political Web graphs (Ackland, 2005), to map political parties network on the Web (Ackland & Gibson, 2004) and to investigate the U.S. political Blogosphere (Ackland, 2005b). Political Web linking is also studied by Foot et al. (2003) during the U.S. congressional election campaign season on the Web. In this aspect, expanding e-democracy boarders will lead to more transparent and participating democracy, which are vital to the society.

FUTURE TRENDS

Nowadays the Web is a rich and huge information repository, where a number of methods and automatic systems have been created for identifying, locating, accessing, and retrieving information. The main open question in Web mining is how to provide information relevant to specific users' needs. Semantic Web (Berners-Lee et al., 2001) works toward this direction and is considered as the next Web generation. The current Web is based on the hypertext mark-up language (HTML), which specifies how to layout Web pages so that they can be readable to humans, thus it is human-centralized. The problem is the retrieval of relevant information by search engines because machines cannot understand Web content to retrieve relevant information. This is expected to change by semantic Web technologies as in semantic Web "information is given well-defined meaning better enabling computer and people to work in cooperation" (Berners-Lee et al., 2001).

Consequently, machine-learning techniques will continue to play the most important role in the semantic Web (Hess & Kushmerick, 2004) for information retrieval and knowledge discovery. Berendt et al. (2002) introduce "semantic Web mining" as the field where semantic Web meets Web mining. It is expected that machine learning techniques and semantic Web mining will be in the focus of research for the next years.

In this chapter the author has introduced areas of societal interest that may be benefited by Web mining and machine learning. The literature review revealed that most research in these areas has just recently been started. The future trend seems to be the convergence of Web mining and machine learning to practical solutions in the six areas of societal benefit: Helpdesks and recommendation systems, digital libraries, e-learning, security and crime investigation, e-government services, e-politics and e-democracy.

CONCLUSION

This chapter has provided a survey on Web mining and machine-learning methods focusing on current Web mining research in societal benefit areas identifying that most of this research has been recently developed. Therefore, one of the current trends of Web mining is toward the connection between intelligent Web services and applications of social benefits, which brings to work closer scientists from various disciplines.

Furthermore, this integrating tendency benefits researchers from various fields.

Social studies on the Web may benefit from machine learning and Web mining methods for providing them with tools and methods to better collect, manage and analyze Web based-phenomena. Moreover, a social interpretation of the meaning of outcomes from computer science Web mining methods is the key question from social and communications studies (Thelwall, 2006). Finally, Web mining and machine learning community may benefit from social and communication expertise on the Web to better interpret their outcomes in the direction of why this happens; or whether mining patterns have meaningful or useful knowledge; or whether hidden knowledge found from Web mining creates a new view that needs further investigation and explanation.

REFERENCES

Ackland, R. (2005). *Estimating the size of political Web graphs*. Revised paper presented to ISA Research Committee on Logic and Methodology Conference. Retrieved April 10, 2008, from http://acsr.anu.edu.au/staff/ackland/papers/political_web_graphs. pdf

Ackland, R. (2005b). *Mapping the U.S. political blogosphere: Are conservative bloggers more prominent?* Paper presented to BlogTalk Downunder, Sydney. Retrieved April 10, 2008, from http://acsr.anu.edu.au/staff/ackland/papers/polblogs.pdf

Ackland, R., & Gibson, R. (2004). Mapping political party networks on the WWW. In *Proceedings of the Australian Electronic Governance Conference*, Melbourne, Australia.

Adafre, S. F., & Rijke, M. D. (2005). Discovering missing links in Wikipedia. In *Proceedings of the 3rd International Workshop on Link Discovery* (pp. 90-97). ACM Press.

Badia, A., & Kantardzik, M. (2005). Graph building as a mining activity: Finding links in the small. In *Proceedings of the 3rd International Workshop on Link Discovery* (pp. 17-24). ACM Press.

Baldi, P., Frasconi, P., & Smyth, P. (2003). *Modeling the Internet and the Web: Probabilistic methods and algorithms.* West Sussex, UK: John Wiley.

Bellaachia, A., Vommina, E., & Berrada, B. (2006). Minel: A framework for mining e-learning logs. In *Proceedings of the 5th IASTED International Conference on Web-based Education* (pp. 259-263). Puerto Vallarta, Mexico.

Berendt, B. (2002). Using site semantic to analyze, visualize and support navigation. *Data Mining and Knowledge Discovery, 6,* 37–59. doi:10.1023/A:1013280719795

Berendt, B., Hotho, A., & Stumme, G. (2002). Towards semantic web mining. *Lecture Notes in Computer Science*, *2342*, 264–278. doi:10.1007/3-540-48005-6_21

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic Web. *Scientific American, 284*(5), 34–43.

Bhattacharya, I., & Getoor, L. (2004). Deduplication and group detection using links. In *Proceedings of the SIGKDD Workshop on LinkAnalysis and Group Detection,* Seattle, WA.

Bishop, C. M. (2003). *Neural networks for pattern recognition.* Oxford University Press.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7th International World Wide Web Conference, Elsevier Science* (pp. 107-117), New York.

Buchner, A. G., Mulvenna, M. D., Anand, S. S., & Hughes, J. G. (1999). Navigation pattern discovery from Internet data. In *Proceedings of the Web Usage Analysis and User Profiling Workshop* (pp. 25-30), San Diego, CA.

Cai, D., Shao, Z., He, X., Yan, X., & Han, J. (2005). Minning hidden community in heterogeneous social networks. In *Proceedings of the 3rd International Workshop on Link Discovery* (pp. 58-65). ACM Press.

Chakrabarti, S. (2000). Data mining for hypertext: A tutorial survey. *ACM SIGDDD Explorations*, *1*(2), 1–11. doi:10.1145/846183.846187

Chakrabarti, S. (2003). *Mining the Web: Discovering knowledge from hypertext data.* San Francisco: Morgan Kaufmann Publishers.

Chen, H., & Chau, M. (2004). Web mining: Machine learning for Web applications. [ARIST]. *Annual Review of Information Science & Technology*, *38*, 289–329. doi:10.1002/aris.1440380107

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., & Chau, M. (2004). Crime data mining: A general framework and some examples. *Computer*, *37*(4), 50–56. doi:10.1109/MC.2004.1297301

Chen, H., Qin, J., Reid, E., Chung, W., Zhou, Y., Xi, W., et al. (2004). The dark Web portal: Collecting and analyzing the presence of domestic and international terrorist groups on the Web. In *Proceedings of the 7th International Conference on Intelligent Transportation Systems (ITSC)*, Washington D.C.

Cirasa, A., Pilato, G., Sorbello, F., & Vassallo, G. (2000). EαNet: A neural solution for Web pages classification. In *Proceedings of the 4th World Multiconference on Systemics, Cybernetics, and Informatics SCI2000*, Orlando, Florida.

Cohen, A., & Nachmias, R. (2006). A quantitative cost effectiveness model for Web-supported academic instruction. *The Internet and Higher Education*, *9*(2), 81–90. doi:10.1016/j.iheduc.2006.03.007

Cohn, D., & Chang, H. (2000). Learning to probabilistically identify authoritative documents. In *Proceedings of the17th International Conference on Machine Learning (ICML2000)* (pp. 167-174), Stanford, California.

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining: Information and pattern discovery on the World Wide Web. In *Procceding of the 9th International Conference on Tools with Artificial Intelligence(ICTAI '97)* (pp. 558-567), New Port Beach, CA: IEEE Computer Society.

Dai, H., & Mobasher, B. (2003). A road map to more effective Web personalization; Integrating domain knowledge with Web usage mining. In [), Las Vegas, Nevada.]. *Proceedings of the International Conference on Internet Computing*, *IC*, 2003.

Demertzis, N., Diamantaki, K., Gazi, A., & Sartzetakis, N. (2005). Greek political marketing on-line: An analysis of parliament members' Web sites. *Journal of Political Marketing*, *4*(1), 51–74. doi:10.1300/J199v04n01_04

Desikan, P., Srivastava, J., Kumar, V., & Tan, P. N. (2002). *Hyperlink analysis: Techniques and applications* (Tech. Rep. TR 2002-0152). Army High Performance Computing Center.

Dietterich, T. G., Lathrop, R. H., & Lozano-Perez, T. (1997). Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, *89*(1-2), 31–71. doi:10.1016/S0004-3702(96)00034-3

Do, T. D., Chang, K., & Hui, S. C. (2004). Web mining for cyber monitoring and filtering. In *Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems Vol. 1* (pp. 399-404). Singapore.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification.* New York: John Wiley.

Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for Web personalization. *ACM Transactions on Internet Technology*, *3*(1), 1–27. doi:10.1145/643477.643478

Esposito, F., Malerba, D., Di Pace, L., & Leo, P. (1999). A learning intermediary for automated classification of Web pages. In *Proceedings of the 16th International Workshop on Machine Learning in Text Data Analysis (ICML1999)* (pp. 37-46).

Faca, F. M., & Lanzi, P. L. (2005). Mining interesting knowledge from Weblogs: A survey. *Data & Knowledge Engineering*, *53*(3), 225–241. doi:10.1016/j.datak.2004.08.001

Fang, X., & Sheng, O. R. L. (2005). Designing a better Web portal for digital government: A Web-mining based approach. In *Proceedings of the 2005 National Conference on Digital Government Research* (pp. 277-278), Atlanta, Greorgia.

Foot, K., Schneider, S., Dougherty, M., Xenos, M., & Larsen, E. (2003). Analyzing linking practices: Candidate sites in the 2002 U.S. electoral Web sphere. *Journal of Mediated Communication*, *8*(4).

Fukuda, H., Passos, E., Pacheco, A. M., Neto, L. B., Valerio, J., Roberto, V. J. D., et al. (2000). Web text mining using a hybrid system. In *Proceedings of the 6th Brazilian Symposium on Neural Networks* (pp.131–136).

Garofalakis, J., Kappos, P., & Mourloukos, D. (1999). Website optimization using page popularity. *IEEE Internet Computing*, *3*(4), 22–29. doi:10.1109/4236.780957

Getoor, L., & Diehl, C. P. (2005). Link mining: A survey. *ACM SIGKDD Explorations Newsletter*, *7*(2), 3–12. doi:10.1145/1117454.1117456

Getoor, L., Segal, E., Tasker, B., & Koller, D. (2001). Probabilistic models of text and link structure for hypertext classification. In *Proceedings of the IJCAI Workshop on Text Learning: Beyond Supervision*, Seattle, Washington.

Gibson, R. K., & Ward, S. J. (2000). A proposed methodology for studying the functions and effectiveness of party and candidate Web-sites. *Social Science Computer Review*, *18*(3), 301–319. doi:10.1177/089443930001800306

Giles, C. L., Bollacker, K., & Lawrence, S. (1998). CiteSeer: An automatic citation indexing system. In *Proceedings of the 3rd ACM Conference on Digital Libraries*, 89-98.

Gillani, B. (1998). The Web as a delivery mechanism to enhance instruction. *Educational Media International*, *35*(3), 197–202. doi:10.1080/0952398980350311

Gleim, R., Mehler, A., & Dehmer, M. (2006). Web corpus mining by instance of Wikipedia. In *Proceedings of the EACL 2006 Workshop on Web as Corpus*, Trento, Italy.

Goldenberg, A., & Moore, A. W. (2005). Bayes net graphs to understand co-authorship networks? In *Proceedings of the 3rd International Workshop on Link Discovery* (pp. 1-8). ACM Press.

Hayes, C., Avesani, P., & Veeramachaneni, S. (2006). An analysis of bloggers and topics for a blog recommender system. In *Proceedings of the Workshop on Web Mining, 7th European Conference on Machine Learning and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, Berlin, Germany.

Herrera-Viedma, E., & Pasi, G. (2006). Soft approaches to information retrieval and information access on the Web: An introduction to the special topic section. *Journal of the American Society for Information Science and Technology*, *57*(4), 511–514. doi:10.1002/asi.20305

Hess, A., & Kushmerick, N. (2004). Machine learning for annotating semantic Web services. In *Proceedings of the AAAI Spring Symposium on Semantic Web Services*, Palo Alto, California.

Hong, G. H., & Lee, J. H. (2005). Designing an intelligent Web information system of government based on Web mining. *Lecture Notes in Computer Science*, *3614*, 1071–1078.

Hu, W. & Meng, B. (2005). Design and implementation of Web mining system based on multi-agent. *Lecture notes on artificial intelligence* (Vol. 3584, pp.491-498).

Irani, Z., Al-Sebie, M., & Elliman, T. (2006). Transaction stage of e-government systems: Identification of its location & importance. In *Proceedings of the 39th Hawaii International Conference on System Sciences,* Hawai.

Kautz, H., Selman, B., & Shah, M. (1997). Referral Web: Combining social networks and collaborating filtering. *Communications of the ACM*, *40*(3), 63–65. doi:10.1145/245108.245123

Kohonen, T. (1990). The self-organizing maps. *Proceedings of the IEEE*, *78*, 1464–1480. doi:10.1109/5.58325

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *ACM*, *2*(1), 1–15.

Kuo, R. J., Liao, J. L., & Tu, C. (2005). Integration of ART2 neural network and genetic k-means algorithm for analyzing Web browsing paths in electronic commerce. *Decision Support Systems*, *40*, 355–374. doi:10.1016/j.dss.2004.04.010

Lappas, G., & Yannas, P. (2006). A framework to evaluate political party Websites. In *Proceedings of the 4th International Conference on Politics and Information Systems: Technologies and Applications Vol. II* (pp. 226-231), Orlando, Florida.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the Web. *Nature*, *400*, 107–109. doi:10.1038/21987

Lempel, R., & Moran, S. (2001). SALSA: The stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*, *19*(2), 131–160. doi:10.1145/382979.383041

Levene, M., & Loizou, G. (1999). *Computing the entropy of user navigation in the Web* (Tech. Rep. No. RN/99/42), University College London.

Li, X. (1998). Web page design and graphic use of three U.S. newspapers. *Journalism & Mass Communication Quarterly*, *75*(2), 353–365.

Liu, J. W., Yu, S. J., & Le, J. J. (2005). Online mining dynamic Web news patterns using machine learn methods. *Lecture notes on artificial intelligence* (Vo. 3614, pp. 462-465).

Lu, Q., & Getoor, L. (2003). Link-based text classification. In *Proceedings of the 3rd International Workshop on Link Discovery* (pp. 1-8). ACM Press.

Margolis, M., Resnick, D., & Tu, C.-C. (1997). Campaigning on the Internet: Parties and candidates on the World Wide Web in the 1996 primary season. *The Harvard International Journal of Press/Politics*, *2*(1), 59–78. doi:10.1177/1081180X97002001006

Martin-Guerrero, J. D., Palomares, A., Balaguer-Ballester, E., Soria-Olivas, E., Gomez-Sanchis, J., & Soriano-Asensi, A. (2006). Studying the feasibility of a recommender in a citizen Web portal based on user modeling and clustering algorithms. *Expert Systems with Applications*, *30*, 299–312. doi:10.1016/j.eswa.2005.07.025

Matsuo, Y., Ohsawa, Y., & Ishizuka, M. (2001). Average-clicks: A new measure of distance on the WWW. In *Proceedings of First Asia-Pacific Conference, Web Intelligence*, Japan.

Maule, R. W. (1998). Content design frameworks for Internet studies curricula and research. *Internet Research: Electronic Networking Applications and Policy*, *8*(2), 174–184. doi:10.1108/10662249810211647

Mayer, M. A., Karkaletsis, V., Stamatakis, K., Leis, A., Villarroel, D., & Thomeczek, C. (2006). MedIQ-Quality labelling of medical Web content using multilingual information extraction. *Studies in Health Technology and Informatics*, *121*, 183–190.

Melville, P., Mooney, R. J., & Nagarajan, R. (2002). Content-boosted collaborative filtering for improved recommendations. In *Proceedings of the 18th National Conference on Artificial Intelligence* (pp. 187-192).

Michalski, R. S., & Tecuci, G. (1994). *Machine learning: A multistrategy approach* (Vol. IV). Morgan Kaufmann

Mitchell, T. (1997). *Machine learning*. McGraw Hill.

Mladenic, D., & Grobelnik, M. (1999). Predicting content from hyperlinks. In *Proceedings of the 16th International ICML99 Workshop on Machine Learning in Text Data Analysis* (pp. 109-113).

Mobasher, B., Cooley, R., & Srivastava, J. (1999). Creating adaptive Web sites through usage based clustering of URLs. In *Proceedings of the IEEE Knowledge and Data Engineering Exchange Workshop* (KDEX99), Chicago, Illinois.

Mobasher, B., Dai, H., Luo, T., Sung, Y., & Zhu, J. (2000). Integrating Web usage and content mining for more effective Web personalization. In *Proceedings of the International Conference on E-Commerce and Web Technologies (ECWeb 2000)* (pp. 165-176). Greenwich, UK.

Mobasher, B., Jain, N., Han, E., & Srivastava, J. (1996). *Web Minning: Pattern discovery from WWW transaction* (Tech. Rep. TR-96050). Department of Computer Science, University of Minnesota, Minneapolis. Retrieved April 12, 2008, from http://citeseer.ist.psu.edu/mobasher96web.html

Mooney, R. J., & Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the 5th ACM Conference on Digital Libraries* (pp. 195-204). ACM Press.

Nasraoui, O., & Pavuluri, M. (2004). Complete this puzzle : A connectionist approach to accurate Web recommendations based on a committee of predictors. In *Proceedings of the 6th WEBKDD Workshop*, Seattle, Washington.

Ngu, D. S. W. & Wu, X. (1997). Sitehelper: A localized agent that helps incremental exploration of the World Wide Web. *Computer Networks*, *29*(8-13), 1249-1255.

Oberle, D., Berendt, B., Hotho, A., & Gonzalez, J. (2003). Conceptual user tracking. *Lecture notes on artificial intelligence* (Vol. 2663, pp. 155-164).

Pal, S., Talwar, V., & Mitra, P. (2002). Web mining in soft computing framework: Relevance, state of the art and future directions. *IEEE Transactions on Neural Networks*, *13*(5), 1163–1177. doi:10.1109/TNN.2002.1031947

Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the Web. *Connections*, *25*(1), 49–61.

Pazzani, M., & Billsus, D. (1997). Learning and revising user profiles: The identification of interesting Web sites. *Machine Learning*, *27*(3), 313–331. doi:10.1023/A:1007369909943

Pierrakos, D., Paliouras, G., Papatheodorou, C., Karkaletsis, V., & Dikaiakos, M. (2003). Web community directories: A new approach to Web personalization. *Lecture notes on artificial intelligence* (Vol. 3209, pp. 113-129).

Pilato, G., Vitabile, S., Vassallo, G., Conti, V., & Sorbello, F. (2003). A concurrent neural classifier for HTML documents retrieval. *Lecture Notes in Computer Science*, *2859*, 210–217.

Reeves, T. C., & Dehoney, J. (1998). Cognitive and social functions of courseWeb sites. In H. Maurer & R.G. Olson (Eds.), *Proceedings of WebNet World Conference 98—World Conference of the WWW, Internet & Intranet*. Orlando, FL: Association for the Advancement of Computing in Education.

Rennie, J., & McCallum, A. K. (1999). Using reinforcement learning to spider the Web efficiently. In *Proceedings of the 16th International ICML99 Workshop on Machine Learning in Text Data Analysis* (pp. 335-343).

Resig, J., Dawara, S., Homan, C. M., & Teredesai, A. (2004). Extracting social networks from instant messaging populations. In *Proceedings of LinkKDD'04*, Seattle, Washington.

Richardson, M., & Domingos, P. (2002). The intelligent surfer: Probabilistic combination of link and content information in PageRank. *Advances in Neural Information Processing Systems*, 14.

Riedl, R. (2003). Design principles for E-government services. In *Proceedings of eGov Day 2003*, Vienna, Austria.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). Analysis of recommendation algorithms for e-commerce. In *Proceedings of the ACM Conference on Electronic Commerce* (pp. 158-162).

Scheffer, T. (2004). Email answering assistance by semi-supervised text classification. *Intelligent Data Analysis*, *8*(5), 2004.

Schneider, S., & Foot, K. (2004). The Web as an object of study. *New Media & Society*, *6*(1), 114–122. doi:10.1177/1461444804039912

Scime, A. (2005). *Web mining: Application and techniques*. Hershey, PA: Idea Group Inc.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. doi:10.1145/505282.505283

Semeraro, G., Basile, P., Degemmis, M., & Lops, P. (2006). Discovering user profiles from papers by using word sense disambiguation. In *Proceedings of the ECML/PKDD Workshop on Web Mining* (pp. 69-79), Berlin, Germany.

Sharma, A., & Woodward, R. (2001). Political economy Websites: A researcher's guide. *New Political Economy*, *6*(1), 119–130. doi:10.1080/13563460020027795

Spiliopoulou, M., & Pohle, C. (2001). Data mining for measuring and improving the success of Web sites. *Data Mining and Knowledge Discovery*, *5*(1-2), 85–114. doi:10.1023/A:1009800113571

Spiliopoulou, M., Pohle, C., & Faulstich, L. (1999). Improving the effectiveness of a Web site with Web usage mining. In [San Diego, CA.]. *Proceedings of*, *WEBKDD99*, 142–162.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web usage mining: Discovery and applications of usage patterns from Web data. *SIGKDD Explorations*, *1*, 12–23. doi:10.1145/846183.846188

Stamatakis, K., Karkaletsis, V., Paliouras, G., Horlock, J., Grover, C., Curran, J. R., & Dingare, S. (2003). Domain-specific Web site identification: The CROSSMARC focused Web crawler. In *Proceedings of the Second International Workshop on Web Document Analysis (WDA 2003)* (pp. 75-78), Edinburgh, UK.

Sun, A., Lim, E. P., & Ng, W. K. (2002). Web classification using support vector machine. In *Proceedings of the Fourth ACM CIKM International Workshop on Web Information and Data Management (WIDM'02)*, McLean, Virginia.

Sutcliffe, A. (2001). Heuristic evaluation of Website attractiveness and Web usability. *Lecture Notes in Computer Science*, *2220*, 183–198. doi:10.1007/3-540-45522-1_11

Tango-Lowy, R., & Lewis, L. (2005). Situation management in crisis scenarios based on self-organizing neural mapping technology. In *Proceedings of the IEEE Military Communications Conference* (pp. 1-7), Atlantic City, New Jersey.

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology*, *57*(1), 60–68. doi:10.1002/asi.20253

Tug, E., Sakiroglu, M., & Arslan, A. (2006). Automatic discovery of the sequential accesses from Web log data files via a genetic algorithm. *Knowledge-Based Systems*, *19*(3), 180–186. doi:10.1016/j.knosys.2005.10.008

Wang, X., Abraham, A., & Smith, K. (2005). Intelligent Web traffic mining and analysis. *Journal of Network and Computer Applications*, *28*, 147–165. doi:10.1016/j.jnca.2004.01.006

Wang, Y., & Hu, J. (2002). A machine learning approach for table detection on the Web. In *Proceedings of the 11th International World Web Conference*, Honolulu, Hawaii.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.

Wolfgang, G., & Lars, S. (2000). Mining Web navigation path fragments. In *Proceedings of the Workshop on Web Mining for E-Commerce (KDD2000)* (pp. 105-110). Boston, MA.

Wu, H., Gordon, M., DeMaagd, K., & Fan, W. (2006). Mining Web navigations for intelligence. *Decision Support Systems*, *41*, 574–591. doi:10.1016/j.dss.2004.06.011

Yannas, P., & Lappas, G. (2005). Web campaign in the 2002 Greek municipal elections. *Journal of Political Marketing*, *4*(1), 33–50. doi:10.1300/J199v04n01_03

Yannas, P., & Lappas, G. (2006). Web candidates in the 2002 Greek prefecture elections. *Journal of E-Government*, *3*(1), 53–67. doi:10.1300/J399v03n01_04

Yao, Y. Y., Hamilton, H. J., & Wang, X. (2002). PagePrompter: An intelligent agent for Web navigation created using data mining techniques. *Lecture Notes in Computer Science*, *2475*, 506–513. doi:10.1007/3-540-45813-1_67

Yu, H., Han, J., & Chang, K. C. (2002). PEBL: Positive example based learning for Web page classification using SVM. In *Proceedings Of The International Conference On Knowledge Discovery In Databases (KDD02)* (pp. 239-248), New York.

Zaiane, O. R. (2001). Web usage mining for a better Web-based learning environment. In *Proceedings of Conference on Advanced Technology for Education* (pp. 60-64). Banff, Alberta, Canada.

Zhang, Y., Yu, J. X., & Hou, J. (2005). *Web communities: Analysis and construction*. Berlin: Springer.

Zhou, Z., Jiang, K., & Li, M. (2005). Multi-instance learning based Web mining. *Applied Intelligence*, *22*(2), 135–147. doi:10.1007/s10489-005-5602-z

Zukerman, I., & Albrecht, D. (2001). Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction*, *11*, 5–18. doi:10.1023/A:1011175525451

This work was previously published in Data Mining Applications for Empowering Knowledge Societies, edited by H. Rahman, pp. 76-95, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.18
Toward Mobile Web 2.0-Based Business Methods:
Collaborative QoS-Information Sharing for Mobile Service Users

Katarzyna Wac
University of Geneva, Switzerland & University of Twente, The Netherlands

Richard Bults
University of Twente, The Netherlands & Mobihealth B.V., The Netherlands

Bert-Jan van Beijnum
University of Twente, The Netherlands

Hong Chen
Altran Netherlands B.V., The Netherlands

Dimitri Konstantas
University of Geneva, Switzerland

ABSTRACT

Mobile service providers (MoSPs) emerge, driven by the ubiquitous availability of mobile devices and wireless communication infrastructures. MoSPs' customers satisfaction and consequently their revenues, largely depend on the quality of service (QoS) provided by wireless network providers (WNPs) available at a particular location-time to support a mobile service delivery. This chapter presents a novel method for the MoSP's QoS-assurance business process. The method incorporates a location- and time-based QoS-predictions' service, facilitating the WNP's selection. The authors explore different business cases for the service deployment. Particularly, they introduce and analyze business viability of QoSIS.net, an enterprise that can provide the QoS-predictions service to MoSPs, Mobile Network Operators (as MoSPs), or directly to their customers (i.e. in B2B/B2C settings). QoSIS.net provides its service based on collaborative-sharing of QoS-information by its users. The authors argue that this service can improve the MoSP's QoS-assurance process and consequently may increase its revenues, while creating revenues for QoSIS.net.

DOI: 10.4018/978-1-60566-366-1.ch006

INTRODUCTION

The last 15 years have been marked by the expansion, global adoption and seamless availability of the (fixed) Internet with a multitude of its ubiquitous services. In parallel, a new mobile era has undergone its preparation phase, driven by miniaturization and personalization of communication devices, as well as the rapid expansion and adoption of mobile voice and data services and heterogeneous communication infrastructures (Hansmann *et al.*, 2003).

In this era, ubiquitous Mobile Service Providers (*MoSPs)* bring to the market users' favorite Internet-services and start offering a wide range of new services. As a *mobile service* we define a data service that is delivered to (or from) a mobile device from (or to) the Internet. The service delivery is supported by the deployment of wireless communication infrastructures, enabling a user the Internet-connectivity while on the move. A MoSP is then an enterprise, which core business processes aim at providing mobile services to its customers (TMF, 2001).

The MoSPs are fully aware that to achieve their goals of gaining customer acceptance, secure their revenues and remain competitive, their services must provide users *Quality of Experience* (QoE) comparable to the QoE provided by the existing Internet-services (Afuah & Tucci, 2000). The QoE is "the overall acceptability of service, as perceived *subjectively* by the user" (ITU-T, 2007), and as a part of it, a MoSP must at least assure meeting user's *Quality of Service* (QoS) requirements, expressed quantitatively in terms of service speed, accuracy, dependability, security level and price related (performance) measures. The QoS is defined as "a collective effect of service performances which determine the (*objective*) degree of satisfaction of a user" (ITU-T, 1993). MoSPs unanimously indicate, that their *QoS-assurance process*, i.e., a business process related to service-management, and responsible for assurance that services provided to users are performing according to their QoS-requirement, is critical to their business viability (Andersson *et al.*, 2006; Nokia, 2004).

The challenge is that, the MoSPs' QoS-assurance process relies on the QoS provided by wireless communication infrastructures, supporting the mobile services delivery. Ideally, there would be infrastructures supporting anytime-anywhere services delivery to a user. However, in reality these infrastructures are beyond the control of MoSPs; they are owned and managed by enterprises called *Wireless Network Providers* (*WNPs*). A WNP core business processes aim at providing Internet-connectivity to its customers. A wireless communication infrastructure exhibits different QoS depending on the WNP it's owned by, a wireless technology used (e.g. WLAN/GPRS) as well as the mobile user's location-time.

Nowadays, in most countries, a number of WNPs coexist, operating different long-range wireless communication technologies. In particular, there exists at least one nationwide WNP called a *Mobile Network Operator (MNO)*. A MNO is an enterprise that, as a WNP, owns and manages its wireless network infrastructures. MNO's core business processes aim at providing voice services (primarily) and Internet-connectivity, using long-range wireless technologies (e.g. GSM/UMTS) (ITU, 2005) to its customers. A MNO can take a role of a MoSP, i.e., can also provide mobile services to its customers.

In parallel, WNPs like public-WLAN providers emerge rapidly, especially in big cities. Moreover, today's mobile devices embed multiple interfaces, thus supporting Internet-connectivity over various wireless technologies. Hence, the communications means tend to be ubiquitously available to mobile service users, and, at least in principle, these users must be able to choose a WNP (and a wireless technology) providing the QoS best meeting their QoS requirements thus facilitating meeting their expected-QoE. However, this assumption is far from the reality. The business strategy of existing WNPs, and particularly

MNOs, is based on a user 'lock-in', i.e., the user can access only wireless networks owned by 'his' WNP (Buschken, 2004). Moreover, the only disclosed information about the QoS provided by WNPs is based on the WNPs' marketing data about their network's theoretical performance—the real QoS (i.e. objectively measured) is unknown to the mobile users. Even mobile services provided by MNOs (taking a role of MoSPs), are based on estimations regarding their provided-QoS. A likely cause is the limited time/budget spent on service-user trials by MNOs (Tan, 2004). Moreover, it is widely accepted that performance tests conduced by WNPs, and particularly by MNOs, are coarse-grained 'drive-tests' over main streets in big cities and main highways (Gomez & Sanchez, 2005). These tests prove the availability of MNOs networks, i.e. their coverage at different locations; and the results are disclosed to a public as coverage maps. Current practices continue, despite the fact that the network's provided-QoS and user perceived-QoE are identified as one of the major factors for customers' loyalty, and moreover, a 'lock-in' practice is admitted to fail as a long-term strategy for customers' retention (Gerpott *et al.*, 2001; M.I.S Trend S.A., 2007; Nokia, 2004).

Effectively, WNPs networks provide 'best-effort' service quality-level to MoSPs. Additionally, MoSPs cannot select at a certain location-time a WNP, which provides a QoS that best meets their customer's required-QoS and their expected-QoE. MoSPs are constrained by the above situation and can only provide mobile services to their customers at a 'best-effort' QoS-level.

We envision that the above mobile user's restrictions on a WNP selection will be removed in the future. Already we see a role of *Mobile Virtual Network Operator (MVNO)* appearing on the market. A MVNO is a WNP (and particularly a MNO) providing voice services and Internet-connectivity over wireless network infrastructures owned/managed by multiple WNPs/MNOs (i.e. MVNO has business contracts with these WNPs/MNOs to resell a use of their infrastructures). A MVNO's customer, i.e., a mobile user can therefore have an Internet-connectivity via set of network infrastructures. With larger number of WNPs/MNOs, with which a MVNO has a contract, a larger WNP/MNO choice becomes available to a mobile user at any location-time. However, this does not solve the problem of how the user will choose the most appropriate WNP/MNO, given his location-time, QoS-requirements and QoE-expectations.

As a solution, we firstly propose a novel *method* facilitating QoS-assurance business process of MoSPs. As a novel method we define a (business) method employed in an enterprise's business process(es), innovative in terms of technology and/or business contracts established between enterprises in a service-value chain (TMF, 2004).

Our method is technologically innovative because it incorporates fine-grained location-time-based predictions about the QoS-observed when given WNP is used, into the MoSP's QoS-assurance process. These predictions are denoted as *QoS-predictions*. They can be used to select a WNP (and a wireless technology) best meeting MoSP's user QoS-requirements. The goal of the proposed method is to facilitate MoSPs' to meet user's expected-QoE and thus increase own revenues. This follows the observation that a happy MoSP user is willing to continue using the service and paying for it (Andersson et al., 2006).

Our method is also innovative for a MoSP in terms of its business contracts. Namely we envision that there is a separate enterprise—a provider of the QoS-predictions (service), called *QoSIS.net*. QoSIS stands from *Quality-of-Service Information Service* embracing the QoS-predictions as part of its service. A MoSP has business relation with QoSIS.net in order to use the QoS-predictions.

Besides a novel method for a MoSP, we also propose a novel method for QoSIS.net. This method facilitates its core business process, i.e. a QoS-predictions service-fulfillment related process, embracing development and operation of the service, as requested by its customers (TMF, 2001).

The method proposed for QoIS.net is innovative technologically because it is based on user collaborative QoS-information sharing; QoS-information is acquired when mobile service users use different WNPs at different locations-times. The method's key feature is that QoSIS.net can provide accurate QoS-predictions only after having a large volume of historical QoS-information acquired from mobile users. Currently, we conduct the QoS-predictions feasibility assessment (including accuracy evaluation) in the mobile healthcare domain (Wac *et al.*, 2008).

To be more specific, QoSIS.net provides mobile service users (as a community) with QoS-predictions, as a value-added service. On behalf of these users, QoSIS.net exchanges between them the QoS-information about QoS observed when using different WNPs. This follows an idea behind all emerging *Web 2.0* services that gain an increasing users' acceptance in different domains (e.g. Facebook, Wikipedia) (Hoegg *et al.*, 2006; Martignoni & Stanoevska-Slabeva, 2007; Pascu *et al.*, 2005). Web 2.0 is a paradigm for "mutually maximizing collective knowledge and added-value for each *participant* (in a community) by formalized and dynamic sharing and creation of user-generated content" (O'Reilly, 2005). Its 'mobile' extension implies an implementation of Web 2.0 service as a mobile service. QoSIS.net, as a provider of QoS-predictions, based on QoS-information acquired from mobile users, adheres to a Mobile Web 2.0 paradigm. QoSIS.net customers are such 'participants' (also called 'prosumers' (Tapscott & Williams, 2006)) contributing to the community with their QoS-information.

In this chapter, the business viability of QoSIS.net is examined in detail, and it discusses possible technical solutions supporting QoS-predictions service fulfillment. We argue that the methods proposed for MoSPs and QoSIS.net have a strong potential to bring benefit to them in terms of revenue creation and customer satisfaction. Therefore, in the next sections we present current solutions for QoS-assurance process of MoSPs, then we present our approach for and the results of- QoSIS.net business viability study, after which we present our vision for MoSPs' QoS-assurance process.

Current Trends

Internet was since its beginning, providing a 'best-effort' QoS-level. In the late 90's, first solutions for Internet-based SPs QoS-assurance business processes have been proposed, particularly for providers of real-time multimedia services (Shepherd *et al.*, 1996) (Hutchison *et al.*, 1997). From the technical perspective, these solutions were based on use of rigorous and complex QoS-management frameworks, including functions like QoS-negotiation and resource reservation (Andersen *et al.*, 2000; Xiao & Ni, 1999). Moreover, from the business perspective, these solutions required business contracts between SPs and Internet-providers (Afuah & Tucci, 2000). They contradicted with the 'open' nature of Internet-services, because they limited the SPs customer-base to that of the Internet-provider's base. In this situation, many SPs ignored the proposed solutions and learned to manage 'best-effort' QoS-level provided by Internet to assure meeting the QoS-requirements (and the expected-QoE) of their service users. The methods, employed in SP's QoS-assurance process, relied on estimations of QoS-provided by the Internet. This approach was feasible due to at least two factors: firstly, QoS-provided by Internet exhibits regularities and long-term estimations for e.g. months can be derived relatively accurately (claffy *et al.*, 1998). Secondly, if necessary, SPs could easily acquire information on provided-QoS via dedicated QoS-monitoring and all that without any degradation in the quality of their provided services (Michaut & Lepage, 2005).

With the dawn of the mobile era, history is repeating itself. The QoS-assurance process has been identified as critical to the business viability of MoSPs already in the late 90's (Chalmers & Sloman, 1999), when only basic voice and data services existed. The most critical factor in QoS-assurance is related to the user's mobility: a mobile user relies on the availability of different WNP/wireless technologies at different locations-times along his trajectory (Dekleva *et al.*, 2007). To deal with this, MoSPs are (again) advised to employ complex QoS-management frameworks in their QoS-assurance business processes. This solution (again) requires a close business relationship between MoSPs and WNPs, resulting in a WNP-centric business models. This solution is explored in research (Calvo *et al.*, 2004; Faber *et al.*, 2003; Robles *et al.*, 2002; Tan, 2004; Tsalgatidou & Pitoura, 2001).

From a technological perspective there are also other proposals. For example a proposal for MoSPs QoS-assurance process to employ predictions of user mobility path acquired from a WNP (Han & Venkatasubramanian, 2006; Soh & Kim, 2003), or to employ a QoS-broker for the reservation of WNP's network resources on behalf of MoSPs users (Nahrstedt *et al.*, 2001). Simultaneously, MNOs propose new concepts like Universal Mobile Access, Generic Access Network or IP-Multimedia-System (Cuevas *et al.*, 2006), aiming at a technical as well as a business solutions for MoSP's QoS-assurance process; aiming for MoSPs to have their core business processes tightly coupled with the business processes of MNOs. Such solutions contradict the 'mobile' nature of MoSPs' services, because business relations with MNOs limit MoSPs customer base and service-usage area to the MNOs customer-base and its coverage-area.

The MoSPs not following the WNP-centric business models emerge on a growing scale (Tan, 2004). Naming few, Skype is a VoIP provider (Osterwalder *et al.*, 2005), MobiHealth.com is a mobile healthcare SP (MobiHealth, 2007), while Digital Chocolate is a mobile gaming SP (Digital Chocolate, 2008). They all struggle to assure meeting their mobile users' QoS-requirements over unpredictable 'best-effort' QoS-provided by WNPs/wireless technologies (Bults *et al.*, 2005). Dedicated QoS-monitoring is not feasible due to the dynamic nature of QoS-provided by WNPs/wireless technologies (Dood, 2005) and the limited resources (e.g. battery) of mobile devices.

At the same time, new WNPs appear and new long-range wireless technologies (e.g. Ultra-Mobile-Wideband) are deployed by the existing WNPs. These are steps towards the vision of 4G, where wireless communication infrastructures are going to be plentifully available for mobile service users (De Vriendt *et al.*, 2002; Dekleva et al., 2007; Ortiz, 2007; Tachikawa, 2003). Yet, any new WNP or wireless technology when launched provides only a 'best-effort' QoS-level (Gomez & Sanchez, 2005) and uses the 'drive-tests' as performance tests.

A new trend of user-centric business models is arising in a mobile business research (Dekleva et al., 2007), aiming in the situation where a MoSP and its mobile service user can access any WNP anytime-anywhere. Towards this end, some technological solutions are proposed for signaling for inter-WNPs handover (Bless *et al.*, 2004; ITU-T, 2006) or for QoS-negotiations between a mobile device and a WNP (Manner *et al.*, 2001). From the business contractual perspective, an idea of 'ad-hoc' smart-business networks (van Heck & Vervest, 2007) between WNPs and MoSPs supporting delivery of service to a mobile user has been presented, while research in large EU projects focus on new methods for MNOs to support MoSPs, without locking-in the MoSPs' customers (Sanchez *et al.*, 2008). Still the question remains: how a MoSP chooses a WNP best matching his user's QoS-requirements and QoE-expectations at a given location-time. To the best of our knowledge, the kind of solution we propose has not yet been proposed in the literature.

Research Goals and Approach

The goal of this chapter is to analyze the business viability of the methods proposed for MoSPs and QoSIS.net. Our approach is as follows. The first step is the choice of the MCM-Business Model Framework of Hoegg et al. for analysis of QoSIS.net as an example of enterprise employing the Mobile Web 2.0 paradigm.

The MCM framework was initially based on a well-established definition of a *business model* being "1) an architecture for the products, services, information flows, including a description of various business actors and their roles, 2) a description of the potential benefits for the various business actors, and 3)a description of the sources of revenues" (Timmers, 1998). Furthermore, Hoegg et al. enhanced this definition to include a social environment component, and that based on examining the other existing frameworks being applied to the existing MoSPs (e.g. 'MediAlert' SP). Along this research they conducted surveys and interviews with mobile service users and domain experts. Finally, the framework presented by them in 2005 (Hoegg & Stanoevska-Slabeva, 2005), has been later-on successfully employed in analysis of existing businesses by other researchers, and in an analysis of (Mobile) Web 2.0 businesses conducted by the authors themselves (Hoegg et al., 2006; Martignoni & Stanoevska-Slabeva, 2007).

The second step of our approach is to apply this framework and conduct QoSIS.net business viability analysis. Particularly because the QoSIS.net is a non-existing enterprise, we conduct this analysis based on the intended QoS-predictions service usage scenarios. Therefore in the following paragraphs we provide the details of the framework, then the QoSIS.net usage scenario, and then its business viability analysis along the framework.

The MCM-Business Model Framework

The MCM-business model framework can be applied for a business viability of a MoSP, and particularly for an enterprise providing (Mobile) Web 2.0-services. The framework has the following components, that collectively expose enterprise's business model and factors influencing its design (Hoegg & Stanoevska-Slabeva, 2005), see also Figure 1:

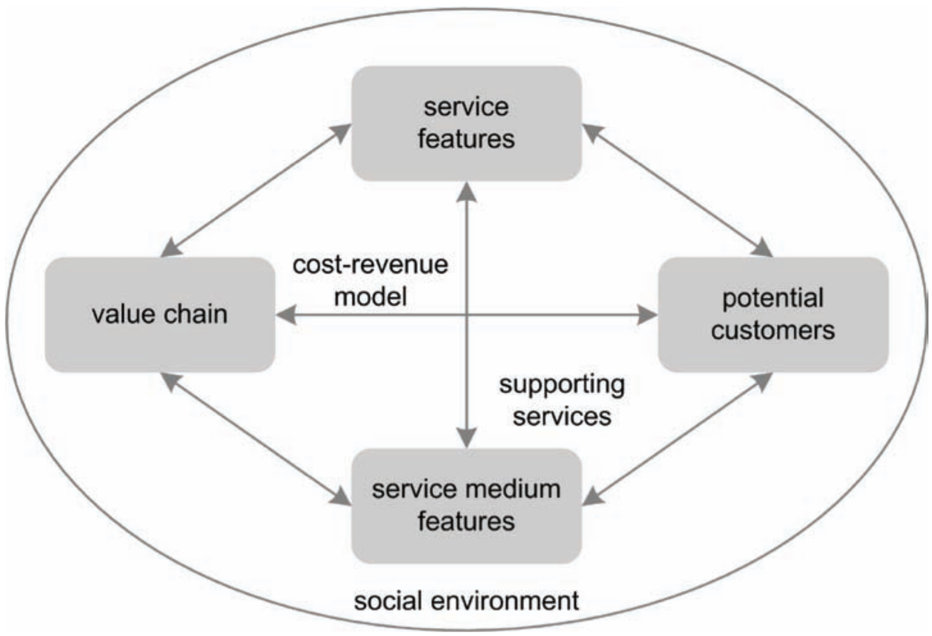
- **Features of the service:** Design and operation of the service provided by an enterprise to its customers (i.e. the enterprise product)
- **Potential customers:** Aspects of target groups of customers (i.e. market segments) and expected service's value-added for them
- **Value chain:** Players involved in delivery of the provided service and their interrelations
- **Cost-revenue model*:*** Financial model explaining contribution of players in value chain
- **Features of the service medium:** Characteristics of means with which service is delivered and that may influence service interactions (e.g. service can be mobile)
- **Flow of supporting services:** Necessary for delivery of the service provided to customers
- **Social environment:** External influences—social, ethical aspects influencing the way the business is designed, implemented and operated.

QoSIS.net Viability Analysis

User Scenario

As every Saturday, Sophie, Maria and Eric meet in Westfield-London shopping center. Sophie is a young COPD (chronic obstructive pulmonary disease) patient and her health state is continuously monitored from the hospital with the MobiHealth's COPD Body-Area-Network (BAN) that she is wearing. She does not have to visit a care professional at the hospital frequently, she feels secure

Figure 1. MCM-business model framework



being remotely monitored. She is less restricted in her active life; she enjoys Saturday's habit of meeting her friends. In case of exacerbations, help from the hospital is dispatched to her, wherever she is. In the mall, her COPD BAN always uses the most suitable WNP/wireless technology as available there. The "*most suitable*" means sending her vital-signs data at the highest speed, i.e. with the lowest possible delay, to the hospital, and in a secure manner (i.e. not using unknown free WLANs). Netwrok usage price does not matter. The WNP/wireless technology selection process is transparent to her. In the mall her BAN uses most of the time the 3G-HSUPA network of Vodafone-UK, and when it is not available, e.g. inside the big shops, the BAN switches to 3G-UMTS network of T-Mobile. In such a way her service-data is always sent at the highest possible speed to the hospital.

Maria and Eric are both typical technology nerds competing with each other. Maria has a newest mobile from 3UK with this fancy SeeMeTV application, which allows her to easily post short videos from her mobile on the SeeMeTV-web. She enjoys making jokes with Sophie and Eric, trying to get some catchy video that can bring her money if other SeeMeTV users watch it or even a prize of £1000 if they vote for it. To access the SeeMeTV-web her mobile always uses the most suitable wireless technology of 3UK, as available in this center. For her, the most suitable means that her videos appear online as quick as possible; price does not matter much to her; she can afford it. This time her mobile uses the 3G-UMTS network, because 3UK's GSM/GPRS network is, as always on Saturdays, saturated with voice/SMS traffic of users shopping in the same center.

Eric is a Facebook fan, and uses this social-networking community to stay in touch with his family and friends back home in Amsterdam. He also enjoys making funny photos and movies along their day and posts them to his blog. He has just bought a fancy t-shirt for his brother, and he has posted a related movie and is now waiting for response—maybe it is not what his brother wanted? To access Facebook, his mobile always chooses the most suitable WNP/wireless technology available in the mall (in his case the cheapest but still as fast as possible!). Because he uses the latest phone with an integrated GPS, he was able to download an application from QoSIS.net website. This little application makes the WNP/wireless technology choice on his behalf; he just needed to set his selection preferences. Since he installed it, he is very satisfied using his Facebook.com services which are provided to him without frustrating connection loses, as it was the case in the past. During this shopping-day, his mobile uses 2.5G-GPRS of O2 in the morning, and Orange-WLAN network in the afternoon, as they moved to the north part of this big center, where it has become available to him.

Features of the Service

The QoS-predictions service is a mobile service that provides MoSP and their users with fine-grained location-time-based QoS-predictions, about QoS-provided by different WNPs/wireless technologies at a given location-time. Each QoS-prediction is an estimation of QoS-provided by a WNP/wireless technology at a given location and time in the future, associated with its accuracy value.

The QoS-predictions service is a value-added service for MoSPs and their users. It is used by a MoSP in a WNP/wireless technology selection process; but it is transparent for a MoSP user like Sophie, Maria or Eric (the only effect a MoSP user may see is a changing WNP name on his/her device screen as an effect of a vertical handover of his mobile-device between different WNPs).

Although it was not shown in the QoSIS.net user scenario, the QoS-predictions service based on the user's private mobility-map (stored on the user's mobile device) denoting nearby locations along with QoS-predictions for these locations in different times. The mobility-map is automatically updated, either at regular time-intervals or at the moment a new QoS-prediction is requested. The information in the map allows the mobile service user to get QoS-predictions even if the user is out of coverage of any WNP (although these QoS-predictions can be outdated depending when the map was last updated); in this situation a user can see where is the closest WNP available for him and what is its predicted QoS. If a service provided by a MoSP is critical, e.g. a MoSP is MobiHealth providing mobile healthcare service, the user can be warned explicitly via sound/vibration generated by the mobile device, when he a leaves the coverage area of any WNP.

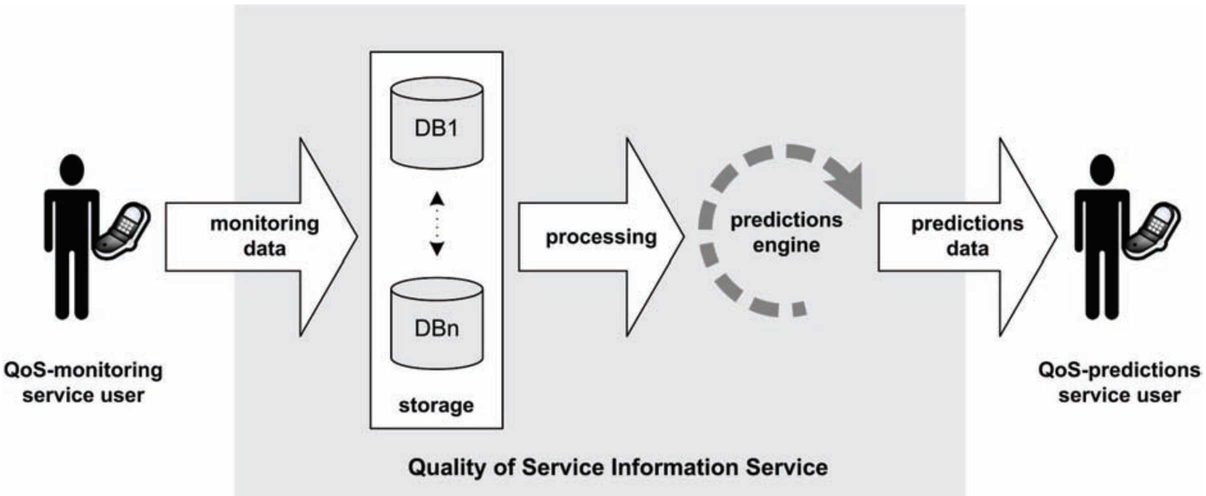
The QoSIS.net operation is based on the QoS-information service; QoS-predictions service is part of it. The QoS-predictions service is based on users' collaborative QoS-information sharing, i.e. information is acquired from MoSPs users using different WNPs/wireless technologies at different locations-times. The collected QoS-information is used to derive QoS-predictions, used (back) by mobile users. The QoS-predictions are then validated along the new collected QoS-information; a feedback is implemented implicitly. Therefore, the QoS-information service contains of (Figure 2):

1. QoS-monitoring service
2. QoS-information processing service and
3. QoS-predictions engine (Pawar *et al.*, 2008; Wac, 2006).

Particularly, QoS-monitoring and QoS-prediction services are executed partially on a mobile device of the service user (see further paragraphs for details).

The *QoS-monitoring service* acquires QoS-information from mobile users (and actually from their devices) based on application-level measurements, i.e., measurements of the mobile service's observed QoS as well as users location-time and other parameters (see further details). Technically, QoS-monitoring service can be realized via passive QoS-measurements in MobiHealth or SeeMeTV service for users like Sophie and Maria, or via measurement-applications like AcbTaskMan

Figure 2. QoSIS.net high-level system architecture



(AcbTaskMan, 2007) or CoSphere (Peddemors, 2008) for Facebook users like Eric.

The QoS-monitoring service is transparent to the MoSP's users. The QoS-information is acquired to QoSIS.net, preferably via a lightweight protocol. The information is send with an hourly or daily frequency or even continuously (however that can be exhausting for mobile device resources). The information can be compressed (using lossless algorithms) before being send to the QoSIS.net, where is stored in databases (DBs).

The quality of the acquired QoS-information is critical to the accuracy of the QoS-predictions' service (and hence to QoSIS.net business viability). We argue, that the QoS-predictions' service cannot be more accurate than the historical QoS-information it is based upon. Particularly the location-time accuracy is critical; a user moving few meters away may observe a different QoS for a given WNP/wireless technology and hence acquire different QoS-monitoring data (Dood, 2005).The QoS-information acquired from MoSPs users includes:

1. the mobile user's location-time, mobility level (e.g. fixed location, walking, moving in a vehicle)
2. the WNP name and wireless technology used
3. the network usage price and security level
4. the mobile service's speed in terms of delay/throughput/loss (ITU-T, 1993, 2006; Seitz, 2003); detailed metrics depend upon the mobile service for which QoS-predictions are provided
5. the characteristics of the mobile device (CPU, memory, battery, network interfaces) and the location-determination system type/accuracy e.g. GPS system

The QoS-information does not have (nor needs) any notion of the mobile user's identity, all information acquired from mobile users is anonymous (but needs to be trustworthy).

The *QoS-processing service* (presented as an arrow in the figure) transforms the stored QoS-information into a form suitable for a QoS-predictions' engine. The transformation may include annotation, rating, sub-sampling, discretization, obfuscation or any other method for information processing commonly used in data mining techniques (Mitchell, 1999).

The *QoS-prediction engine* continuously processes large quantities of collected historical QoS-information. The engine uses data mining techniques (e.g. Bayesian Networks or trees (Witten & Frank, 2005)) to discover QoS-information patterns. A QoS-prediction service request retrieves instantaneously the requested QoS-predictions for a given location-time. The result is returned to the QoSIS.net customer (i.e. a MoSP service user) in form of mobility-maps.

One of the mobile service users' characteristics, which we envision can facilitate QoS-predictions accuracy, is the fact that these users tend to have fixed mobility patterns, e.g. home, work, shopping-center at particular week-days on particular hours. Scientific research indicates that 45% of the people stay in a location area of a radius of ~10 km, while 73%—of radius of ~30 km and their mobility patterns can be learned for a history of 3 months (Gonzalez *et al.*, 2008). We anticipate that if the user's mobility pattern is highly predictable, a collection of large amount of QoS-information for WNPs/wireless technologies available to this a user along his movements, facilitates providing accurate QoS-predictions to this user.

Features of the Service vs. Mobile Web 2.0 Paradigm

The QoS-predictions' service is based on users-collaborative QoS-information creation and sharing. Hence, QoSIS.net can be seen as a specific example of *stand-alone Mobile Web 2.0 Service Provider* (Hoegg et al., 2006; O'Reilly, 2005). The content produced by MoSPs users is the QoS-information acquired by the QoS-monitoring service. This content is then processed in QoS-predictions engine and it is consumed by mobile users via the QoS-predictions service. Mobile users are content 'prosumers' (Tapscott & Williams, 2006). We argue that while implementing the QoS-predictions service as Mobile Web 2.0, where users collaboratively collect QoS-information, QoSIS.net may benefit from the fact that MoSPs' users are very likely to be in a close location-time span, e.g. in one city, and therefore they collect overlapping QoS-information, which in turn increases the QoS-predictions service accuracy.

Potential Customers

The QoSIS.net market contains two types of customers:

1. customers in a Business-to-Business market segment, containing MoSPs and WNPs/MNOs/MVNOs acting as MoSPs and
2. customers in a Business-to-Customer market segment containing mobile service end-users (i.e. customers of MoSPs).

Recalling that in our scenario, Sophie is a MobiHealth.com customer and a user of its mobile health monitoring service. MobiHealth.com is an example MoSP (van Halteren *et al.*, 2004). Maria is a MNO-3UK customer and SeeMeTV mobile service user; her MNO acts as a MoSP (3UK, 2008). Sophie and Maria have role of QoSIS.net service users. Eric is a Facebook.com (MoSP (Facebook, 2007)) customer and user, and a customer and user of QoSIS.net.

Business-to-Business (B2B)

MoSPs like MobiHealth.com or MNOs like 3UK are potential business customers of QoSIS.net. They are particular mobile service providers in e.g. mobile information, healthcare or entertainment domains. Services provided by these types of MoSP require frequent exchange of service-data between (Internet-based) servers and mobile users and hence frequent use of WNPs. Mobile user's satisfaction depends on the QoS-provided by the WNP used for service delivery. For MoSP, QoS-predictions can be used in a WNP/wireless technology selection process. For a MNO acting as a MoSP, QoS-predictions can be used in a selection process of a wireless technology (WLAN/GPRS/UMTS/HSxPA) at a given location-time for a mobile service user. The goal of MoSPs is to best meet users' QoS-requirements and facilitate meeting their expected-QoE. QoSIS.net can play an instrumental role in meeting this goal.

The QoSIS.net fulfills the MoSP's need of knowledge of QoS-provided by different WNPs/wireless technologies at mobile user location-time. For a MNO, it adds to its existing coarse-grained knowledge of its network; a MNO can analyze its provided-QoS to mobile service users at given locations-time and be willing to improve this QoS in order to stay competitive amongst the WNPs. The QoS-information can support the MNO's network planning, dimensioning and management, as well as better design of own mobile services provided to mobile users. Moreover, once encouraged, a WNP can even provide QoSIS.net with some additional information regarding its network configuration, which would facilitate more accurate QoS-predictions service. Without the existence of the QoSIS.net, MoSPs/MNOs try to assure QoS to its users by using coarse-level QoS-estimations (c.f. Current Trends). The advantage for QoSIS.net of entering in a business relation with a MoSP would be that they can be an equal copartner in designing the QoS-predictions and QoS-monitoring services (e.g. mobile device configuration), have integrated billing for both services, co-marketing and co-branding strategies, bringing even more customers to the MoSP.

When using QoS-predictions, a MoSP better manages user's resources involved in its services delivery e.g. lower service-data delays/improved throughput or optimizing user's device battery life. The biggest incentive for a MoSP to use QoSIS.net would be the fact that by using QoSIS.net, it can increase its customers' satisfaction and hence its revenues.

Although it sounds contradictory to what we have said earlier, we claim that MoSPs and particularly MNOs can use QoSIS.net services as a way to additionally 'lock-in' its mobile service users to their services/networks. It is because a customer satisfaction and QoE is a strong predictor of customer loyalty (Eshghi *et al.*, 2007), and QoSIS.net facilitates its improvement. Namely, if users are satisfied from the services provided to them, and they experience them at the expected-QoE level, they are more likely to use even more services from this MoSP. Additionally, the cost of switching to other MoSP for this user can be high: new MoSP may not use QoSIS.net, and even if they do, the user's historical QoS-information might not be portable to this new MoSP. In this situation a user need to accept that only after some service-launch-time (necessary for collecting of historical QoS-information), the service will be provided to him at the expected-QoE level, bringing him back satisfaction related to the service use.

Business-to-Customer (B2C)

Mobile service users like Eric are potential customers of QoSIS.net, but only if they have a mobile device with localization capabilities, e.g. an integrated GPS. Because their QoE while using their mobile services is related to the QoS-provided by the WNP/wireless technology used, they are interested in the use of QoSIS.net's value-added service in order to facilitate (proactive and automatic) choice of the WNP/wireless technology that fits best to their needs anytime-anywhere. The user grants responsibility to the QoSIS.net on how to use QoS-predictions, i.e., when to demand new QoS-predictions and when to choose another WNP/wireless technology.

QoSIS.net can equally target mobile users living a village or city suburbs with limited WNP/wireless technologies choice, as well as big number of users living in the city center with plentiful of available WNPs/wireless technologies.

A QoSIS.net user saves resources involved in e.g. Facebook.com services delivery e.g. money while choosing a WNP with cheaper tariffs, or mobile device's battery life. The incentive for a mobile user to use QoSIS.net is the improvement of his QoE when using his mobile services. Without the existence of QoSIS.net, Eric would be using an arbitrary WNP/wireless technology as available in his location-time, probably experiencing lower quality services and paying more than he expects.

QoSIS.net's Critical Success Factors

We identify that a critical success factor for the QoSIS.net service is to reach a critical mass of QoS-predictions service users. This is because QoSIS.net business thrives on large quantities of QoS-information for its QoS-predictions service. We envision that the primary market segment is B2B, and the secondary is the B2C. The B2B market entry strategy is to convince MoSPs and MNOs that QoSIS.net adds an accurate fine-grained location-time-based QoS-predictions service to their infrastructure that facilitates the improvement of QoE of their customers, hence increases their own revenues. This way the QoSIS.net user-base will be then equal to MoSPs/MNOs customer base. However, the initial hurdle of getting enough historical QoS-information for accurate QoS-predictions remains. QoSIS.net needs to agree with its customers on service-launch-time duration, after which it will provide QoS-predictions service with a given accuracy. We envision that the service-launch-time will be a function of mobility patterns of mobile service users (c.f. discussion on QoS-predictions engine) (Gonzalez et al., 2008). Once the QoSIS.net B2B market is substantial enough (i.e. sufficient QoS-information is available for QoS-predictions), the B2C market can be targeted.

Value-Chain

Value chain of QoSIS.net is different for B2B and B2C market segments and is presented in the following paragraphs.

B2B

In the value chain for business relation between QoSIS.net and a MoSP we distinguish QoSIS.net, MoSP like MobiHealth.com, and a MoSP user—Sophie (Figure 3a), who is also a QoSIS.net user. MobiHealth.com can have a business relation with one or more different WNPs/MNOs (via a MVNO). For a relation between QoSIS.net and a MNO acting as a MoSP, we distinguish: QoSIS.net, MNO like 3UK (three.co.uk) and its SeeMeTV service user—Maria (Figure 3b), who is also a QoSIS.net user. In both cases QoSIS.net is a 3rd Party SP, i.e., value-added SP for MobiHealth.com and three.co.uk (QoSIS.net customers). They provide QoSIS.net services to their customers as an integrated offer; these services are transparent for Sophie and Maria.

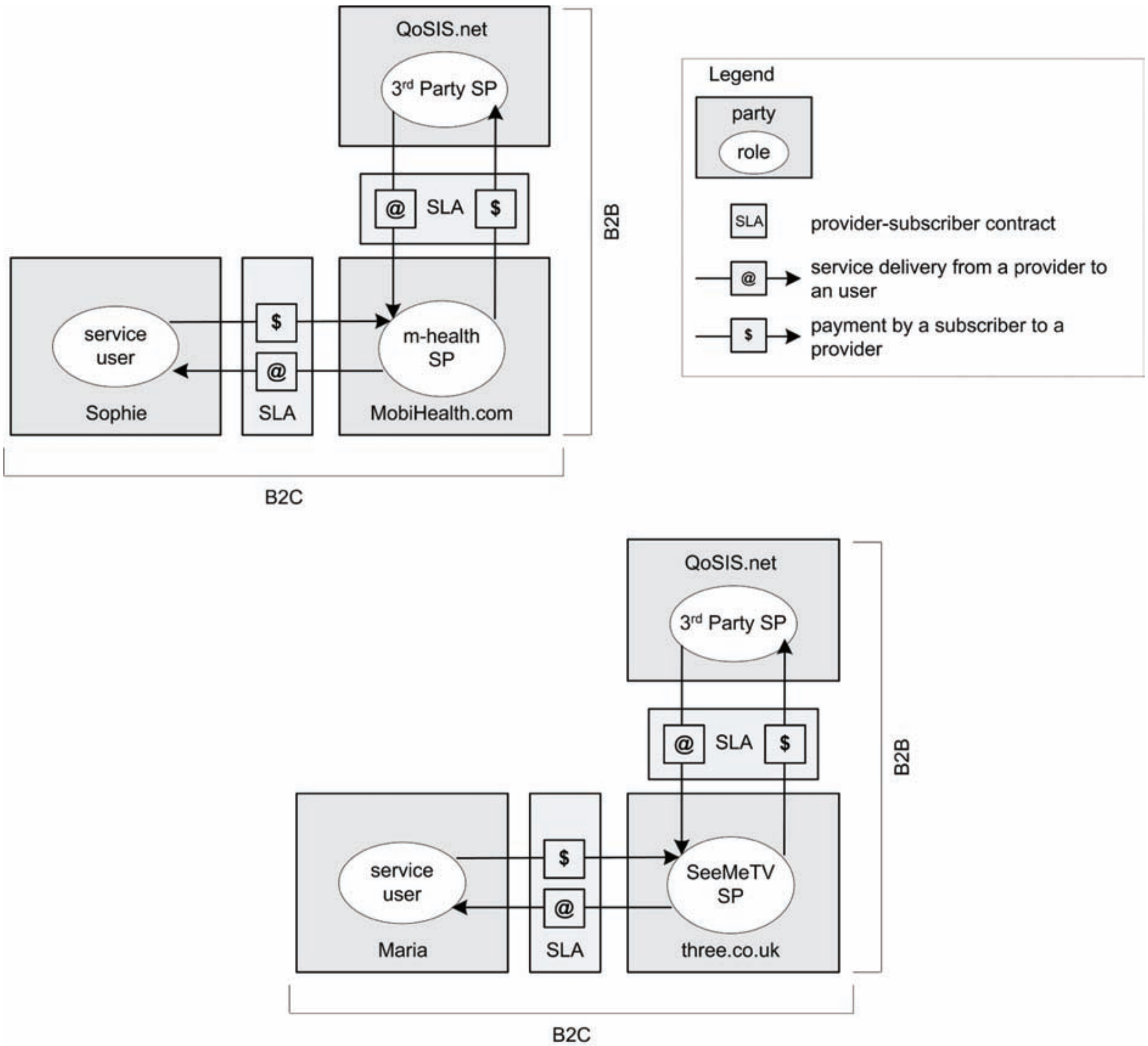
The *business relation* between QoSIS.net and MobiHealth.com/3UK is defined in a *contract* relationship. It is a negotiated and agreed between them formal document defining the terms and conditions for the delivery of the services (e.g. service availability of 99%), detailed services' usage specifications (e.g. user-pull or service-push) as well as the payment specifications (e.g. monthly, per-transaction post-paid) by MobiHealth.com and three.co.uk to QoSIS.net. this contract is also called a Service Level Agreement (SLA) (TMF, 2001).

Due to the nature of the service provided by QoSIS.net, a strong partnership trust (Ratnasingam & Phan, 2003) as well as technology trust (Ratnasingam *et al.*, 2002) is required between MobiHealth.com, 3UK and QoSIS.net in order to assure the success of all businesses. Critical technological details of QoS-monitoring and QoS-predictions service delivery need to be included in SLA, maybe even in legal terms, e.g. which enterprise is responsible if QoS-predictions are not accurate enough and Sophie finds herself outside any WNP at the moment she requires urgent medical assistance?

B2C

In the value chain for business relation between QoSIS.net and a MoSP user, we distinguish: QoSIS.net, Facebook.com and Facebook.com's user—Eric (Figure 4), who is also a QoSIS.net user. Eric can have a business relation with one or a more WNPs/MNOs (via a MVNO).

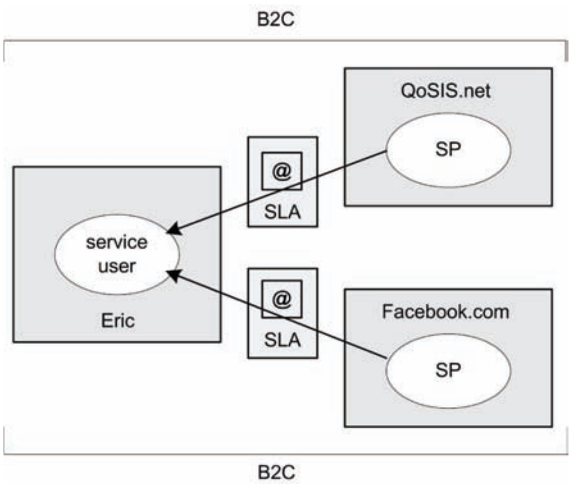
Figure 3. (a) Top QoSIS.net as 3rd party service provider for MobiHealth.com (a MoSP) 3b.on bottom QoSIS.net as 3rd party service provider for 3UK (a MNO in a role of a MoSP)



QoSIS.net is a value-added SP for Eric, and their business relation is defined by a SLA. The payment relationship with Eric states that QoSIS.net services are free (as his Facebook.com services). It is important to notice that the terms and conditions specified in this SLA need to be agreed upon by Eric, before the QoSIS.net application can be used by him, i.e. installed on his mobile device. Moreover, Eric is responsible to make sure that by using QoSIS.net he does not violate an existing SLA between him and Facebook.com, especially with regards to the (default) WNP use or use of any mobile device resources, for which these two SPs would now compete.

Facebook.com does not need to be in a business relationship with QoSIS.net, but it may benefit (in terms of revenues) from the services provided by it; satisfied Eric is more likely to use even more of Facebook.com services while on the move.

Figure 4. QoSIS.net as a service provider



Cost-Revenue Model

In B2B settings, QoSIS.net has costs related to the setup and maintenance of its services, and to the marketing aiming to get new MoSPs/WNPs as customers. QoSIS.net's customers have costs related to QoS-predictions service usage (as agreed in SLA). We argue that any reasonable pricing model proposed by QoSIS.net can be beneficial for QoSIS.net's revenues, increasing also the MoSPs/WNPs revenues, depending nevertheless on the MoSP's application area and criticality on QoSIS.net service to the MoSP's core business.

QoSIS.net customers in B2C settings do not pay for a service usage. In these settings, the QoSIS.net marketing costs are small, as the major advertisement policy is a "word-of-mouth" going from one mobile service user to another (Tapscott & Williams, 2006). QoSIS.net can also launch an affiliation program, rewarding mobile users who helped to acquire a new QoSIS.net customer via their website.Any QoSIS.net service user has costs:

1. Related to ownership of a mobile device with one or multiple network interfaces for wireless technologies like WLAN/GPRS/UMTS/HSxPA to be used via different WNPs (or at least one WNP) and an integrated location-determination e.g. GPS module
2. Related to the mobile device's resources: computation, storage and communication capacity, as well as its battery used for the service delivery.

We would like to emphasize that particularly in Europe devices manufacturers are committed to increase the number of mobile devices with integrated location-determination technologies, like GSM-cell triangulation or GPS. Nokia committed itself that by 2012 50% of its phones will have GPS-integrated (Nokia, 2004). In the US having these technologies in a mobile-phone is mandated due to 911 number and the location-identification requirements, while in Asia it is widely accepted as a new fancy technology. Market analysts predict that overall, by 2011, 29.6% of all mobile phones will have GPS (comparing to only 11.1% for 2006) (ABI Research, 2007).

Once QoSIS.net has reached a critical mass of customers, it can reconsider its cost-revenue model. For example, possible revenue for QoSIS.net can be generated by selling its QoS-information acquired in B2C-settings to MoSPs (like Facebook.com) or MNOs, or any other interested enterprises. It can mean selling (anonymized) user profiles and information about used mobile services.

QoSIS.net can also differentiate its prices for B2B and for B2C (introducing 'premium' service). The transaction fee can depend on:

1. The number of WNPs/wireless technologies available to a mobile user at a given location-time (i.e. the price increases with the number of WNPs/wireless technologies, because the richer the choice, the higher probability that a MoSP can use WNP/wireless technology matching its user QoS-requirements, thus improving its user's-QoE),
2. The actual accuracy of the QoS-predictions, where this accuracy can be checked against

the QoS-information acquired from user QoS-monitoring data
3. The accuracy of QoS-predictions, which would be (on purpose) lower for lower accuracy QoS-predictions, and higher for higher-accuracy QoS-predictions.

Service Medium

The QoSIS.net service is a mobile service—provided to its users from server on the Internet, accessible via a wireless medium. The features of this medium depend on WNPs/wireless technologies available at a particular location-time and used for QoS-predictions delivery. The QoSIS.net service may intentionally or accidentally be disabled/disturbed (Camponovo & Pigneur, 2003; Tsalgatidou & Pitoura, 2001), due to:

1. User communication-autonomy: (a) the user can deliberately configure his mobile device to use one particular WNP and wireless technology (b) user can be unreachable due to WNP lack-of-coverage or device's empty battery
2. Vulnerability of the user's mobile device (damage/loss) and its limited storage, processing, communication capabilities
3. WNP's wireless technology characteristics, e.g. asymmetrical throughput characteristics, variable delay characteristics and restrictions on volume of QoS-information exchange (hence lightweight protocols and lossless compression of exchanged QoS-information can be required).

Supporting Services

From a business perspective, all QoSIS.net's services: QoS-monitoring, QoS-processing and QoS-predictions services are necessary to create a value to the QoSIS.net's service user. Supporting processes include processes for a WNP/wireless technology selection (i.e. handover to) and its enforcement. A technical realization of intra- and inter-WNP (i.e. vertical) handovers is an ongoing research issue (Chen & Shu, 2005; Dekleva et al., 2007; Pawar et al., 2008), outside of the scope of this chapter.

From the QoSIS.net perspective, its web-service provided to customers like Eric is a service supporting acquisition of new customers in the B2C settings. Moreover, due to the nature of services delivered by QoSIS.net, business-partnership management is one of the supporting services but still in the core activities of QoSIS.net, e.g. for generating service reports for its customers.

Social Environment

There are several influences of the QoSIS.net services provision rising from customer competition, legislation and social/ethical constraints. Firstly, competition amongst QoSIS.net customers (MoSPs/WNPs) requires QoSIS.net to be a trustworthy enterprise. It should apply strong security mechanisms to prevent competitive customer information being disclosed to others. Any information regarding MoSP service usage statistics or its customer's profiles should be protected. Similarly, WNPs' competitive market situation poses strong security requirements on the QoSIS.net information. Any information regarding QoS provided by a WNP should not be altered in favor of this WNP. Moreover, in order to secure QoSIS.net revenues and its competitive advantage on the market, details of its QoS-information databases or QoS-predictions engine should not be disclosed to its customers.

An important social aspect of the provided service is related to the user-privacy consent. QoS-information acquired from QoSIS.net customers contains detailed location-time information of mobile service users. Therefore, it is required (at least in Europe) that a mobile service user is legally informed of the fact that this privacy sensitive information is acquired (Gorlach *et al.*, 2004), even if it is in anonymous form. In the B2B case, it is the responsibility of a MoSP as the QoSIS.net customer to provide user-privacy informed-consent and then to exchange all QoS-information respecting security, authentication/authorization mechanisms. In the B2C case, all this is the responsibility of QoSIS.net.

QoSIS.net-based Method's Future

There are many possible scenarios, in which QoSIS.net as a part of the proposed novel business method for a MoSP QoS-assurance can evolve. For example, regarding the cost-revenue model, QoSIS.net can provide QoS-monitoring and QoS-predictions services as separate services, and reward a user of QoS-monitoring service (i.e. producing QoS-information) while charging a user for using QoS-predictions service (i.e. consuming QoS-information). This however brings to QoSIS.net a risk of not having enough contributing users in B2C case; statistics for Web 2.0 indicate that at the launch of any service, only 1% of users is willing to contribute and generate the content (Arthur, 2006).

The other possible future scenario in the B2C settings is related to the situation where a user would like to use QoSIS.net, however, he is not a frequent MoSP's service user (e.g. like Eric who is using his Facebook.com daily) and hence will not generate lots of 'real' mobile traffic. This kind of mobile user can be offered to use a QoSIS.net QoS-monitoring service as his mobile phone 'screen-saver'. Namely, when he would not use his mobile for a while, the QoS-monitoring service would take a role of an active mobile service user, using some WNP/wireless technology and imitating e.g. busty web-browsing, thus acquiring QoS-monitoring information at a given location-time. In such a way, QoSIS.net would acquire QoS-information enriching its databases and such a mobile user could be paid for information generation. However, the critical issue would be related to mobile device resources usage for this QoS-monitoring service (battery, capacity, storage etc.).

A scalable solution for QoSIS.net would be to limit its scope of operation to a particular city, region or country, limiting the scope of WNPs/wireless technologies for which QoS-predictions can be provided. Scoping can be dictated by the need of limiting the QoS-information to be processed by the QoS-prediction engine, or the need for higher accuracy of QoS-predictions provided for a restricted location-area. Furthermore, QoSIS.net can be limited in terms of WNPs/wireless technologies for which it acquires QoS-information. In the case of different QoSIS.net's location-based instances, they can form (short-term or a long-term) "smart-business-networks" for the purpose of delivering better services to their users roaming in locations areas belonging to different instances of QoSIS.net. All QoSIS.net instances collaborating in the business network could benefit from a larger QoS-information base for their QoS-predictions service.

Moreover, in order to enhance the QoS-information base and improve the QoS-predictions accuracy, we envision that QoSIS.net customers—MoSPs/WNPs whose users are likely to be in overlapping location-areas, can have a business relationship in which they agree to collaboratively share their QoS-information bases. All collaborating MoSPs could benefit from larger QoS-information base for QoS-predictions.

CONCLUSION

Mobile and ubiquitous service providers (MoSPs) emerge, struggling to provide to their users a QoE that is at least comparable to one the user is familiar with from the Internet. To bridge the gap regarding the lack of information about QoS-provided by different wireless network providers (WNPs) over diverse wireless technologies in a mobile users' location-time, in this chapter we have proposed methods enabling firstly a creation of an enterprise (QoSIS.net) providing such an information to MoSPs, and, secondly, usage of this information by an MoSP in its QoS-assurance business process. We emphasize that the aim of the proposed methods is to facilitate fulfillment of MoSP's user's QoS-requirements and his expected-QoE, hence securing the revenues of a MoSP, while creating revenues to QoSIS.net.

Future research opportunities within the domain of our topic relate to further advancements of the proposed methods for QoSIS.net and its customers, e.g. understanding the dependencies between the QoS-provided by a WNP/wireless technology and the MoSP user's QoS-requirements and his QoE-expectations, and this for MoSPs in different application domains. Related to this, a second research opportunity can focus on the efficient market entry approach for QoSIS.net as enterprise, such that it overcomes initial hurdle of attracting the required critical mass of users and start generating revenues. A third research opportunity lies in understanding the partnership-trust required in the QoSIS.net's value chain, as well as the challenges in QoSIS.net's customer management in the B2B and B2C market segments. Finally, we indicate a need for research on trust in technology. This research investigates dependability features of architectural QoSIS-system design and technical details included in SLAs established between enterprises (QoSIS.net, WNPs/MNOs), e.g. security of QoS-information exchange between enterprises.

The future research opportunities relate particularly to research on new competitive methods that can be employed in existing management, operational (i.e. core) or supporting business processes of mobile and ubiquitous service providers as enterprises. These methods need necessarily aim in satisfying their customer, while increasing enterprise revenues. We propose these methods to be based on emerging trend of short- and long-terms business inter-dependencies (i.e. "smart-business-networks") between different enterprises, bringing into a value network different, but complementary expertise. This has a high risk of failure but also a huge potential to increase revenues of all of the involved enterprises and that by increasing a customer's QoE anywhere-anytime-anyhow.

Moreover we propose research on novel methods employing the user-collaborative information sharing paradigm, i.e. Mobile Web 2.0. The risk of these methods is mainly related to attracting a critical mass of contributing users. These methods would aim at creating new revenue streams from user-generated content-manipulation and enrichment. The enriched content could then be become a part of an enterprise service, consumed back by the users. The Mobile Web 2.0-based methods however require careful research upon the content type to be generated-consumed by the users. Therefore, answers for research questions like: what is the pre-existing offline information possessed by users, which, if available to be manipulated and shared online amongst them could empower them in some way?, Does this information violates in any way user's privacy?, And what is user's willingness to share this information online and with whom?, are critical.

We envision that in order to fulfill the dream of providing mobile services meeting users' QoE-expectations and generating revenues, MoSPs need to take a necessary risk and employ novel methods in their existing business processes.

REFERENCES

AcbTaskMan. (2007). AcbTaskMan software.

Afuah, A., & Tucci, C. T. (2000). *Internet business models and strategies: Text and cases*. McGraw-Hill Higher Education.

Andersen, D., Bansal, D., Curtis, D., Seshan, S., & Balakrishnan, H. (2000). *System support for bandwidth management and content adaptation in internet applications.* Paper presented at the 4th Symposium on Operating Systems Design and Implementation (OSDI), San Diego, CA, USA.

Andersson, C., Freeman, D., James, I., Johnston, A., & Ljung, S. (2006). *Mobile media and applications, from concept to cash: Successful service creation and launch*. West Sussex, England: Wiley.

Arthur, C. (2006). What is the 1% rule?

Bless, R., Hillebrand, J., Prehofer, C., & Zitterbart, M. (2004). Quality-of-service signaling for next-generation ip-based mobile networks. *IEEE Communications Magazine*, *42*(6), 72–79. doi:10.1109/MCOM.2004.1304235

Bults, R., Wac, K., van Halteren, A., Konstantas, D., & Nicola, V. (2005). *Goodput Analysis of 3G wireless networks supporting m-health services.* Paper presented at the 8th International Conference on Telecommunications (ConTEL05), Zagreb, Croatia.

Buschken, J. (2004). *Higher Profits Through Customer Lock-In*: Thomson Texere.

Calvo, M., Rodríguez, C., & Dillinger, M. (2004). *Business models for reconfigurable communication systems.* Paper presented at the 13th IST Mobile & Wireless Communications Summit, Lyon, France.

Camponovo, G., & Pigneur, Y. (2003). *Business model analysis applied to mobile business.* Paper presented at the 5th Intl Conference on Enterprise Information Systems (ICEIS03), Angers, FR.

Chalmers, D., & Sloman, M. (1999). A survey of Quality of Service in mobile computing environments. *IEEE Communications Surveys and Tutorials*, *2*(2).

Chen, W. T., & Shu, Y. Y. (2005). *Active application oriented vertical handoff in next-generation wireless networks.* Paper presented at the Wireless Communications and Networking Conference (WCNC05).

claffy, k., Miller, G., & Thompson, K. (1998). *The nature of the beast: recent traffic measurements from an Internet backbone.* Paper presented at the International Networking Conference (INET98), Geneva, Switzerland.

Cuevas, A., Moreno, J. I., Vidales, P., & Einsiedler, H. (2006). The IMS Platform: A Solution for Next Generation Network Operators to Be More Than Bit Pipes. *IEEE Commun. Mag. Advances in Service Platform Technologies*, *44*(8), 75–81.

De Vriendt, J., Laine, P., Lerouge, C., & Xu, X. (2002). Mobile Network Evolution: A Revolution on the Move. *IEEE Communications Magazine*, *40*(4), 104–111. doi:10.1109/35.995858

Dekleva, S., Shim, J. P., Varshney, U., & Knoerzer, G. (2007). Evolution and emerging issues in mobile wireless networks. *Communications of the ACM*, *50*(6), 38–43. doi:10.1145/1247001.1247003

Digital Chocolate. (2008). Seize the Minute.

Dood, A. (2005). *The Essential Guide to Telecommunications*. US: Prentice Hall PTR.

Eshghi, A., Haughton, D., & Topi, H. (2007). Determinants of customer loyalty in the wireless telecommunications industry. *Telecommunications Policy*, *31*(2), 93. doi:10.1016/j.telpol.2006.12.005

Faber, E., Ballon, P., Bouwman, H., Haaker, T., Rietkerk, O., & Steen, M. (2003). *Designing business models for mobile ICT services.* Paper presented at the 16th BLED Electronic Commerce Conf.—eTransformations, Bled, Slovenia.

Facebook. (2007). A social utility that connects you with the people around you.

Gerpott, T. J., Rams, W., & Schindler, A. (2001). Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. *Telecommunications Policy*, *25*(4), 249. doi:10.1016/S0308-5961(00)00097-5

Gomez, G., & Sanchez, R. (2005). *End-to-End Quality of Service over Cellular Networks: Data Services Performance Optimization in 2G/3G*: John Wiley & Sons, Ltd.

Gonzalez, M., Hidalgo, C. A., & Barabasi, A. (2008). Understanding individual human mobility patterns. *Nature*, *453*(1), 779–782. doi:10.1038/nature06958

Gorlach, A., Heinemann, A., & Terpstra, W. (2004). *Survey on location privacy in pervasive computing.* Paper presented at the Workshop on Security and Privacy in Pervasive Computing (SPCC04) at PERVASIVE2004.

Han, Q., & Venkatasubramanian, N. (2006). Information Collection Services for QoS-aware Mobile Applications. *IEEE Transactions on Mobile Computing*, *5*(5), 518–535. doi:10.1109/TMC.2006.65

Hansmann, U., Merk, L., Nicklous, M., & Stober, T. (2003). *Pervasive Computing: The Mobile World*: Springer.

Hoegg, R., Martignoni, R., Meckel, M., & Stanoevska-Slabeva, K. (2006). *Overview of business models for Web 2.0. communities.* Paper presented at the GeNeMe, Dresden, DE.

Hoegg, R., & Stanoevska-Slabeva, K. (2005). *Towards Guidelines for the Design of Mobile Services.* Paper presented at the 18th BLED Electronic Commerce Conf., Bled, Slovenia.

Hutchison, D., Mauthe, A., & Yeadon, N. (1997). Quality-of-service architecture: Monitoring and control of multimedia communications. *Electronics & Communication Engineering Journal*, *9*(3), 100. doi:10.1049/ecej:19970303

ITU. (2005). *The Internet of Things (Statistical Annex)*.

ITU-T. (1993). General aspects of Quality of Service and Network Performance in Digital Networks, including ISDNs (Vol. I.350): ITU.

ITU-T. (2006). Framework for achieving end-to-end IP performance objectives (Vol. []: ITU.]. *Y (Dayton, Ohio)*, 1542.

ITU-T. (2007). Vocabulary for performance and quality of service: Appendix I—Definition of Quality of Experience (QoE) (Vol. P.10/G.100): ITU-T.

Manner, J., Burness, L., Hepworth, E., Lopez, A., & Mitjana, E. (2001). *Provision of QoS in heterogeneous wireless IP access networks.* Paper presented at the Intl Symposium on Personal, Indoor and Mobile Radio Communications.

Martignoni, R., & Stanoevska-Slabeva, K. (2007). *Mobile Web 2.0.* Paper presented at the 20th BLED Electronic Commerce Conf.—eMergence, Bled, Slovenia.

Michaut, F., & Lepage, F. (2005). Application-oriented network metrology: Metrics and active measurement tools. *IEEE Communications Surveys & Tutorials*, *7*(2), 2–24. doi:10.1109/COMST.2005.1610543

M.I.S Trend S.A. (2007). *Study of use of mobile services in telecommunication (realized for Division of Telecommunication of the Federal Communication Office in Switzerland).* Lausanne, CH: Institute for market and opinion surveys study.

Mitchell, T. (1999). Machine Learning and Data Mining. *Communications of the ACM*, *42*(11), 30–36. doi:10.1145/319382.319388

MobiHealth. (2007). Putting care in motion.

Nahrstedt, K., Xu, D., Wichadakul, D., & Li, B. (2001). QoS-aware middleware for ubiquitous and heterogeneous environments. *IEEE Communications Magazine*, *39*(11), 140–148. doi:10.1109/35.965372

Nokia. (2004). *Quality of Experience (QoE) of mobile services: Can it be measured and improved?*

O'Reilly, T. (2005). What is Web 2.0?

Ortiz, S. (2007). 4G Wireless Begins to Take Shape. *IEEE Computer*, *40*(11), 18–21.

Osterwalder, A., Ondrus, J., & Pigneur, Y. (2005). *Skype's Disruptive Potential in the Telecom Market: A Systematic Comparison of Business Models*. Lausanne, CH: University of Lausanne.

Pascu, C., Osimo, D., Ulbrich, M., Turlea, G., & Burgelman, J. C. (2005). The potential disruptive impact of internet 2 based technologies. *First Monday—peer-reviewed Journal on the Internet*.

Pawar, P., Wac, K., van Beijnum, B. J., Maret, P., van Halteren, A., & Hermens, H. (2008). *Context-Aware Middleware Architecture for Vertical Handover Support to Multi-homed Nomadic Mobile Services*. Paper presented at the 23rd Annual ACM Symposium on Applied Computing (ACMSAC08), Ceará, Brazil.

Peddemors, A. (2008). CoSPhere NAL software.

Ratnasingam, P., Pavlou, P., & Tan, Y. (2002). *The Importance of Technology Trust for B2B Electronic Commerce*. Paper presented at the 15th BLED Electronic Commerce Conf.—eReality: Constructing the eEconomy, Bled, Slovenia.

Ratnasingam, P., & Phan, D. (2003). Trading Partner Trust in B2B E-Commerce: A Case Study. *Information Systems Management*, *20*(3), 39–50. doi:10.1201/1078/43205.20.3.20030601/43072.5

Research, A. B. I. (2007). *GPS-enabled mobile devices—Key drivers and latest trends pushing GPS penetration in CDMA, GSM and WCDMA Handsets* (White Paper).

Robles, T., Mitjana, E., & Ruiz, P. (2002). *Usage scenarios and business opportunities for systems beyond 3G*. Paper presented at the IST Mobile and Wireless Telecommunications Summit 2002, Thessaloniki, GR.

Sanchez, A., Carro, B., & Wesner, S. (2008). Telco Services for End Customers: European Perspective. *IEEE Communications Magazine*, *46*(2), 14–18. doi:10.1109/MCOM.2008.4473072

Seitz, N. (2003). ITU-T QoS standards for IP-based networks. *IEEE Communications Magazine*, *41*(6), 82–89. doi:10.1109/MCOM.2003.1204752

Shepherd, D., Scott, A., & Rodden, T. (1996). Quality-of-service support for multimedia applications. *IEEE MultiMedia*, *3*(3), 78–82. doi:10.1109/93.556542

Soh, W. S., & Kim, H. S. (2003). QoS provisioning in cellular networks based on mobility prediction techniques. *IEEE Communications Magazine*, *41*(1), 86–92. doi:10.1109/MCOM.2003.1166661

Tachikawa, K. (2003). A perspective on the evolution of mobile communications. *IEEE Communications Magazine*, *41*(10), 66–73. doi:10.1109/MCOM.2003.1235597

Tan, S. (2004). *Evolution of mobile technology and business models (technical report)*: Center for Information and Communication Technologies, Lyngby, DK.

Tapscott, D., & Williams, A. D. (2006). *Wikinomics: How mass collaboration changes everything*. New York, USA: Portfolio.

Timmers, P. (1998). Business models for electronic markets. *Journal on Electronic Markets*, *8*(2), 3–8. doi:10.1080/10196789800000016

TMF. (2001). SLA Management Handbook (Vol. GB917-2001): TMF.

TMF. (2004). *Shared Information/Data (SID) Model; GB922 Addendum 0—SID Primer (ver.4.0)*. Morristown, New Jersey, US: TeleManagement Forum.

Tsalgatidou, A., & Pitoura, E. (2001). Business models and transactions in mobile electronic commerce: requirements and properties. *Computer Networks*, *37*(2), 221–236. doi:10.1016/S1389-1286(01)00216-X3UK. (2008).

van Halteren, A., Bults, R., Wac, K., Konstantas, D., Widya, I., & Dokovsky, N. (2004). Mobile Patient Monitoring: The MobiHealth System. *The Journal on Information Technology in Healthcare*, *2*(5), 365–373.

van Heck, E., & Vervest, P. (2007). Smart business networks: how the network wins. *Communications of the ACM*, *50*(6), 28–37. doi:10.1145/1247001.1247002

Wac, K. (2006). *QoS-predictions service: infrastructural support for proactive QoS- and context-aware mobile services.* Paper presented at the On the Move to Meaningful Internet Systems 2006: OTM Workshops, Intl Workshop on Context-Aware Mobile Systems (CAMS), Monpellier, France.

Wac, K., Hilario, M., Konstantas, D., & van Beijnum, B. J. (2008). Data mining on application-level QoS traces: The MobiHealth system case study (forthcoming). *IEEE Trans. on Mobile Computing*.

Witten, I., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Xiao, X., & Ni, L. M. (1999). Internet QoS: a big picture. *IEEE Network*, *13*(2), 8–18. doi:10.1109/65.768484

This work was previously published in Mobile and Ubiquitous Commerce: Advanced E-Business Methods, edited by J. Wang, pp. 80-100, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.19
The Pedagogical Implications of Web 2.0

Matthias Sturm
ICT Consultant, Canada

Trudy Kennell
ICT Consultant, Canada

Rob McBride
ICT Consultant, Canada

Mike Kelly
ICT Consultant, Canada

ABSTRACT

Web 2.0 tools like blogs, Wikis, and podcasts are new to the vocabulary of language acquisition. Teachers and students who take full advantage of these emerging tools will participate in more dynamic, immediate, and communicative environments that provide opportunities for meaningful experiences through social constructivist learning. This chapter aims to bring perspectives rooted in educational theory to a domain too often dominated by the technological implications of its tools and argues that social constructivism is the pedagogical paradigm for learning and teaching facilitated by the next generation of Web technology. It reviews basic theoretical tenets and discusses their implications. Social constructivism lays the foundation for learning environments that foster the participation of students and teachers in today's knowledge and information-based society to their full potential.

DOI: 10.4018/978-1-60566-190-2.ch020

INTRODUCTION

Language learning in the 21st century has new tools at hand. Tools like blogs, wikis, and podcasts are new to the vocabulary of language acquisition. Language learning environments are evolving into more dynamic, immediate, and communicative environments. The traffic on second language Web sites like *BBC Learning English* demonstrates the growing popularity and reach of online language learning. Concurrent with the development of the Internet over the past 20 years, learning has become intertwined with learning online; more and more people are looking for flexibility and independence in their language learning experiences.

The emergence of new technologies has always been accompanied by promises of the transformation of learning and teaching. In *Teacher and Machines*, Cuban (1996, p. 3) states that “educators [have] searched for means of communicating knowledge in simple, inexpensive, and timely ways” while “making instruction both productive and enriching” (p. 3), all in the name of transforming education to serve more students more efficiently. Cuban continues to say that “because teachers believe that interpersonal relations are essential in student learning, the use of technologies that displace, interrupt, or minimize that relationship is viewed in a negative light” (1996, p. 61). Although Cuban has argued that technologies have been oversold, he also makes a case for computers being underused in modern education. Apart from the promise of “more efficient and productive” educational institutions, the “transformation of teaching and learning into an engaging and active process connected to real life,” and “the preparation of the current generation of young people for the future workplace” have been major goals of educational technology reform (Cuban, 2002, pp. 13-15).

The promise of Web 2.0 technologies is different. Their impact on the learning process and the practice of teaching is truly revolutionary in that it does not promise more efficiency but it extends the relations between teachers and students beyond the two-dimensional models of instruction to multidimensional networks that resemble the world we live in closer than ever before. However, the role of technology represents a site of struggle with effects on the quality of learning opportunities. As Warschauer (2006) argues, “Educational reformers suggest that the advent of new technologies will radically transform what people learn, how they learn, and where they learn, yet studies of diverse learners’ use of new media cast doubt on the speed and extent of change” (p. 1).

BACKGROUND

Many educators consider correspondence education the precursor of distance education. Correspondence education developed in the mid-19th century and this was the only way to reach students who were physically separated from their instructor. By the mid-20th century, education models had evolved to computer systems built to also increase the efficiency of instruction by delivering learning packages to a large number of students, for example via PLATO (Programmed Logic for Automatic Teaching Operations) (Berners-Lee & Caillau, 2000, p. 85). In the late 1960s, a computer-assisted instructional system called TICCET (Time-shared, Interactive, Computer-Controlled Educational Television) was developed by combining computer with television technology to deliver large amounts of individually controlled instructional material to students. It was not until the 1980s that progress in the areas of speech recognition, machine-assisted translation, Artificial Intelligence and generally Natural Language Processing was made to a significant extent. While computers became more available to the average consumer and the World Wide Web was invented they didn’t enter the public sphere until the early 1990s. From this the first generation of the Web as an environment for learning emerged, giving teachers the tools to create and disseminate electronic and digitized learning materials in more efficient ways much like correspondence courses once did with print-based resources. The learning paradigm as such remained unchanged.

Many influences have had an impact on the world of learning and on shaping the online learning landscape of the second generation of the World Wide Web, the Web 2.0 era. Richardson’s (2006) book, *Blogs, Wikis, Podcasts, and other Powerful Web Tools for Classrooms* has been recognized for the use of blogs in the classroom as a “trendsetter in education” by the *New York Times*. The underlying concepts of the use of Web tools in this book are spreading throughout the online education community. Richardson states that "every educator needs to understand the potential impact of these tools [and the implications of] the social connections that students are now making on the Web, the ability to share and contribute ideas and work, the new expectation of collaboration, and the ability to truly extend the walls of our classrooms" (2006, p. viii). Throughout the book, Richardson describes these free and easy to use Web-based services to create blogs, wikis, RSS feeds, podcasts, and social bookmarks as learning tools for a constructivist and collaborative pedagogy with considerable relevance to curriculum standards.

A hugely popular blog created by the Public Library of Charlotte & Mecklenburg County called *Learning 2.0* lists *23 Things* or small exercises developed for their staff to explore and expand their knowledge of the social Web. The blog was developed to provide professional development on Web 2.0 tools using Web 2.0 tools and has been copied, commented on, contributed to by many people, much in the same way that Richardson describes the potential of these tools for education.

While there is still some disagreement about just what Web 2.0 means, O'Reilly, a supporter of the Open Source Movement and credited with coining the term Web 2.0, states that "Web 2.0 is the network as platform, spanning all connected devices … creating network effects through an 'architecture of participation,' and going beyond the page metaphor of Web 1.0 to deliver rich user experiences" (O'Reilly, 2005, n.p.). Downes, a senior researcher at Canada's National Research Council, who has been regularly writing about Web 2.0 and its impact on education for some time, echoes O'Reilly's definition of Web 2.0 but underlines that most importantly, Web 2.0 is "a social phenomenon embracing an approach to generating and distributing Web content itself, characterized by open communication, decentralization of authority, freedom to share and re-use, and the market as a conversation" (Wikipedia, 2007, n.p.).

Web 2.0 tools have now become so widely available that many students are already using them (Prensky, 2001, 2004). What's more, students are using these tools in substantially different ways than any learning tools before: "The possibilities of what *Digital Natives* can do online are growing exponentially" in ways that "online life has become an entire strategy for how to live, survive and thrive" and "it's as though the cognitive structure were parallel and no longer sequential" (Richardson, 2006, pp. 6-7). Godwin-Jones (2007), a frequent contributor to *Language Learning and Technology* states that "the much-ballyhooed Web 2.0 is essentially a transition from online consumer to consumer/producer/participant" (p. 8), a self-construct that many students embody quite naturally.

Web 2.0 as an emerging social technology phenomenon is at the heart of what we have come to understand as e-Learning 2.0, environments encouraging collaborative learning via the use of innovative, community driven technologies and tools. Because of the Open Source Movement, a set of principles and practices that promotes access to the development of design and the production of software, user-generated content can now be created through individual effort or through collaboration among instructional designers.

MAIN FOCUS OF THE CHAPTER

With the arrival of technologies that provide a variety of media connecting students to teachers, peers and learning materials, online learning has come to describe a new way of reaching students and connecting them with one another. Today's instructional models have to address the complexities of online learning environments, meeting the needs of students and teachers to learn and teach in an effective way as well as the demands online learning places on them to rediscover and redefine learning and teaching practices.

Online learning students and teachers have different motivations; they may prefer working independently from their peers or they may face barriers of transportation, scheduling, and/or accessibility to services that prevent participation in a traditional school. More and more students and teachers, however, make a conscious choice to learn and teach online, with new technology in mind that makes new ways of learning and teaching possible. For instance, in the case of immigrants who are awaiting passage to new countries, online learning provides an excellent opportunity to develop the language skills and acquire cultural knowledge they will need in their new country. Today, education brings together life-long learning theory with the ideas of distributed, blended, and flexible learning that have emerged within the context of technology-based instruction over the last two decades

As new computer and communications technology has emerged, together with software applications such as browsers and other clients, distance learning has become synonymous with learning online. However, while distance learning carries the interpretative baggage of its principal defining characteristic, that is the physical separation between student and instructor, online learning is often too narrowly defined by the extent of networks such as the Internet or intranets. While both promise learners to be able to learn anytime and anyplace to a lesser or greater degree, the implications for how students learn and teachers teach run deeper. Overall, it can be stated that online learning uses technology to breach the distance where there is a separation of student and instructor in time and space.

In exploring the potential use of technologies as a medium for learning, authors and academics have looked at the challenges for students and teachers. These challenges have defined our understanding of Web-based learning and teaching and guided us closer and closer to the paradigm shift exemplified by Web 2.0 environments, but a common understanding about aspects of quality and the different perspectives available in the design of online and distance learning environments is still needed.

The following statements may best reflect common convictions about learning that online learning environments should bring to life from the perspectives of developing a program primarily based on the technological infrastructure, the availability of learning content, and providing flexibility to learners:

- **Technology perspective:** Learning is *distributed* in that it "makes use of mixed or multimedia tools to bridge the distance between teacher and learner" (Utah Education Network, 2007, n.p.).
- **Content perspective:** Learning is *blended* in that it "employs multiple strategies, methods, and delivery systems" including e-based and print-based resources (Node, 2001, p. 5).
- **Learner perspective:** Learning is *flexible* in that it "expands the choice on what, when, where, and how people learn" (Australian National Training Authority, 2003, p. 3).

Over the past 20 years, at the same time as education has become intertwined with learning online, the Internet has evolved from information and communication technology environments to a network of virtual spaces built on the dynamics of social communities. This second generation of the World Wide Web has come to be known as Web 2.0 bringing social aspects to the foreground while technology steps into the background and fulfills its intended purpose as a medium for learning and teaching.

The emphasis on community and social networks in Web 2.0 has a strong connection to theories of social constructivism and the learners' need to create meaning. Within this context social-constructivist learning theory has reemerged as an approach to learning independently and embedded within a social community. Social constructivism is the approach for online and distance learning for the e-Learning 2.0 paradigm and the Web 2.0 era.

As the tools of e-Learning 2.0 make their way into the hands of users, creating community and working with others online has become easier. Language learning blogs are being used to connect ideas and people around the world, educational podcasts and the communities that pop up to develop and listen to them are a vibrant and valuable phenomenon, *YouTube* has found its way into everyday language. Ideas about the decentralization of authority, freedom to share and re-use information fit perfectly with modern notions of learning as less of a transfer of knowledge from a teacher and more of students learning from each other. *WikiEducator* has begun a listing of free classroom handouts and is planning to encourage teachers to collaborate on developing free textbooks using the wiki platform.

A New Way of Learning and Teaching

Face-to-face instruction often assumes the teacher's ownership of knowledge and transmission of it to the learner, while online learning should be built with the student at the centre of the learning environment. A social-constructivist approach helps focus resources and support for learners to enable them to actively use new material rather than passively absorb information presented to them. According to Warschauer (Berners-Lee & Caillau, 2000, p. 93), the evolution of Computer-Assisted Language Learning can be divided into 3 trends: Behavioral, Integrative, and Communicative, mirroring the evolution of technology and the evolution of linguistic and instructional sciences. The shift to a social-constructivist approach signifies a fundamental change to "willful, reflective, active, conscientious and constructive" learning (Jonassen & Land, 2000, p. v) and collaborative learning represents a key concept between teachers and students within this approach.

Berners-Lee and Caillau (2000) state that:

"in the past, methods have tended to be specialized and exclusive in technique, banishing what preceded to the scrapheap of failed technology [but] it is now generally believed that successful language learning involves competence in a large number of complex and integrated skills and that successful language teaching is more likely to result from using a combination of several different language teaching approaches as no single approach can be said to be entirely successful on its own." (p. 101)

As the epistemological foundations on which social-constructivist convictions are built, differing from cognitive theories of learning (Jonassen & Land, 2000) by expanding on them, there is much room to accommodate different teaching approaches. Social constructivism presumes that learning is a process of individual interpretation and meaning making based on a variety of experiences, and that knowledge is constructed from these experiences (Jonassen, 1991). Additionally, social-constructivist learning processes state that social interaction or social constructivism is not merely supportive of but an essential ingredient in cognitive development (Duffy & Cunningham, 1996). These fundamental changes envision learning as a social dialectic process of meaning. A learning event is characterized by internal and social negotiation between participants of an activity situated in a community. This represents a paradigm shift in the way we think of the process of learning, and the learning medium. The locus of knowledge shifts from the teacher to the learner.

This approach values authentic activities that allow learners to explore, discuss, and construct concepts and relationships relating to real-world problems and projects. Content must be relevant and meaningful to the learner (Donovan et al, 1999). This type of learning is situated in communities of learning and practice as opposed to within the minds of individuals (Lave & Wenger, 1991). Collaborative learning is a key concept between instructors and students, and "two-way interaction is critical in learning a second language" (Ariza & Hancock, 2003, p. 2). The interactions between student-teacher, student-student and student-content are the main media for learning to take place, and specifically for language learning the two-way interaction is important (Pica, 1996). Ariza and Hancock (2003) explain that "while Krashen (1994) believes that only one-way comprehensible input is required for Second Language Acquisition" (pp. 2-3), Lightbrown and Spada (1999) believe that students learn best "when [they] are given the opportunity to engage in meaningful activities [and] are compelled to negotiate for meaning, that is, to express and clarify their intentions, thoughts, opinions, etc., in a way which permits them to arrive at a mutual understanding. This is especially true when the learners are working together to accomplish a particular goal" (p. 22). The possibility and desire to include two-way interactions represent one of the important developments over the history of using technology in language learning and in the history of approaches to language learning in general.

Especially for second language learning, the immersion of learners in authentic environments is paramount. Authentic environments are those that make learning meaningful to students, often because they have a common goal either to achieve a similar outcome or to engage in a similar learning process. Authentic environments need to be situated in real world contexts that learners can relate to because of their previous life, education, or work experiences. As an alternative or even at the same time, authentic environments can be situated learning environments that create contexts similar to the real world in which learners will have to apply the skills that they are learning. The terms "authentic" and "situated" learning are often used interchangeably.

This experience of immersion into a new learning environment and the adaptation to a new way of learning is a profoundly social-constructivist experience where learning through knowledge acquisition and problem solving can take place. In order to arrive at a successful skill transfer to and application of knowledge in another concrete context, learning must occur in authentic environments and be based on the learning of general skills that are transferable to a variety of situations (Winn, 1993).

In authentic learning environments such as these, knowledge and skills are acquired through a process of social communication and discourse. Activities that focus on social communication and discourse also focus on the connections with the community and the patterns of participation, that is connections with the community comprised of all participants who contribute to the learning experience and in the ways they contribute (Duffy & Cunningham, 1996).The following basic tenets may be derived for the practice of pedagogical design:

- Learning is a process of construction based on and situated in experience.
- Learning is based on the instruction of authentic and transferable skills.
- The student is at the centre of the learning process.
- The student is a distributed and multidimensional participant.

e-Learning 2.0

The term e-Learning 2.0 appeared first in an online article by Downes in 2005, referring to a second phase of e-learning embedded in Web 2.0 environments. It describes a new generation of e-based learning environments that allow students to create content, collaborate with peers to form a learning network with distribution of content creation and responsibilities.

According to Karrer in *Understanding e-Learning 2.0*, the second generation of the Web is one of the primary forces behind this new learning paradigm. The key components to Web 2.0 are its tools that allow for collaboration and social interaction to take place. Karrer points to the emergence of a collective intelligence as a result from e-Learning 2.0 that unfolds naturally. Web 2.0 allows learning content to be aggregated together from various sources using various tools but the central idea of e-Learning 2.0 runs deeper. In many respects, collaboration and social interaction result in the creation of content rather than building collaborative learning activities around existing material.

As a result, e-Learning 2.0 also reverses the notion of traditional learning models in terms of the roles of its participants. That is, learning content is no longer produced by publishers and organized into structured courses by teachers. Students take an active part in all aspects of their learning experience, including content creation and learning management on an ongoing basis.

Teachers Learn to Teach

Within a social-constructivist learning approach teachers are generally regarded as facilitators guiding learners through their interaction with the learning material and supporting the collaboration with other learners. The teacher has the expertise and the skills to bring the student to the appropriate learning and will work to create a learning environment where knowledge building is fostered through social exchange.

On the one hand, online teachers share these characteristics with classroom teachers, and like them, must have sufficient knowledge of their subject domain and can be expected to convey enthusiasm for the subject and for their task as a learning motivator; and both types of teachers must have access to appropriate learning activities.

On the other hand, an online learning environment requires different approaches and methods. And different qualities define an online teacher. Most importantly, online teachers must also have sufficient technical skills to navigate and contribute effectively within the online learning context, access necessary hardware, and sufficient Internet efficacy to function within the inevitable technical challenges of these new environments (Anderson & Elloumi, 2004).

Special challenges confront teachers at a distance: these are also believed to hold for online learning environments in general. For example, the teacher must (Gottschalk, 1995, p. 2):

- Develop an understanding of the characteristics and needs of online students with little first-hand experience and limited, if any, face-to-face contact.
- Adapt teaching styles taking into consideration the needs and expectations of multiple often diverse, audiences.
- Develop a working understanding of delivery technology, while remaining focused on their teaching role.
- Function effectively as a skilled facilitator as well as content provider and manager.

As Gottschalk further suggests, “The instructor often finds it beneficial to rely on a site facilitator to act as a bridge between the students and the instructor. To be effective, a facilitator must understand the students being served and the instructor’s expectations” (Gottschalk, 1995, p. 2).

Teachers have an important place in these programs. Current research on distance and online learning indicates that interaction between learners and teachers through face-to-face, telephone, or electronic means is vital to the learning process in these programs (Porter & Sturm, 2006). During the course of this research, the following skills teachers need to know when working online with students emerged:

- How to use assessment tools with learners at a distance
- How to use synchronous and asynchronous online tools
- How to give appropriate feedback online
- How to motivate and encourage students at a distance
- How to create a positive learning environment online
- How to manage discussions in chat rooms
- How to organize and monitor project work online
- How to retain students online
- How to integrate technology into language learning programs
- How to facilitate learning in a Web-based environment
- How and where to grow and update their skills as new online communications and student support technologies evolve

Are teachers prepared to integrate Web 2.0 tools into their programs? The specific needs of teachers in online language learning programs are rarely mentioned. The management of online learning environments encompasses everything from administration to lesson planning to assessment and evaluation, with teachers often involved in these processes every step of the way. Warschauer (2000) noted that he sees distance education as a realm in which the role of technology will be a site of struggle in increasing or lowering the quality of learning opportunities, which could also bring to a head issues about the professional standing of educators in the field of online learning (White, 2003). Although it is often helpful and practical for teachers to develop their tech skills informally, many of these skill areas are not easy to develop without guidance.

Furthermore, teachers face additional challenges in the shape of structural and attitudinal barriers. Common structural barriers may be due to the way the computers are used within a program, if they are integrated in a way that maximizes the use of the resources available to teachers, if opportunities to practice new technology skills are available, and if the educational institution values this learning process. Common attitudinal barriers are often rooted in the way technology forces change that disrupts programs. In the early stages of technology integration, teachers are often overwhelmed and need to be motivated to continue their learning process that will result in achieving their goals of improving student independence (Kennell, 2004).

An excellent resource when considering online teacher competencies is Salmon's *E-Moderator Online Competencies*.Table 1, adapted from Salmon (2002, p. 41), presents a summary of the qualities and characteristics that Salmon states are key competencies for the changing role of a teacher in an online learning environment. She also makes the point that good face-to-face teachers do not necessarily make the best online teachers and notes that face-to-face subject teachers who are used to being "experts" might have difficulty adapting to the levelling effect and informality of online discussion (Salmon, 2002, p. 42).

The Milken Exchange's Professional Competency Continuum: Professional Skills for the Digital Age Classroom provides an interesting assessment for instructors to see where their technology skills lie on a continuum that looks at five target areas of skill: core technology skills; curriculum, learning, and assessment; professional practice; classroom and instructional management; and administrative competencies (Milken Family Foundation, 2008).

Learners Teach to Learn

Online learning skills are necessary in the wider world beyond language training. Increasing opportunities for online learning increases the learner's ability to participate in a multitude of learning opportunities and to take a leadership role among their peers. The world of work integrates online collaboration into more and more of the structure

Table 1. Key competencies of an online teacher

Quality/ Characteristic	**1. CONFIDENT**	**2. CONSTRUCTIVE**	**3. DEVELOPMENTAL**	**4. FACILITATING**	**5. KNOWLEDGE SHARING**	**6. CREATIVE**
Understanding of online process	Personal experience as an online learner, flexibility in approaches to teaching & learning. Empathy with the challenges of becoming an online learner	Able to build online trust & purpose for others, Understand the potential of online learning & groups	Ability to develop & enable others, act as catalyst, foster discussion, summarize, restate, challenge, monitor understanding & misunderstanding, take feedback	Know when to control groups, when to let go, how to bring in non-participants, know how to pace discussion & use time on line, understand the 5-stage scaffolding process & how to use it	Able to explore ideas, develop arguments, promote valuable threads, close off unproductive threads, choose when to archive	Able to use a range of approaches from structured activities (e-tivities) to free wheeling discussions, & to evaluate & judge success of these
Technical skills	Operational understanding of software in use reasonable keyboard skills; able to read fairly comfortably on screen, good, regular, mobile access to the Internet	Able to appreciate the basic structures of bullet boards, forums, & the WWW & Internet's potential for learning	Know how to use special features of software for e-moderators, e.g. controlling, weaving, archiving, Know how to "scale up" without consuming inordinate amounts of personal time, by using the software productively	Able to use special features of software to explore learner's use e.g. message history	Able to create links between online & other features of learning programs	Able to use software facilities to create & manipulate conferences & to generate an online learning environment, able to use alternative software & platforms
Online communication skills	Courteous & respectful in online (written) communication, able to pace & use time appropriately	Able to write concise, energizing, personable online messages	Able to engage with people online (not the machine or the software), respond to messages appropriately, be appropriately "visible" online, elicit & manage students' expectations	Able to interact through e-mail & conferencing & achieve interaction between others, be a role model Able to gradually increase the number of learners dealt with successfully online, without huge amounts of extra personal time	Able to value diversity with cultural sensitivity, explore differences & meanings	Able to communicate comfortably without visual cues, able to diagnose & solve problems & opportunities online, use humour online, use & work with emotion online, handle conflict constructively
Content expertise	Knowledge & experience to share, willingness to add own contributions	Able to encourage sound contributions from others, know of useful online resources for their topic	Able to trigger debates by posing intriguing questions	Carry authority by awarding marks fairly to students for their participation & contributions	Know about valuable resources (e.g. on the WWW) & refer participants to them, and use them as sparks for participation	Able to enliven conferences through use of multi media & electronic resources, able to give creative feedback & build on participants' ideas
Personal Characteristics	Determination & motivation to become as an e-moderator	Able to establish an online identity as e-moderator	Able to adapt to new teaching contexts, methods, audiences & roles	Show sensitivity to online relationships & communication	Show a positive attitude, commitment & enthusiasm for online learning	Know how to create & sustain a useful, relevant online learning community

of professional and personal life, much like many public and academic libraries have done adopting the *23 things* learning paradigm. This could be the kind of on the job training that language students might also encounter, especially internationally trained professionals and trades people.

Online learning built on e-Learning 2.0 principles, that is learning environments that employ social-constructivist principles and include Web-based environments for communal spaces of learning and teaching, make special demands on students and teachers. Students will be working with little or no face-to-face support, and without encountering fellow students in physical spaces but depending on the suite of Web 2.0 tools at their disposal they may be connected to their teacher and peers more or less in *real-time*. For instance, internationally trained individuals are often expected to connect with mentors in their field by contributing to a blog or by listening to a podcast.

In synchronous environments, that is *real-time* exchange between the participants, student support can be implemented similarly to and in some ways better than in face-to-face classrooms. The immediacy of responses that face-to-face environments provide, for better or for worse depending on the cues of teachers acknowledging or discounting students' willingness to participate, is as much the key in student retention in online environments but the net of connections between students, their peers and their teacher can be cast wider. The tools at the disposal of all participants provide for more distributed participation patterns with the potential of supporting students adequately. Also, it is easier for students to seek support from their peers in a less disruptive manner to the group when using tools like chat or instant messaging concurrently with virtual classroom applications, such as *Centra* and *Elluminate*.

In asynchronous environments, when students are also separated in time from the teacher and their peers and cannot expect to get an immediate response, they must be able to maintain their motivation on their own even though feedback from teachers and fellow students may be delayed. Online learning options that merely distribute learning materials by electronic means for the sake of access, expediency, and convenience without adding the value to Web 2.0 tools to enrich students' experiences suffer the same fate as correspondence courses and remain one-dimensional and undistributed. When designed and employed at the service of students and teachers within a community, asynchronous tools cannot only give anytime anywhere access to learning but can also empower students to self-manage their learning experiences, and take on roles of expertise and leadership among their peers at their own pace.

Often an ideal learning environment is built around both synchronous and asynchronous elements that provide students and teachers with an array of communication and learning tools. In order to be most effective in any online program, students need to be willing to work towards becoming more autonomous and self-directed to fulfill their multi-dimensional and distributed connections with other participants in the same space. This is substantially different from traditional classrooms where one-to-one and group-to-one interactions are mandated by the limitations of a physical environment. In order to participate to the fullest and be successful, students must see that they have an essential part to play to bring the benefits of online learning environments about. According to the Illinois Online Network (2007, n.p.), students need to:

- Be open minded about sharing life, work, and educational experiences as part of the learning process
- Be self-motivated and self-disciplined
- Be willing to "speak-up" when problems arise
- Be willing and able to meet the minimum requirements of the program
- Accept critical thinking and decision making as part of the learning process

- Be able to think ideas through before responding
- Feel that high quality learning can take place without going to a classroom

Naturally, some students begin with more skills for online learning than others. Students without online learning experience may perform lower overall than students with online learning experience. Inexperienced students may communicate more about difficulty with software, the course management system, or mistakes in submitting work online. In contrast, experienced students may perform better overall and their more frequent communications may be indicative of seeking clarification of course concepts at higher levels. At any stage, students need to engage in self-assessment as a key step in the process of learning online successfully. Furthermore, "multi-modal, multi-tiered products … provide students a vehicle for drawing on varied strengths and making positive contributions, regardless of whether they are fluent in [a second language]" (Warschauer, 2007, p. 2536).

Some studies have shown that "women may be more successful in online environments than men because they frequently create a sense of community by connecting with other learners" (Imel & Jacobson, 2006, p. 2; Rovai, 2003). In a learning environment that relies on students to communicate not just with a teacher but with other students as well, this is important to keep in mind for teachers customizing an e-Learning 2.0 environment. Furthermore, younger adults are more likely to drop out of online classes than older classmates (Jonassen et al., 1999) because they may not have the same learning persistence. On the other hand, many younger students are more adept at using a computer and are more familiar with e-mail, chat rooms and the Internet in general than older students. With a variety of Web 2.0 tools at hand, teachers need to select the ones that meet the needs of their students and keep up with the demands of the learning material.

The ability of learners to manage their own learning is a key competency that studies have shown to be related directly to successful participation in online learning environments. Porter and Sturm (2006, p. 105) used the following self-management survey to evaluate learners competency level by asking them about how well they thought they did in the these areas:

1. Staying focused or concentrate on what they are doing
2. Sticking with a task or problem
3. Figuring things out for themselves before asking for help
4. Asking for help when they're stuck
5. Making decisions for themselves
6. Solving problems by themselves
7. Feeling they can do things and accomplish things
8. Organizing their work and life
9. Learning things on their own
10. Setting goals for themselves
11. Managing their time
12. Evaluating their own progress and how they are doing
13. Trying or learning new things
14. Learning on their own without help
15. Accepting responsibility for themselves
16. Seeking constructive criticism of their work
17. Trying to actively try new things

The Pedagogy of Blogs and Wikis

Years before the emergence of Web 2.0 tools, Jonassen, Peck and Wilson (1999) argued that using technology as storage places for learning material does not exploit the capabilities of technology, teachers and students, but that technology can amplify students' ability to construct knowledge. Jonassen, Carr, and Yueh (1998) described "computer applications that, when used by learners to represent what they know, necessarily engage them in critical thinking about the content they are studying" as mind tools, and continues to say that "mind tools scaffold different forms of reasoning about content. That is, they require students to think about what they know in different, meaningful ways. … Students cannot use mind tools as learning strategies without thinking deeply about what they are studying" (p. 1). Web technology should be used to engage students in critical thinking and enable them to become intelligent designers when computers and networks serve as catalysts for facilitating planning, decision-making and self-management skills when they are used in ways to promote reflection, discussion, and problem-solving by the teacher.

Blogs and wikis exemplify these possibilities in powerful ways for students and teachers. Richardson (2006) calls blogs "a truly constructivist tool for learning" (p. 27) because their content is part of a wider body of knowledge accessible and potentially relevant to an audience outside the classroom. A high degree of *information literacy* is required. The American Library Association defines *information literacy* as "the ability to access needed information effectively and efficiently; evaluate information and its sources critically; incorporate selected information into one's knowledge base; use information effectively to accomplish a specific purpose; and understand the economic, legal, and social issues surrounding the use of information" (Warschauer, 2007, p. 2512). Furthermore, blogs and wikis facilitate reflection and metacognitive analysis through archiving, they support different learning styles, and they provide students with opportunities to acquire the new literacy skills needed in a more and more knowledge and information-based society (Richardson, 2006). Writing on a blog is not merely writing using another medium; the medium transforms the writing process and extends its reach by making it more conversational, collaborative and in the end democratic.

Wikis take the democratization of learning even further by allowing everyone to be an editor and thus exemplary collaborative learning to emerge, including astounding self-regulatory practices of quality control. *Wikipedia*, the online encyclopedia build on the wiki paradigm, is the "poster child for the collaborative construction of knowledge and truth that the new, interactive Web facilitates" (Richardson, 2006, p. 61). Wikipedia has grown rapidly into one of the largest reference Web sites since 1995 and has sparked the emergence of other tools based on the wiki paradigm of collaborative knowledge management like the *WikiEducator*, an evolving community intended for the collaborative planning of education projects linked with the development of free content.

Because of the potential of open source products like wikis, many education institutions have been moving away from their exclusive use of more restrictive commercial systems, their recurring licensing and upgrading fees being only part of the problem. Many still retain these systems due to the need for the learning management features they provide, but more and more are trying out and reporting success with open source systems like *Moodle* that is designed to help educators create online courses with opportunities for rich interaction and manage groups of students effectively by providing a free online learning platform supported by a global community of developers. *Moodle* is designed to be extremely flexible for instructors and learners, and can be downloaded and used on nearly any mainstream computer; it readily scales from single-instructor or departmental Web sites to 50,000-student universities (Instructional Technology Resource Center, 2006).

Although quality control in Web 2.0 learning environments is a major challenge, for instance Richardson (2006) adds that because of the wiki's democratic process of knowledge creation students begin to teach each other when put to the task to negotiate to agree on correctness, meaning, and relevance with their peers. With a system as openly accessible as *Wikipedia*, its potential for collaborative learning appears to be matched by its potential for unethical use. When anyone can make changes, information can be easily falsified, and the collaborative community serves as a watchdog. Richardson argues that "giving students editorial control can imbue in them a sense of responsibility and ownership" (p. 64).

Many Web 2.0 tools come together seamlessly within blogs, wikis, and learning management systems like *Moodle*, such as automatically updated RSS (Rich Site Summary) feeds from news Web sites and other blogs, audio and video podcasts (vodcasts) from amateur and professional producers, or social bookmarking features, to name just a few tools that extend the reach of distributed learning applications. The community building properties of blogs and wikis, along with other synchronous or asynchronous Web-based communication and content creation tools, build communities of practice into the paradigm of an online learning environment. In traditional classrooms, there are often social communities or cliques but they are rarely based on common learning goals since knowledge building is seen as rooted in the individual student's learning process.

FUTURE TRENDS

Language learning environments that traditionally provide opportunities for repetitive practice can still help students with lower language proficiency master the skills they need before they can handle Web 2.0 tools like blogs and wikis effectively, but it is also important to keep in mind that students' language proficiency does not necessarily parallel their familiarity and proficiency with technology. Many students may have had varied experience with Web 2.0 tools using them for professional or personal reasons. Given the growing extent that these tools are integrated in today's working environments, communal and personal spaces, the more they are integrated in learning environments the more authentic learning experiences are and the higher rate of skills transfer students experience.

It has become more and more likely for students to be asked to create and manage content for Web 2.0 environments. Richardson (2006) points out that there is a clear disconnect between the traditional teacher who grades independent study assignments aimed at a very limited audience and the students who need to be critical readers of Web-based content, literate in Web 2.0 publishing, comfortable with virtual collaboration, and good information managers. It is this disconnect that may present a major barrier in student achievement.

According to Warschauer (2006), "the ability to learn autonomously will indeed be critical in the digital future" and he continues to say that "strong mentorship is required for students to achieve this autonomy" (p. 46) at the same time. It is through the multi-dimensional ways that Web 2.0 learning environments allow learners to connect and collaborate with teachers and other learners that independent learning and meaning making embedded in a social context and constructs is made possible. Richardson (2006) identifies the social, collaborative construction of meaningful knowledge as one of the paradigm shifts that Web 2.0 technologies demand a reexamination of the way we learn and teach. Producing work in truly collaborative ways for large audiences creates a new social context that requires teachers to rethink the demands placed on the students. Instead of completion of an assignment, contribution to project becomes more and more the ultimate goal. As Warschauer (2006) comments, "New technologies do not replace the need for strong human mentorship, but, indeed, amplify the role of such mentorship" (p. 48). Teachers need to see themselves as *connectors* not only between students and the learning content but also with their peers. They also need to become *content creators* using Web 2.0 tools, *collaborators* in the sense of learning alongside their students, and *coaches* modeling skills students need as well as motivating them to take responsibility and ownership for their own performance. Last but not least, teachers need to become *change agents* using Web 2.0 tools to move towards a new way of learning and teaching (Richardson, 2006, pp. 132-133).

CONCLUSION

In this chapter, social constructivism has been proposed as the foundation for online language learning environments that foster the participation of students and teachers in today's knowledge and information-based society to their full potential through the use of Web 2.0 tools like blogs and wikis. It has been argued that teachers and students need to take full advantage of these emerging tools to participate in more dynamic, immediate, and communicative environments that provide opportunities for meaningful experiences through social constructivist learning.

While putting the pedagogy of blogs and wikis to work may take some time, the following recommendations can assist program designers and teachers in making the first step towards participatory and collaborative online learning of the Web 2.0 era:

- Where possible, orientation "events" should be held to introduce students to the requirements of the technology and the expectations for student-to-instructor and peer-to-peer communication. The orientation process can be a critical factor in the success of online learning programs (Johnston et al., n.d.).
- When possible, begin the course by providing traditional face-to-face instruction and then blend it with online education. Face-to-face instruction can provide students with a little online learning experience with support and help them develop confidence in their ability to succeed using on-line learning tools (Johnston et al., n.d.).
- Help students develop their ability to engage in self-directed learning. If the online learning program is complemented with some traditional instruction, teachers can spend time in class working with students on self-management strategies and help change their perceptions of themselves as students and allowed them to take ownership of their learning (D'Amico & Capehart, 2001). Provide opportunities for students to take leadership and engage in peer tutoring.
- In an online learning program, ongoing support should be provided for students through frequent contact with teachers via multiple modalities, e.g. e-mail, instant messenging, chat, or telephone contact. Learners should be offered opportunities to participate in online learning at the earliest possible point in their language learning.
- A tool for students to self-assess their online learning skills should be accessible to them. This tool should assess students' familiarity with technology, their experience in online learning, their problem solving skills, their ability to motivate themselves, their level of self-directedness as well as their level of English fluency. The results should be shared with the student to help in determining their preference for learning environments.
- Encourage students and fellow teachers to explore and experiment with Web 2.0 tools and their potential for learning and teaching. Professional development activities should take advantage of the same technologies so that teachers learn about e-Learning 2.0 the same way students would and understand the implications of the demands placed on students and their expectations better.

REFERENCES

Anderson, T., & Elloumi, F. (2004). *Theory and practice of online learning*. Athabasca University. Retrieved January 10, 2008, from http://cde.athabascau.ca/online_book/index.html

Ariza, E. N., & Hancock, S. (2003). Second language acquisition theories as a framework for creating distance learning courses. *International Review of Research in Open and Distance Learning*, October. Retrieved January 10, 2008, from http://www.irrodl.org/index.php/irrodl/article/view/142/710

Australian National Training Authority. (2003). Definition of key terms used in E-Learning (Version 1.00), in *Australian Flexible Learning Framework Quick Guides series*. Backroad Connections Pty Ltd. Retrieved January 10, 2008, from http://pre2005.flexiblelearning.net.au/guides/keyterms.pdf

Berners-Lee, T., & Caillau, R. (2000). *The history of computer assisted language learning Web exhibition*. Retrieved January 10, 2008, from http://www.eurocall-languages.org/resources/history_of_call.pdf

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). "The semantic Web". *Scientific American Magazine*. Retrieved May 8, 2008, from http://www.sciam.com/article.cfm?id=the-semantic-web

Cuban, L. (1996). *Teachers and machines*: *The classroom technology since 1920*. Teachers Colleges Press.

Cuban, L. (2002). *Oversold and underused: Computers in the classroom*. Harvard University Press.

D'Amico, D., & Capehart, M. A. (2001). Letting learners lead: Theories of adult learning and TV411. Focus on Basics, 5(B), 35-37, 39, in *California Adult Education Research Digest No. 44 – Technology – May 2006*. Retrieved January 10, 2008, from http://www.calpro-online.org/documents/100924DistanceEducationProof.pdf

Donovan, S., Bransford, J., & Pellegrino, J. (Eds.). (1999). *How people learn: Bridging research and practice*. National Academy of Sciences. Retrieved January 10, 2008, from http://bob.nap.edu/html/howpeople2/

Duffy, T. M., & Cunningham, D. J. (1996). Constructivism: Implications for the design and delivery of instruction. In D. H. Jonassen (Ed.), *Handbook of Research for Educational Communications Technology* (pp. 170-198). Simon & Schuster Macmillan.

Godwin-Jones, R. (2007). Emerging technologies: E-texts, mobile browsing, and rich Internet applications, in *Language Learning and Technology*, *11*(3), October, 8-13. Retrieved January 10, 2008, from http://llt.msu.edu/vol11num3/pdf/emerging.pdf

Gottschalk, T. H. (1995). Distance education: An overview. *Distance Education at a Glance, Guide 1*. University of Idaho Engineering Outreach. Retrieved January 10, 2008, from http://www.uidaho.edu/eo/guide1.pdf

Illinois Online Network. (2007). *What makes a successful online student?* Illinois Online Network and the Board of Trustees of the University of Illinois on the World Wide Web. Retrieved January 10, 2008, from http://www.ion.uillinois.edu/resources/tutorials/pedagogy/StudentProfile.asp

Imel, S., & Jacobson, E. (2006). *Distance education and E-Learning: New options for adult basic and English language education*. Research Digest No. 4, Technology. California Department of Education. Retrieved January 10, 2008, from http://www.calpro-online.org/documents/100924DistanceEducationProof.pdf

Instructional Technology Resource Center. (2006). *LMS pilot report (Moodle).* Idaho State University. Retrieved January 10, 2008, from http://www.isu.edu/itrc/resources/LMS_Pilot_Report_FINAL_Fall06.pdf

Johnston, J., Petty, L. I., & Young, S. J. (n.d.). *Research overviews: Technology in adult education.* Institute for Social Research, University of Michigan. Retrieved January 10, 2008, from http://www.calpro-online.org/documents/100924DistanceEducationProof.pdf

Jonassen, D. H. (1991). Objectivism versus constructivism: Do we need a new philosophical paradigm? *Educational Technology Research and Development*, *39*(3), 5–14. doi:10.1007/BF02296434

Jonassen, D. H., Carr, C., & Yueh, H.-P. (1998). Computers as mind tools for engaging learners in critical thinking. *TechTrends*, *43*(2), 24–3. doi:10.1007/BF02818172

Jonassen, D. H., & Land, S. M. (2000). Preface. In D. H. Jonassen & S. M. Land (Eds.), *Theoretical foundations of learning environments*. Lawrence Erlbaum Associates.

Jonassen, D. H., Peck, K. L., & Wilson, B. G. (1999). *Learning with technology: A constructivist perspective*. Merril/Prentice Hall.

Kennell, T. (2004). *Guide to blended learning*. Skills at Work Series. AlphaPlus Centre, from http://alphaplus.ca/pdfs/SkillsatWorkGBL.pdf

Krashen, S. (1994). The input hypothesis and its rivals. In N. Ellis (Ed.) *Implicit and explicit learning of languages* (pp. 45-77). Academic Press.

Lave, J., & Wenger, E. (1991). *Situated learning - legitimate peripheral participation*. Cambridge University Press.

Lightbrown, P. M., & Spada, N. (1999). *How languages are learned.* Oxford University Press.

Milken Family Foundation. (2008). *The Milken Exchange's Professional Competency Continuum: Professional Skills for the Digital Age Classroom*. Retrieved January 10, 2008, from http://www.mff.org/publications/publications.taf?page=280

Node (2001). *The Node's guide to blended learning: Getting the most out of your classroom and the Internet*. The Node Learning Technologies Network.

O'Reilly, T. (2005). *Web 2.0: Compact definition*. Retrieved January 10, 2008, from http://radar.oreilly.com/archives/2005/10/web_20_compact_definition.html

Pica, T. (1996). Second language learning through interaction: Multiple perspectives. *Working Papers in Educational Linguistics*, *12*(1), 1–22.

Porter, P., & Sturm, M. (2006). *Crossing the great divides: Distance learning and flexible delivery in adult basic education*. AlphaPlus Centre. Retrieved January 10, 2008, from http://distance.alphaplus.ca

Prensky, M. (2001). Digital natives, digital immigrants. Retrieved January 10, 2008, from http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf

Prensky, M. (2004). *The emerging online life of the digital native: What they do differently because of technology, and how they do it*. Retrieved January 10, 2008, from http://www.marcprensky.com/writing /Prensky-The_Emerging_Online_Life_of_the_Digtal-Native-03.pdf

Richardson, W. (2006). *Blogs, Wikis, podcasts, and other powerful tools for classrooms*. Corwin Press.

Rovai, A. P. (2003). In search of higher persistence rates in distance education online programs. *The Internet and Higher Education*, *6*(1), 1–16. doi:10.1016/S1096-7516(02)00158-6

Salmon, G. (2002). *E-moderating: The key to teaching and learning online*. Kogan Page.

Utah Education Network. (2007). Distance Education: Glossary. *Utah Education Network* in partnership with the *Utah State Office of Education* and the *Utah System of Higher Education* on the World Wide Web. Retrieved January 10, 2008, from http://www.uen.org/distance_ed/glossary.shtml

Warschauer, M. (2000). The changing global economy and the future of English teaching. *TESOL Quarterly*, *34*, 511–535. doi:10.2307/3587741

Warschauer, M. (2006). The paradoxical future of digital learning. *Learning Inquiry*, *1*, 41–49. doi:10.1007/s11519-007-0001-5

Warschauer, M. (2007, November). Information literacy in the laptop classroom In . *Teachers College Record*, *109*(11), 2511–2540.

White, C. (2003). *Language learning in distance education*. Cambridge University Press.

Wikipedia (2007). Web 2.0. *Wikipedia*. Retrieved January 10, 2008, from http://en.wikipedia.org/wiki/Web_2

Wikipedia (2008). Wiki. *Wikipedia*. Retrieved January 10, 2008, from http://en.wikipedia.org/wiki/Wiki

Winn, W. (1993). A constructivist critique of the assumptions of instructional design. In T. M. Duffy, J. Lowyck, & D. H. Jonassen (Eds.), *Designing environments for constructive learning*. (pp. 189-212). Springer-Verlag.

KEY TERMS

Blended Learning: The term Blended Learning describes the design of a learning environment from the viewpoint of how the delivery of learning materials to the students is best accomplished by a variety of means available, be they technological or non-technological in nature. By choosing the appropriate vehicle for the student to access the learning content, a number of different strategies are used to provide hybrid learning environments. Blended Learning is closely related to Distributed Learning and Flexible Learning.

Distributed Learning: This term refers to learning environments that use a mixture of tools to navigate the distance between teachers and learners. From a design viewpoint of a learning environment, building a variety of connections between the participants and the learning content is the main objective, as is allowing patterns of participation to develop between teachers, students and learning materials. Technological tools allow these connections to be made easily. Distributed Learning is closely related to Blended Learning and Flexible Learning.

e-Learning 2.0: The term e-Learning 2.0 refers to the second generation of eLearning making use of the social collaboration and information sharing tools embedded in Web 2.0 environments. It describes a new generation of e-based learning environments that allow students to create content, and collaborate with peers on the creation of content distributed by technological tools. e-Learning

2.0 provides a new learning paradigm naturally unfolding collective intelligences.

Flexible Learning: This term describes a learning design perspective deeply rooted in the needs of students, with the main objective being to provide them with the most flexibility about the learning content, schedules, access, and learning styles as possible. A flexible learning design customizes learning environments to meet the needs of learners, using both technological and non-technological tools. Flexible Learning is closely related to Blended Learning and Distributed Learning.

PLATO: Programmed Logic for Automated Teaching Operation, refers to one of the first computer assisted instruction systems, dating from the early 1970s and running until 2006. PLATO was one of the first systems to test applications such as e-mail, discussion forums, and chat rooms.

TICCET: This stands for Time-shared, Interactive, Computer-Controlled Educational Television. The project ran at the same time as PLATO and was funded by the University of Texas at Austin and Brigham Young University. In place of expensive hardware, the system used television technology with minicomputers to deliver interactive educational content.

Wiki: This is a Web-based environment designed to enable readers to become creators of content and editors of previous entries. Wikis are paradigm examples of Web 2.0 tools that are effectively used to design constructivist learning environments and engage learners in collaborative learning environments. Much like blogs, wikis integrate different types of media from audio to video files, which can be played on demand, as well as podcasts to vodcasts, which readers can subscribe to. Wikis can be an integrated part of a larger learning management system.

This work was previously published in Handbook of Research on Web 2.0 and Second Language Learning, edited by J. Wang, pp. 367-384, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.20
Developing Digital Literacy Skills with WebQuests and Web Inquiry Projects

Susan E. Gibson
University of Alberta, Canada

ABSTRACT

This article identifies digital literacy as an important aspect of new media literacy at the K-12 level. Digital literacy includes developing the skills of information location and application as well understanding how to use available evidence to assist in problem solving and decision making about important questions and issues that have no clear answers. Two web-based examples of instructional strategies – WebQuests and Web Inquiry Projects—are suggested as ways to develop these and other important 21st century learning skills.

WHAT IS DIGITAL LITERACY?

Over the last decade the term ' literacy' has evolved to include an ever increasing, and diverse range of skills. "The new literacies of the Internet and other ICTs include the skills, strategies and dispositions necessary to successfully use and adapt to the rapidly changing information and communication technologies and contexts that continuously emerge in our world and influence all areas of our personal and professional lives" (Leu, Kinzer, Coiro & Cammack, 2004, p. 1572). According to Jamie McKenzie (2005), "Literacy is about wrestling understanding from chunks of information, whether these chunks be numerical, textual, visual, cultural, natural or artistic" (p. 7). One form of literacy, 'digital' literacy, can be defined as "a person's ability to perform tasks effectively in a digital environment, with "digital" meaning information represented in numeric form and primarily for use by a computer... [and] includes the ability to read and interpret media (text, sound, images), to reproduce data and images through digital manipulation, and to evaluate and apply new knowledge gained from digital environments" (Jones-Kavalier & Flannigan, 2006, p. 9).

Developing the skills of information location and application is one aspect of **digital literacy**. These skills include the ability to find, evaluate, synthesize, and use information to answer questions and make informed decisions. Digitized information comes in many forms, and students need to acquire the ability to read, interpret, understand, and use all of these media formats. They need to understand that

DOI: 10.4018/978-1-60566-120-9.ch026

everything on the Web represents an individual's point of view and that all sources need to be carefully and critically examined for authenticity and bias. They also need to recognize that no one source of information can adequately represent all there is to know about a particular topic; multiple sources on any topic should always be consulted and their information compared. **Digital literacy** also involves understanding how to use the available evidence to assist in problem solving and decision making about important questions and issues that have no clear answers. Furthermore students benefit from opportunities in which they are encourage to transform information in new ways to advance their own and other's thinking, rather than simply consuming what others have produced. Finally, students need to develop a critical attitude toward computer technology in our society in terms of its present and future impact on humanity. The overall goal of **digital literacy** is to develop knowledgeable, skilled, and responsible users of computer technologies.

The Partnership for 21st Century Learning [http://www.21stcenturyskills.org/index.php?option=com_content&task=view&id=254&Itemid=120] calls for an emphasis in schooling on all of these literacy skills to ensure that students will be successful in the 21st century. The International Society for Technology in Education's (ISTE) Standards for Educational Technology (2007) also include creativity and innovation, communication and **collaboration**, research and information literacy, critical thinking, problem solving and decision making, digital citizenship, and technology operations and concepts. Addressing all of these components of **digital literacy** is a major undertaking for schools and all teachers, grade levels and subject areas have important roles to play.

This chapter begins by reviewing what we currently know about effective computer use to support and enhance teaching and learning. **Constructivism** is then examined as a promising theoretical framework for that use. The remainder of the chapter looks at **WebQuests** and their extension, **Web Inquiry Projects**, as approaches that have the potential to effectively address both constructivist learning principles and digital literacy, **higher level thinking**, problem solving and communication skills.

WHAT DOES THE RESEARCH TELL US ABOUT WHAT MAKES EFFECTIVE AND MEANINGFUL TECHNOLOGY INTEGRATION?

Before examining ways to address **digital literacy** skills in teaching with technology, it is important to review what we know about effective technology use. Computers are now more readily available in many schools worldwide and the Internet is often hailed as an innovation with unprecedented potential for the improvement of teaching and learning. Although some critics claim that the use of computer technologies has had minimal to no affect on learning outcomes (Cuban, 2001, Oppenheimer, 2003), there have been positive affects identified in the research literature. "Several recent research reviews and meta-analyses published in the United States and Britain suggest that, when measured across the board, educational technology yields "small, but significant" gains in learning and student engagement" (Viadero, 2007 p. 1). Learner motivation has been identified in numerous studies as being particularly evident with the use of computer technologies (Sterling, 2007). As for learning gains, Wan, Fang, and Neufeld (2007) found that, "Technology can influence learning processes by facilitating cognitive information processing activities such as search, scanning, transformation or comparison of information" (p. 187). Higgins (2004) found advances in reasoning, understanding and creativity using computers. Viadero (2007) identified positive affects for writing with the use of word processors, and for generating deeper understanding and increasing knowledge through the use of simulations. Balanskat, Blamire and Kefala (2006) analyzed the evidence from 17 impact studies and found that using information and communication technologies had a positive impact on children's' learning of basic skills such as calculation, reading and writing, and on communication and process skills, while also allowing for greater differentiation to address individual needs and learning styles, and giving more responsibility for the learning to the student. The greatest benefits, according to Balanskat et al. (2006), were seen in primary education. Significant benefits have also been found in the areas of special needs (Hartley, 2007) and English as a Second Language (Lee, 2006).

While the benefits from the use of computer technologies are evident, K-12 teachers continue to be at varied levels of awareness about the possibilities for employing these technologies in effective and efficient ways to enhance teaching and learning. "Effectively integrating new technology into educational practice is not just a matter of learning how to use technology. It is also a process of reflecting on how to teach and how students can learn most effectively in today's world" (Wiske, Franz, & Breit, 2005, p. 3). Where the greatest challenge for teachers lies is in thinking differently about teaching and learning. According to David Thornburg, "The main thing that's holding technology back is...a fear--a well-placed fear, I might add--that if technology becomes ubiquitous, it will totally transform the practice of education. There are a lot of people who don't want the practice of education transformed because they're very comfortable with it" (cited in Brumfield, 2006, p. 1). Computer technologies can help teachers to develop new approaches to teaching and learning, but teachers need to be exposed to these new understandings and new capabilities. They also need to determine where technologies fit into their philosophy of teaching. As noted by Doolittle and Hicks (2003), "A philosophical and theoretical foundation provides answers to the questions of why and how specific pedagogy, including the application of technology, should be employed" (p. 76). The key to best use is not the fact that computers are being used, but how they are being used. Where success has been most apparent, has been in cases where teaching is transformed through the use of computer technologies and where learning is happening in ways that were impossible or difficult without the use of these technologies. "Education can be transformed using ICT which brings new capabilities and capacities to learning. For example, ICT has the potential for enabling teachers and students to construct rich, multi-sensory, interactive environments with almost unlimited teaching and learning potential" (Balanskat et al., 2006, p. 12)

"Researchers are just now understanding how much greater the payoffs can be when digital-learning programs combine specific academic content with lessons from cognitive science and developmental psychology on how children learning in those subjects" (Viadero, 2007, p. 1). Computer use needs to go beyond low-level tasks such as students being able to demonstrate understanding of how to operate the various technologies with proficiency, to tasks that encourage more advanced learning by actively engaging students in learning, by releasing of agency from teachers to students, and through collaborative knowledge building around authentic or ill-defined problems. According to Dunn (2007) the best uses of computers gain learners' attention, engage learners through productive work, increase learners' perceptions of control, help learners visualize problems and solutions, link them to information resources and to learning tools, encourage shared intelligences through collaborative and cooperative learning, and encourage **higher level thinking**. Using an **inquiry** approach to learning with computers can be an effective way of creating a learning environment that places less emphasis on acquiring and presenting information and more on constructing knowledge, making meaning, drawing on personal life experience, and taking responsibility for learning.

The merging of technology and **constructivism** offers many possibilities for framing the design of such innovative learning environments.

HOW CAN CONSTRUCTIVIST LEARNING THEORY HELP TEACHERS TO DESIGN MEANINGFUL, COMPUTER-ENHANCED LEARNING ENVIRONMENTS?

Ferdig (2006) identifies the importance of "tying innovation to learning theory to create authentic and engaging activities for students" (p. 750). Research on effective integration of computer technologies in schools points to uses that support constructivist learning principles (Jonassen, Howland, Moore, & Marra, 2003). Constructivism is a theory about how people learn in which learning is not just about acquiring more knowledge but rather "it's the mental act of reformulating what we thought we knew into something new and different...Learning occurs through conceptual change" (Brooks, 2003, p. 13). This conceptual change occurs through an active and social process. The new learning always begins with and builds upon the learners' previously stored knowledge; as the learners elaborate upon and interpret the new information, their initial ideas are reshaped, and misconceptions in prior knowledge can be addressed through the formation of alternate conceptions (Tarhan, Ayar-Kayali, Urek & Acar, 2007). They are routinely asked to apply knowledge in diverse and authentic contexts, to explain ideas, interpret texts, predict phenomena, and construct arguments based on evidence (Windschitl, 2002). Learning occurs most effectively when it is situated in experiences that are authentic and meaningful to the learner and when they engage in task-oriented dialogue with one another. **Constructivism** as a framework for using computer technologies in the classroom has been advocated now for over a decade, but adoption has been slow. One of the reason is that it requires a significant shift in thinking about teaching and learning for many teachers from knowledge instruction to knowledge construction. Teachers who support this view recognize the importance of the active involvement of their students in learning and the need for a learning environment that encourages students' independent exploration of ideas. Smith, Clark and Blomeyer (2005) see the greatest benefits in "constructivist approaches that use interaction within a situational context to encourage learners to think and reflect while constructing their own personal meaning" (p. 11). However, teachers need to remember that the technology does not teach students, but rather the students only learn when they construct their own knowledge and think and learn through their experience. The computer is simply a tool that can assist students in their knowledge construction.

Technology use that is shaped by constructivist learning principles supports a more student centered, **inquiry** oriented approach to teaching. What is needed in classrooms are technology uses that help students to build knowledge and develop higher order thinking and problem solving skills by providing opportunities for them to think critically and analytically about information and represent their new understandings in multiple ways in an engaged setting (Marlow & Page, 2005). According to active learning principles, which emphasize **constructivism**, students must engage in researching, reasoning, critical thinking, decision making, analysis and synthesis during construction of their knowledge" (Tarhan, Ayar-Kayali, Urek & Acar, 2007, p. 286). Ferdig (2006) identifies five components of a social constructivist innovation design: authentic, interesting and challenging academic content; a sense of ownership by the learner; active participation, **collaboration** and social interaction; opportunities for creation of artifacts in a variety of ways; and publication, reflection and feedback (p. 750). Teachers also need to recognize that the four classroom walls no longer bind learning. "When children collaborate, they can and do scaffold each others' thinking" (Ferdig, 2006, p. 751). Every classroom has the potential to be a global learning environment. In this way, computer technologies can help to bridge the gap between the artificial world of school and the outside lives of young people by engaging then in projects that investigate real world issues, that draw on multiple perspectives and that encourage **collaboration** with experts and other students from around the world. Such global collaborative activities with peers in classrooms around the world can help to promote understanding and appreciation of multiple perspectives and encourage students to become global thinkers (Boss & Krauss, 2007).

Thus, constructivist uses of computer technologies need to provide learning opportunities that are based on authentic tasks and environments and include opportunities for exploring and doing as well as for feedback and reflection. These learning environments should be learning spaces in which students have control over the learning activities and are able to use a variety of information resources and tools to solve problems. The i**nquiry** should begin with students' prior background knowledge and experience, and engage them in creatively applying the resultant new knowledge. This learning environment should represent as much as possible the complex real world of problem solving, however, students need to be taught the skills to work in such environments. This is where a more structured type of learning environment such as **problem based learning** can provide initial assistance in developing the requisite skills by providing a guided process.

What is problem based learning and how it is an example of a learning environment based on constructivist learning principles?

In order to prepare students for today's' complex world, some schools engage students in problem-based learning... trying to hone the students' skills in applying what they learn to the kinds of problems they are likely to face. (Sternberg, 2008, p. 14)

Problem-based learning (PBL) is an instructional model that exemplifies constructivist learning principles (Ochoa & Robinson, 2005). One of the main characteristics of problem-based learning is situating the learning in the examination of authentic, real-life problems and questions of relevance to the learner in order to engage them in the learning. Rather than 'teaching' the student in the sense of presenting or even assigning information, the goal is encourage student driven **inquiry** in which they activate their prior knowledge and investigate the problem from a number of different perspectives in order to develop equally viable alternative solutions to the problem (Ochoa & Robinson, 2005). "Teachers who value thinking and habits of mind would ensure that students confront the problem with a questioning attitude, arm themselves with attendant data, explore alternatives to the status quo, and predict the consequences of each of those alternatives" (Costa, 2008, p. 21). Learning abstract ideas in this way becomes more concrete and realistic for students (Frazier & Sterling, 2008). Effective PBL environments also involve communication and c**ollaboration** that require students to articulate their ideas in ways that strengthen and assist the knowledge construction process as well as activities that encourage the learners to reflect on their learning. Organizing content around significant questions or problems can also assist students in developing higher order thinking skills, flexible understanding and lifelong skills (Ruiz, 2008).

The teacher in PBL does not teach the students what they should do or know and when they should do it or go about learning it. Rather the teacher is there to support the students in developing their critical thinking skills, self-directed learning skills, and content knowledge in relation to the problem. The teacher should acknowledge and support the students' thinking rather than impose structure on it but should also provide experiences that challenge that thinking. Initially the teacher needs to determine what the key concepts and procedures are that the child needs to know and then design a learning experience that requires students to use the that information in authentic tasks. Scaffolds also need to be built in to help students to organize and represent what they know as to provide the teacher with opportunities to probe students' knowledge and thinking skills. The **inquiry** needs to focus on using information as a means to develop information-processing skills and problem solving skills.

Computer technologies can be effective vehicles for introducing problems for student investigation because they "allow students to experience a shared context in which they engage in sustained thinking about complex problems and engage in interpretive learning experiences" (Hmelo-Sivler, 2004). **WebQuest** and Web Inquiry Projects are two examples of how online learning environments that are problem based can be designed.

HOW IS A WEBQUEST AN EXAMPLE OF A PROBLEM-BASED LEARNING APPROACH TO TECHNOLOGY USE?

A **WebQuest** is one example of how to design Internet-based learning experiences that promote **digital literacy** as well as the development of essential **higher level thinking**, problem solving and communication skills. There is a growing body of literature on the value of **WebQuests** as an instructional approach to integrate structured **inquiry** and the use of technology (Hicks, Sears, Gao, Goodmans & Manning, 2004). A **WebQuest** is an inquiry-oriented activity in which most or all of the information used by learners is drawn from the Web (Dodge, 2005). **WebQuests** are designed to efficiently use learners' time, to focus on using information rather than looking for it, and to support learners' thinking at the levels of analysis, synthesis and evaluation (March 2004; Dodge, 2005). Such a use of the Internet supports a view of students as active creators and shapers of their own knowledge who are able and willing to think for themselves. Through a **WebQuest,** students can actively explore issues and problems from a number of different perspectives, as well as searching for solutions and making moral and ethical decisions about real contemporary world problems. In an authentic **WebQuest** there is no single correct answer. While engaged in the **inquiry** through a **WebQuest,** students are constructing their own personal meaning about the problem under investigation.

The rationale for using a structured **inquiry** approach such as a **WebQuest** design can be traced back to Bruner's cognitive development theory. For Bruner, the most important outcome of cognitive development is thinking and the process that students undergo to acquire knowledge, not the product (Bruner, 1966). Bruner's discovery learning and **inquiry** teaching methods envision the learners creating their knowledge by "rearranging or transforming evidence in such a way that one is enabled to go beyond the evidence so assembled to additional new insights" (Bruner, 1961, p. 22). This requires an activity structure that scaffolds learners' experience so that they must move beyond simply finding information to using that information to think through and resolve a problem or issue. Also, the question posed to students cannot be answered simply by collecting and spitting back information. A well designed **WebQuest** requires students to transform information into something else. Some of the thinking skills analogous with **WebQuests** are "comparing, classifying, inducing, deducing, analyzing, constructing, abstracting, and analyzing perspectives" (Norton & Wiburg, 2003, p. 180).

WebQuests have been used successfully to develop subject specific content in middle and high school social studies (Hung 2004; Leite, McNulty & Brooks, 2005; Lipscomb, 2003; Stickland, 2005); creativity in art (Kundu & Bain, 2006), conceptual understanding in elementary math

(Orme & Monroe, 2005); thinking skills (Murry, 2006), with at risk students (Wilson, 2006); and ESL students (Goodwin-Jones, 2004); and for developing reading and literacy skills (Ikpeze & Boyd, 2007). "Using a WebQuest can help to bring reading alive, address essential questions that bring meaning to learning, engage in information processing, problem solving, **collaboration**, alleviate the concern of how to address reading needs of all students, taking on role- meet individual needs and differing learning styles" (Teclehaimanot & Lamb, 2004).

WebQuests can also enhance students' communication skills as many involve working in cooperative groups and role-playing. Working either independently or in groups, the students explore an issue or problem in a guided and meaningful manner. Some **WebQuests** have the students take on roles that help to make the group work together more efficiently and effectively. These roles can include a group leader, recorder, communicator, encourager and evaluator, among others. Other **WebQuests** have the learners assume the roles of particular players in a role-playing setting where they access, analysis and synthesize the information provided from the perspective of that player.

The most authentic **WebQuests** engage students in perspective taking on a particular problem or issue. Students investigate the context and the issue from an individual's perspective in order to build a better understanding of the person, the event and the setting. The goal is for students to use the information collected to construct an argument based on evidence. They then publicly share their findings with the class and the class tries to come to some kind of resolution to the problem under investigation. This resolution may mean arriving at class consensus or if there is a conflict of resolutions, then agreeing to disagree. Role-playing can be particularly beneficial for teaching students the importance of perspective taking when problem solving. Here is where **WebQuests** have the greatest potential for addressing the multicultural literacy aspect of **digital literacy**. Investigating problems from a number of different cultural perspectives can help learners to better understand the wide diversity of views on any one issue as well as the important cultural foundations of those views. This can lead to learning to respect and appreciate diversity.

WebQuests and **problem based learning** fit well together as they both address constructivist learning principles, critical thinking, **scaffolding**, learner motivation, cooperative learning and authentic assessment (Levine, 2002). Both **WebQuests** and problem based learning encourage **higher level thinking** including analysis, critical thinking and creative thinking; both include a introduction that sets the stage and provides some background information; both place students in a scenario where they must solve a fuzzy problem; both actively engage students in the learning and empower them to determine the outcome; both have no one right answer; both require compiling information from a variety of sources in order to arrive at a solution; and, both use authentic assessment strategies such as rubrics. Where these two strategies differ is in how structured they are and in who imposes that structure. **Problem based learning** is less structured than a **WebQuest** and provides the students with a larger decision making role in terms of defining the problem to be investigated, setting the conditions for resolution, determining strategies for addressing the problem, deciding on the roles to be taken and the end product of the investigation, and in selecting the resources to be used.

WHAT ARE THE COMPONENTS OF A WEBQUEST?

Usually a **WebQuest** consists of the introduction, task, process, resources, evaluation and conclusion. The first part, the introduction, lays out the task or the problem to be investigated, provides some background information and acts as a motivator to get the students interested in the activity. The task outlines the overall challenge the students will be engaged in and explains what they will be doing to represent what they have learned from completing the **WebQuest**. The task also provides the focus questions that frame the investigation and facilitate the learning process. The process provides a description of what needs to be done in order to accomplish the task in a step-by-step fashion. Here, students are usually assigned roles or provided with differing perspectives on the issue or problem being investigated. The resource section provides information sources that are needed for solving the task. Most of the resources used for the **inquiry** are other Websites that have been vetted by the teacher and linked directly to the **WebQuest**. Many **WebQuests p**rovide direct access to individual experts, current news sites and searchable databases for information sources. The evaluation section provides information for students on how they will be assessed. The assessment tool often included is a rubric for providing feedback on the outcome of the i**nquiry**. Other formative types of assessment can be used throughout the **inquiry** including personal reflective logs, skills checklists, and self and group feedback on the effectiveness of their group work. The conclusion brings closure to the **WebQuest** by reviewing and summarizing the learning from the experience and often challenges learners to extend their learning in new ways.

WebQuests can be either short term, on the average one to three classes, with the goal of acquiring and making sense of new information or longer term in which a student analyzes a body of information, transforms it in some way and demonstrates an understanding of that information in a public way. Longer-term **WebQuests** can take anywhere from a week to a month (Norton & Wiburg, 2003). Throughout the **WebQues**t, the teacher acts as the facilitator checking to see that students understand the role that they are to take and that they stay on task.

WHERE CAN WEBQUESTS BE FOUND?

A **WebQuest** can be chosen from a series of pre-designed **WebQuest** collections [see <http://Webquest.org/index.php> or < http://www.kn.att.com/wired/fil/tips/Webquest_instructions.html >] or one can be created by the teacher to address a specific topic of study. The latter allows for more active student involvement in deciding what problem they might like to investigate and in designing an interesting and relevant learning experience around that problem. An example of a pre-designed **WebQuest** from one of the databases mentioned earlier entitled "Does the Tiger Eat its Cubs" [http://www.kn.pacbell.com/wired/China/childquest.html] explores the way children in orphanages in China are treated.

In this **WebQuest,** students investigate the question "What's the truth about how children are treated in China?" They are directed to investigate the question from a number of perspectives. They are divided into three teams. One team reads international news reports, another reads responses from the Chinese people and a third examines the government of China's position as stated in China's One Child Policy. The class then comes back together and discusses their findings with the challenge of arriving at consensus decision on the issue. The culminating activity is to write a letter to the government expressing their opinion on what they feel should be done about the situation.

In the **WebQuest**, Children of conflict [http://www.accessola.com/osla/bethechange/Webquest/conflict/index.html], students work in teams as part of a task force to investigate how conflict affects children in different parts of the world. After researching their particular areas, the groups come back together to present their recommendations to a special parliamentary committee to decide what Canada should do to help to protect children around the world.

DNA for Dinner **WebQuest** [http://dnafordinner.blogspot.com/] engages students in an **inquiry** about the issue of genetically altered food. The issue to be investigated is "Should genetically engineered food crops be specifically labeled for consumers and why." Using the resources provided, students are to read up on the issue and then draft a law based on their investigation. They are then encouraged to email a representative in the federal government detailing their investigation and their concern over the issue and explaining their proposed solution. A **WebQuest s**uch as this one is an example of how the learning activity can be designed to increase students' motivation to want to learn by connecting what is learned in school to real world experiences.

WebQuests can also be a powerful way for students to be immersed in historical events and to have the opportunity to work with historical documents. In the Scrooge for Mayor **WebQuest** [http://www.coollessons.org/Dickens.htm], students work in teams to develop a campaign proposal for Scrooge using information about labour, education, industrialization and quality of life issues in nineteenth century England as represented in Charles Dicken's work of fiction "A Christmas Carol". Each campaign team is made up of a team manager, research analyst, public relations person and political strategist. Students are directed to focus on how Scrooge's viewpoint on daily life in London will need to change and what solutions to London's problems and programs he will need to support in his run for mayor. Each person on the campaign team is responsible for writing an article for a newspaper describing what they found out including what life was like in the area in the 1840's, the conditions that made it necessary to bring about change, what changes were proposed and how those changes would better things as well as an editorial on the topic "Is the industrial revolution a good thing?" The team also is directed to create a campaign poster, a pamphlet and a PowerPoint presentation that are to be used to communicate their ideas to Scrooge. This **WebQuest i**s an excellent example of how **WebQuests** can be used to integrate various subject areas in meaningful ways. It could be used to address the learning outcomes of social studies, reading, language arts and science.

In the Ancient Egypt **WebQuest** [http://www.iWebquest.com/egypt/ancientegypt.htm] students take on a series of missions to learn about King Tut, early Egyptian daily life, and the study of archeology.

Using the Middle Ages Storytelling Quest [http://www.iWebquest.com/middleages/Default.htm], students learn about the history of the Middle Ages then create their own story to teach their peers what they have learned about this historical time period.

Some WebQuests encourage students to take on cooperative learning roles to make their group work more efficient. The Big Wide World **WebQuest** [http://www.kn.pacbell.com/wired/bww/index.html] is an example of one that combines cooperative roles and focus topics to engage primary students in an investigation about their world. The "A "No-bullying Proposal" **WebQuest** [http://www.gecdsb.on.ca/d&g/nobullying/index.html] involves children in role taking from different perspectives on the issue of bullying. The groups then come up with a plan for how to address bullying in their school.

As well as selecting from thousands of predesigned **WebQuests**, teachers can design a **WebQues**t to meet their own personal needs using available templates [see for example < http://Webquest.sdsu.edu/LessonTemplate.html.]. Students can also be encouraged to try developing their own WebQuests and sharing them with classmates. A database of sample student developed WebQuests can be found at the ThinkQuest Library site [http://www.thinkquest.org]. Having students create their own WebQuests challenges them "to explore a topic, summarize what the most important events or facts are in relation to the topic, and then put together the links and questions or other students to follow" (Whitworth &

Berson, 2003, p. 480). When students engage in creating their own WebQuests, it can also enhance the development of their critical, creative and **higher level thinking** skills. The two Websites noted previously provide templates that students can use for creating their own quests.

WHAT ARE SOME OF THE LIMITATIONS OF WEBQUESTS?

The **WebQuest** approach is intended to capitalize on the possibilities provided by the Internet for guided **inquiry** learning while eliminating some of the disadvantages such as time wasted looking for resources, learners accessing inappropriate resources, and the lack of sufficient experience with the research process (Milson, 2002). There are some limitations to using WebQuests, however, that teachers need to be aware of. Maddux and Cummings (2007) caution that "simply because a lesson is cast in a **WebQuest** format is no guarantee that the lesson makes use of cooperative learning, advanced organizers, **scaffolding, problem-based learning**, nor does it guarantee that these concepts and techniques are effectively, or even merely competently, applied in a way that is consistent with the huge literature base underlying each of them" (p. 121). One problem is that not all **WebQuests** encourage **higher order thinking** and must be carefully scrutinized in order to assess how well they accomplish this. Many **WebQuests** are merely designed as fact-finding exercises that do little to engage students in problem solving. No attempt is made to engage students in role taking or learning to view problems from multiple perspectives. Fewer still actually engage students in learning the important problem solving skills of conflict resolution, compromising or agreeing to disagree. Others lack clear direction to the user that can detract from the ability of students to take control of the learning experience.

There are a number of Websites that provide rubrics for determining the quality of WebQuests. [See for example, <http://bestWebquests.com/bwq/matrix.asp> and <http://Webquest.sdsu.edu/Webquestrubric.html >]. The criteria included in these assessments are: engaging opener; clear question and tasks; learner roles match the issues and resources; **higher level thinking** built in; opportunities for feedback provided; and a conclusion that ties in to the introduction, makes the students' cognitive tasks overt and suggests how this learning could transfer to other domains/issues.

Another limitation of **WebQuests** is that students are most often removed from the process of selecting resources on which to base their investigation. There is now more information than teachers, textbooks and the curriculum can dispense. Consequently, students need to learn the skills to become information managers themselves. New computer technologies have much to offer teachers and students in terms of enhancing their information access, use and evaluation skills to encourage more effective and thoughtful consumption of information. As current information becomes easily accessible online, it is increasingly important that students have the opportunity to develop their critical analysis capabilities (Mason, Alibrandi, Berson Diem, Dralle, Hicks, Keiper & Lee, 2000). Also educators are warned not to simply rely on Internet filtering software but rather to focus on teaching students critical thinking skills so that they can learn to make informed decisions and judgments about the information they encounter on the Internet (Whitworth & Berson, 2003, p. 480). The use of such filtering tools can also be a problem as many sites that would be relevant to the study of a topic, such as war and conflict, would be inaccessible to students.

Locating useful and accurate information on the Web can be a struggle for students. The abundance of things to access via the Internet can cause students to be easily side tracked and spend a great deal of time off task. Information gathering can easily become a mindless exercise in which quantity overrides quality. This sort of information-gathering exercise does little to promote deeper thinking and understanding. Students need to be instructed in and have opportunities to practice how to critically examine and make informed choices about the information they are accessing. Critical information literacy skills need to be carefully taught and monitored to ensure students are developing proficiency in their use. In addition to learning the skills of locating and evaluating information on the Web, students also need to learn how to select relevant pieces of information and synthesize and organize it in order to apply it to the learning activity and communicate it to others.

Because there is an inclination to accept the computer as an authority and view the information accessed as the "truth," students need to be taught to recognize that the information on the Web represents a particular viewpoint, as does any other resource. They need to be encouraged to conscientiously use critical thinking skills to make both appropriate and ethical choices when using computer-generated information. Students need to be taught how to apply the skills of actively interpreting the information provided, drawing conclusions from data, seeing several points of view, distinguishing fact from opinion, and finding meaning in information, as they interact with digital technologies. In order to develop students' critical thinking skills, they should be taught to look for authorship/source, objectivity/biases, and validity of content, bibliography/reference links, currency and quality of writing. Questions such as the following can be helpful for students and teachers to use in judging the effectiveness of Websites:

- Where did this document come from and how reliable a source is it?
- Is the information presented objectively or with an obvious bias?
- How current is the information?
- How comprehensive is the coverage of the topic on the Website?
- How trustworthy is the data provided and how accurately does it depict the phenomenon?
- Does the site deepen my understanding of the topic?
- How useful is the site to me in assisting with the **inquiry**?

Critical literacy skills need to be carefully taught and monitored to ensure students are developing proficiency in their use. Children need to be instructed in and have opportunities to practice how to critically examine and make appropriate, ethical and informed choices about the information they are accessing. They need to be taught to recognize that the information on any Website represents a particular viewpoint and that it is important to examine several points of view on any issue. They also need to be taught how to distinguish fact from opinion.

A third limitation is that **WebQuests** lead students through a **scaffolded inquiry** experience that specifies the task, the roles and perspectives to be taken, the resources to be used and the guides for organizing the learning with little opportunity for the students to set the direction and plan for the investigation. Being heavily **scaffolded**, **WebQuests** prevent learners from participating in higher-level **inquiry** activities (Molebash, Dodge, Bell, Mason & Irving, n.d.). While these initial scaffolds are very important for helping children to develop problem solving strategies, there needs to be opportunities for releasing some of the control into the hands of the learners. Molebash and Dodge (2003) note that the support of the **WebQuest** can be removed in stages by allowing more flexibility in how and what student are to produce in the task, by gradually providing fewer URLs and expecting the learner to find more, by gradually removing the **scaffolding** such as note taking guidelines, information organizing structures, writing prompts, etc., and by putting more resources in the conclusion for learners to explore on their own later.

WHAT ARE WEB INQUIRY PROJECTS?

In order to promote higher levels of **inquiry** in the classroom, less specific guidance can be given to students. **Web Inquiry Projects** (WIPs) are one example of a way to extend the **WebQuest** idea beyond structured **inquiry** to more open **inquiry** that promotes **higher levels of thinking** and student engagement. **Web Inquiry Projects** are "open **inquiry** learning activities that leverage the use of uninterpreted [primary source] online data and information" (Molebash, 2004, p. 2). Unlike **WebQuests**, which provide students with a procedure and the online resources needed to complete a predefined task, **WIPS** place more emphasis in having students determine their own task, define their own procedures, and play a role in finding the needed online resources. More often the **inquiry** is sparked by the interest of the students. The teacher's role is to "insert the necessary **scaffolding** at each stage in the process to ensure that students are successful" (Molebash 2004, p. 2). According to Molebash, **WIPS** have seven stages: a hook to capture students' interest, question generating, deciding on procedures for guiding the investigation, data investigation of possible online sources, analysis of data, findings reporting including drawing conclusions based on the evidence, and lastly the generation on new questions resulting from the investigation to encourage further i**nquiry**.

Numerous examples of **Web Inquiry Projects** can be viewed at http://edWeb.sdsu.edu/wip/examples.htm. In the **WIP** entitled "The AIDS Epidemic: Can It Be Stopped?" [http://edWeb.sdsu.edu/wip/examples/aids/index.htm], for example, students are presented with the following hook:

The HIV/AIDS Epidemic is still occurring today. Currently medical research in finding a cure for AIDS have not progressed beyond prolonging HIV before it turns into AIDS. Although we don't see HIV/AIDS in the news today, it is still a problem around the world. Many people feel that they are not at risk for contracting this disease, but it is important for individuals to realize that they may be at risk. The first case of HIV/AIDS was diagnosed in the United States in the early 1980's. When will the last case be diagnosed?

In order to address this challenge, students need to determine what investigative tools to use, what types of data they will need and how they will manipulate that data in order to predict an answer. As a part of their investigation they also conduct detailed research on AIDs in order to increase their understanding of the issues surrounding AIDs and HIV.

In another example, North American Perspectives [http://eprentice.sdsu.edu/F034/sjohnson/teacher_template2.html], students are hooked into the i**nquiry** through a series of questions that they are to answer initially from their own perspective then from "behind Native American eyes". They are encouraged to think of some questions related to this topic that they might like to investigate as well as being provided some teacher-initiated ones. There are some pre-selected resources provided but students are encouraged to locate their own as well. Some ideas for how to re-present their learning are made available but once again students are encouraged to come up with their own ideas too. Each of these examples allows for a greater degree of student control over the learning experiences.

CONCLUDING REMARKS

Attention to **digital literacy** has become an essential aspect of children's education for the 21st century. This chapter began by defining **digital literacy** as well as highlighting other important 21st century skills including **higher level thinking**, problem solving, communication and **collaboration**. Included under the umbrella of digital literacy are such skills as understanding how to operate a particular technology, knowing how and why technologies can be used, and recognizing the ramifications of their use. The research on best uses of technology for learning has identified a number of effective ways for infusing **digital literacy** skills and other technology outcomes throughout a child's educational experiences. Emerging from this review of the research is an acknowledgment of the learning theory of **constructivism** as a way of framing learning experiences with computer technologies. **WebQuests** and their extension, **Web Inquiry Projects**, are two approaches that have the potential to effectively model constructivist learning principles while also addressing **digital literacy**, thinking, problem solving and communication skills. What makes these approaches to technology use in schools most effective is the emphasis on student directed learning and active student engagement. The level of student control over the decision making about the learning varies from a lesser degree in the more structured **inquiry** usually found in **WebQuests** to a greater degree in the open **inquiry** of **Web Inquiry Projects**. Other essential features of effective technology use found in both **WebQuests** and **Web Inquiry Projects** that were identified were: a) **problem based learning** focused on real world authentic issues and questions of interest to students and, in the case of **WIPs**, generated by the students; b) a focus on **collaborative** learning both within and beyond the classroom walls; and, c) an emphasis on learning to manage information and to work with that information at a **higher level of thinking** and understanding. All of these features support the call for learning experiences that attend to **digital literacy** and to developing the thinking, problem solving and communication skills of today's learners.

REFERENCES

Balanskat, A., Blamire, R., & Kefala, S. (Dec. 11, 2006). The ICT impact report: A review of studies of ICT impact on schools in Europe. *European Schoolnet*. Retrieved from http://ec.europa.eu/education/doc/reports/doc/ictimpact.pdf

Boss, S., & Krauss, J. (2007). Real projects in a digital world. *Principal Leadership*, *8*(4), 22–26.

Brooks, J. (2003). *Schooling for life: Reclaiming the essence of learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

Brooks, J., & Brooks, M. (2001). *In search of understanding: The case for constructivist classrooms*. Upper Saddle River, NJ: Prentice-Hall.

Brumfield, R. (April 4, 2006). Thornburg: Ed tech stalled by 'fear'. *eSchool News*. Retrieved from http://www.eschoolnews.com/news/top-news/index.cfm?i=36912&CFID=906452&CFTOKEN=45974173

Bruner, J. (1961). The act of discovery. *Harvard Educational Review*, *31*, 21–32.

Bruner, J. (1966). *Toward a theory of instruction*. Cambridge, MA: Belknap.

Costa, A. (2008). The thought-filled curriculum. *Educational Leadership*, *65*(5), 20–24.

Cuban, L. (2001). *Oversold and underused: Computers in the classroom*. Cambridge, MA: Harvard University Press.

Dodge, B. (1996). *Active learning on the Web*. Retrieved May 23, 2008 from http://edWeb.sdsu.edu/people/bdodge/active/ActiveLearningk-12.html

Doolittle, P., & Hicks, D. (2003). Constructivism as a theoretical foundation for the use of technology in social studies. *Theory and Research in Social Education*, *31*(1), 72–104.

Dunn, R. (2007). Integrating technology into pedagogical practice. *Digital Commons*. Available at http://digitalcommons.liberty.edu/educ_fac_pubs/60

Ferdig, R. (2006). Assessing technologies for teaching and learning: understanding the importance of technological pedagogical content knowledge. *British Journal of Educational Technology*, *37*(5), 749–760. doi:10.1111/j.1467-8535.2006.00559.x

Frazier, W., & Sterling, D. (2008). Motor mania: Revving up for technological design. *Technology Teacher*, (February): 5–12.

Godwin-Jones, R. (2004). Emerging technologies: Language in action from WebQuests to Virtual Realities. *Language Learning & Technology*, *8*(3), 9–14.

Hartley, J. (2007). Teaching, learning and new technology: A review for teachers. *British Journal of Educational Technology*, *38*(1), 42–62. doi:10.1111/j.1467-8535.2006.00634.x

Hicks, D., Sears, P., Gao, H., Goodmans, P., & Manning, J. (2004). Preparing tomorrow's teachers to be socially and ethically aware producers and consumers of interactive technologies. *Contemporary Issues in Technology & Teacher Education*, *3*(4), 470–481.

Hmelo-Silver, C. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review*, *16*(3), 235–266. doi:10.1023/B:EDPR.0000034022.16470.f3

Hung, C. C. (2004). *The use of WebQuests as a constructivist learning tool in secondary school geography in Singapore*. Paper presented at ISTE Conference. Retrieved May 9, 2008 from http://www.iste.org/Content/NavigationMenu/Research/ NECC_Research_Paper_Archives/NECC_2004/Chang-Chew-Hung-NECC04.pdf

Ikpeze, C., & Boyd, F. (2007). Web-based inquiry learning: Facilitating thoughtful literacy with WebQuests. *The Reading Teacher*, *60*(7), 644–654. doi:10.1598/RT.60.7.5

ISTE. (2007). *National Educational Technology Standards*. Retrieved from http://www.iste.org/AM/Template.cfm?Section=NETS

Jonassen, D., Howland, J., Moore, J., & Marra, M. (2003). *Learning to solve problems with technology*. Upper Saddle River, NJ: Merrill Prentice Hall.

Jones-Kavalier, B., & Flannigan, S. (2006). Connecting the digital dots: Literacy of the 21st century. *EDUCAUSE Quarterly*, *2*, 8–10.

Kundu, R., & Bain, C. (2006). Utilizing technology in a constructivist manner to facilitate meaningful preservice learning. *Art Education*, *59*, 2, 6–12.

Lee, R. (2006). Effective learning outcomes of ESL elementary and secondary students utilizing educational technology infused with constructivist pedagogy. *International Journal of Instructional Media*, *33*(1), 87–93.

Leite, M., McNulty, A., & Brooks, D. (2005). Learning from WebQuests. *International Society for Technology in Education Research Paper*. Retrieved May 8, 2008 from http://www.iste.org/Content/NavigationMenu/Research/NECC_Research_Paper_Archives/NECC_2005/Leite-Martonia-NECC05.pdf

Levine, R. (2002). Comparing problem based learning and WebQuests. Retrieved May 20, 2008 from http://www.coollessons.org/compare.htm

Lipscomb, G. (2003). "I guess it was pretty fun": Using WebQuests in the middle school classroom. *Clearing House (Menasha, Wis.)*, *76*(3), 152–155.

Mackenzie, J. (2005). Singular displeasure: Technology, literacy and semantic power plays. *From Now On: The Educational Technology Journal* 14(5) Retrieved from http://fno.org/jun05/singular.html

Maddux, C., & Cummings, R. (2007). WebQuests: Are they developmentally appropriate? *The Educational Forum*, *71*(2), 117–127. doi:10.1080/00131720708984925

March, T. (2004). The learning power of WebQuests. *Educational Leadership*, December / January, 42–47.

Marlow, B., & Page, M. (2005). *Creating and sustaining a constructivist classroom*, (2nd Ed). Thousand Oaks California: Corwin Press.

Mason, C., Alibrandi, M., Berson, M., Diem, R., Dralle, T., Hicks, D., et al. (2000). Waking the sleeping giant: Social studies teacher educators collaborate to integrate technology into methods' courses. *Society for Information Technology and Teacher Education International (SITE) Conference*, 2000(1), 1985-1989.

Milson, A. J. (2002). The Internet and inquiry learning: Integrating medium and method in a sixth grade social studies classroom. *Theory and Research in Social Education*, *30*(3), 330–353.

Molebash, P. (2002). *Web Inquiry Projects*. Retrieved from http://Webinquiry.org/

Molebash, P. (2004). Web Historical Inquiry Projects. *Social Education*, *68*(3), 226–230.

Molebash, P. (2005). Web Inquiry Projects: A Paper Submitted as Part of the Symposium "Multimedia, Historical Inquiry and Preservice Teacher Education". In C. Crawford et al. (Eds.), *Proceedings of Society for Information Technology and Teacher Education International Conference 2005* (pp. 3854-3855). Chesapeake, VA: AACE.

Molebash, P., & Dodge, B. (2003). Kickstarting inquiry with WebQuests and Web Inquiry Projects. *Social Education*, *67*(3), 158–162.

Molebash, P., Dodge, B., Bell, R., Mason, C., & Irving, K. (n.d.). *Promoting student inquiry: WebQuests to Web inquiry projects (WIPs)*. Retrieved from http://edWeb.sdsu.edu/wip/overview.htm

Murry, R. (2006). WebQuests celebrate 10 years: Have they delivered? *Action Research Exchange*. Retrieved from http://teach.valdosta.edu/are/vol5no1/Thesis%20PDF/MurryR_ARE.pdf

North Central Regional Educational Laboratory. (2005). *Critical issue: Technology - A catalyst for teaching and learning in the classroom*. Retrieved from http://www.ncrel.org/sdrs/areas/issues/methods/technlgy/te600.htm

Norton, P., & Wiburg, K. (2003). *Teaching with technology: Designing opportunities to learn.* Belmont CA: Thomson Wadsworth.

Ochoa, T., & Robinson, J. (2005). Revisiting group consensus: Collaborative learning dynamics during a problem-based learning activity in Education. *Teacher Education and Special Education*, *28*(1), 10–10. doi:10.1177/088840640502800102

Oppenheimer, T. (2003). *The flickering mind: The false promise of technology in the classroom and how learning can be saved*. New York: Random House

Orne, M., & Monroe, E. (2005). The nature of discourse as students collaborate on a mathematics WebQuest. *Computers in Schools*, *22*(1/2), 135-146. Partnership for 21st century skills. (2004). *Framework for 21st century learning*. Retrieved from http://www.21stcenturyskills.org/ index.php?option=com_content&task=view&id=254 &Itemid=120

Ruiz, E. (2008). Problem-based learning: A pedagogical strategy for active learning. *Community College Journal of Research and Practice*, *32*(3), 251–257. doi:10.1080/10668920701836349

Sterling, D. (2007, December). Modeling problem-based instruction. *Science and Children*, *4*, 50–53.

Sternberg, R. (2008). Interdisciplinary problem-based learning. *Liberal Education*, (Winter): 12–17.

Strickland, J. (2005). Using WebQuests to teach content: Comparing instructional strategies. *Contemporary Issues in Technology & Teacher Education*, *5*(2), 138–148.

Tarhan, L., Ayar-Kayali, H., Urek, R., & Acar, B. (2008). Problem-based learning in 9th grade chemistry class. *Research in Science Education*, *38*, 285–300. doi:10.1007/s11165-007-9050-0

Teclehaimanot, B., & Lamb, A. (2004, March/April). Reading, technology, and inquiry-based learning through literature-rich WebQuests. *Reading Online*, *7*(4). Retrieved from http://www.readingonline.org/articles/art_index.asp?HREF=teclehaimanot /index.html

Viadero, D. (2007, March 29). Collecting evidence. *Education Week*, *26*(30), 30, 32–33.

Wan, Z., Fand, Y., & Neufeld, D. (2007). The role of information technology in technology-mediated learning: A review of the past for the future. *Journal of Information Systems Education*, *18*(2), 183–192.

Whitworth, S., & Berson, M. (2003). Computer technology in the social studies: An examination of the effectiveness literature (1996-2001). *Contemporary Issues in Technology & Teacher Education*, *2*(4), 472–509.

Windschitl, M. (2002). Framing constructivism in practice as the negotiation of dilemmas: An analysis of the conceptual, pedagogical, cultural, and political challenges facing teachers. *Review of Educational Research*, *72*(2), 131–175. doi:10.3102/00346543072002131

Wiske, M., Franz, K., & Breit, L. (2005). *Teaching for understanding with technology.* San Francisco, CA: Jossey Bass.

KEY TERMS AND DEFINITIONS

Digital Literacy: The skills of information location and application including understanding how to use available evidence to assist in problem solving and decision making.

Constructivist Learning Theory: A learning theory that acknowledges the learner as the holder and creator of their own knowledge.

Inquiry: An approach to learning that directly engages learners in constructing their own knowledge and understanding.

Problem Based Learning: An approach to learning in which learners inquire into problems about important questions and issues that have no clear answers.

WebQuest: A Web-based structured inquiry approach to learning.

Web Inquiry Project: A Web-based open inquiry approach to learning.

This work was previously published in Handbook of Research on New Media Literacy at the K-12 Level: Issues and Challenges , edited by L.T.W. Hin; R. Subramaniam, pp. 403-418 , copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.21
The EduOntoWiki Project for Supporting Social, Educational, and Knowledge Construction Processes with Semantic Web Paradigm

Corrado Petrucco
University of Padua, Italy

INTRODUCTION

The Web is going to produce a revolution in learning and teaching: the debate on the role of ICT in educational processes leads to a reconsideration of how we deal with information and knowledge. The widespread use in educational contexts is also due to the ease with which learning resources can be retrieved and shared: for example, the recent introduction of learning objects means that the contents which reside in different e-learning platforms is easy to find and access. But knowledge is also deeply embedded in millions of Web pages. Nonetheless, searching for information on the Web is not a simple task and the great number of documents found using search engines, such as Google, is beyond the human cognitive capacity to deal with this information overflow. Teaching information literacy skills or stimulating collaborative information filtering that supports the discovery of resources in a way that is responsive to the context of users may help, but there is a need for more efficient cognitive tools to search, organize, and discuss information in order to codify it in shared knowledge structures.

In a more and more complex world we need support to think at a high level so the technologies let us develop strong knowledge structures that do not have the representational problems of the old schemas. An attempt in this direction is the Semantic Web: if we succeed in making the Semantic Web available and useful for education, it could revolutionize the way we think about teaching and learning with ICT. Our current research is aimed at the development, experimentation and evolution of an integrated

DOI: 10.4018/978-1-59904-845-1.ch026

learning environment called EduOntoWiki that is backed up by a semantic structure based on the active consent of communities of practice.

BACKGROUND

Current research suggest that it is not correct to assume that the introduction of ICT necessarily changes the way students learn. We have to acknowledge that the teacher plays a critical pedagogical role in creating the conditions for technology-supported learning through selecting and evaluating appropriate technological resources and designing learning activities (Galliani, Costa, Amplatz, & Varisco, 1999). We can distinguish between two approaches to ICT: a technology-centered approach and a learner-centered approach (Mayer, 2005). The former generally fails to lead to lasting improvements in education: looking back at the many predicted educational revolutions, in which the current "new" technology (radio, television, computer, multimedia, the Web) would have been the "killer" application for teaching and learning processes, we see that they failed to materialize (Cuban, 1986) and so was the case with the claims and worries, during the 1960s, that computers-as-tutors would replace teachers (Cognition and Technology Group at Vanderbilt, 1996). A learner-centered approach can, on the other hand, help students and teacher to learn and teach through the aid of technology with a focus on how ICT can be used as an aid to human cognition and consistent with the way the mind works solving complex tasks and dealing with today's information overflow. The quantity and kind of information students today need to assess has expanded exponentially in the last few years, due mainly to the World Wide Web and improvements in the capabilities of search engines. In this context, it is important to consider both student and teacher roles using a constructivist approach that can stimulate collaborative formalization and knowledge building.

SEMANTIC WEB AND ONTOLOGIES

The Web has arrived at an important epistemological crossroad and there is a need to integrate the current dialogic-informative model, which allows us to interact with people and search for documents on the Web, with a model based on the contextual knowledge domains within which we operate: the Semantic Web approach (Berners-Lee et al., 2001). Both models are strongly based on a learner-centered approach so the applied research, in particular in the field of ICT and educational technologies, is moving in two directions:

1. The development of solutions for information exchange, and in general, for intelligent knowledge management;
2. The development of a collaborative/cooperative approach to knowledge building.

The Semantic Web was coined by Tim Berners-Lee to refer to a vision of the next evolution of networks that can add meaning to the navigational context of the current World Wide Web. It is the new-generation Web that makes it possible to express information in a machine-interpretable form, ready for software agents to process, as well as to understand what the terms describing the data mean both on the syntactic and semantic levels (Hendler, 2001). An important role in the development of the Semantic Web is played by ontologies (Gruber, 1993). The term is borrowed from philosophy but it is used in a different and more pragmatic sense: they are an explicit specification of a conceptualization, that is, a formal description of concepts and relationships that can exist in a knowledge domain, that is intended as a knowledge base to be shared and re-used in the real world. These ontological structures will, for instance, allow us to no longer surf the universe of documents on the Web through hypertext links from text to text, but from concept to concept; or even to retrieve information in a relevant way without the "noise" that characterizes search engines. In order to achieve this aim, formalized languages have been created (XML, RDF) to mark texts semantically. These languages, which are able to codify knowledge through domain ontologies, can be easily understood both by humans and by ad hoc programs such as semantic browsers (Dzbor, Domingue, & Motta, 2003) or by specific software agents.

The importance of ontologies has been recognized in different research fields, and even from an operational point of view the current application areas are different: from medicine to knowledge content standardization, from legal information systems to biological and geographical information systems, from e-commerce to natural language processing, and finally education (Devedzic, 2004). Our current research project is aimed at extending and integrating the construction and evolution of a semantic learning space that is backed up by ontological structures relative to educational sciences (Petrucco, 2003) based on an active consent of communities of practice.

THE EDUONTOWIKI PROJECT

The different training and background of those who contribute to educational theory, the different cultures they belong to, and the rapid development of scientific work today require the development of a series of shared conceptual schemas. It is important then to generate these schemas not as general principles but as justified, motivated, documented and finally usable schemas as control "criteria" of pedagogic discourse. As an ontology is basically a conceptual organizer of scientific discourse, it is a formidable support to hermeneutic work.

Within this context we developed the idea to build an ontology of education. The project takes into account the state of the art of educational research in Italy, France, Spain, Germany, England and Spain. The three thematic areas studied, at least in this first step of the project, are: *didactic planning*, *educational communication*, and *assessment* and *evaluation*. The ontology is "negotiated" in working exchanges and dialogical moments in order to develop a circularity of information flow within the virtual community of the experts involved in the project and other actors participating.

The project has been developed with the immediate aim of building an integrated semantic learning environment called "EduOntoWiki" (http://multifad.formazione.unipd.it/eduonto), a wiki-based environment where it is possible to construct, discuss, and contextualize ontologies suitable for describing the actors, processes and technologies of educational sciences. A wiki was chosen because it enables easy and immediate insertion, modification and sharing of texts and materials by a community of users (Wikipedia is a good example) and because it gives freedom over the knowledge creation process to users. The recent promising research in the application of the semantic Web to wiki software (Campanini, Castagna, & Tazzoli, 2004; Hepp, Bachlechner, & Siorpaes, 2005; Scaffert, Gruber, & Westenthaler, 2005) confirm this decision.

Indeed, our initial vision conceived the instrument as a tool to help in the creation of an ontology and the description of a specific knowledge domain mediated by a discussion within a community of practice. To be really useful an ontology requires the active consensus of a committed community of practice in a knowledge domain (Domingue, 1998; Trentin, 2004) as experts of do not always completely share the same categorizations, interpretations and distinctions. Often this is not only because of the reciprocal irreducibility of fundamental theoretical orders, which is both physiological and necessary, but rather because of the confusion created by the different meanings given to "key" terms in the discipline in question.

If it were possible to have an "ontological" reference model with shared lexis and semantics, as regards both terms and their relations, this would probably help to reduce conflicts which arise from misunderstandings and incomprehension. Ontologies created in this way would also have a significant side-effect for all the actors involved: first of all the definition of a common lexis (Wenger, 1998), then a strong push towards the conceptualization of tacit knowledge, and finally the sharing of a metamodel in which processes, knowledge and relations are shared. Defining ontologies which support educational applications based on the Web is therefore no simple task, above all because of the difficulty in formally conceptualising a domain which has always played on the idiosyncratic interpretation of each philosophical/pedagogical approach.

Ontologies would be useful not only to the academic community, but as far as their didactic use is concerned, we can think of an ontology or a series of "educational" ontologies, that could be used and discussed by students, teachers, and people interested in the real world of applications and training contexts. This "open" ontology paradigm can offer considerable advantages. For example, it could provide a medium which would foster the sharing of the basic knowledge in a discipline and a place where students could easily find educational resources (learning objects) with a strong context relation to the subject. Today the learning objects paradigm means that the contents which reside in the different systems used in e-learning platforms, need to become reusable, accessible and interoperable. Each object therefore needs to be described beforehand by others through unambiguous formalisms (so-called LOM, learning objects metadata) so that people can retrieve them more easily. But this retrieval paradigm is often wrong: the meaning of the metadata must be shared by most communities of users possible and this is not the common case. The practice of describing resources may at first appear to be simple and straightforward, however, when a system of description is analyzed deeply it becomes evident that it is actually ambiguous. The use of metadata presupposes not only a set of logical relations but also a specific vocabulary generally agreed upon by a linguistic community (Downes, 2004). Ontologies, integrated with social tagging processes (i.e., folksonomies) (Mathes, 2004), could indeed offer a strong support for solving this problem because every learning object would be *embedded* in the structure of the ontology itself; in this sense there is a side-effect that consists of the collaborative setting up of a learning object repository that uses the ontological base for "intelligent" consultation. Learning objects have always existed in teaching: in their practice, educators operate within a deconstruction and reconstruction process of materials and resources, but what is missing is often a conceptual and disciplinary framework to go back to, which, in today's world, could be easily accessible and consultable through the Web. It is precisely this that the various domain ontologies could provide.

The wiki interface of the ontologies is well suitable for developing a constructivist environment, where people of a learning community can add and modify the concepts dialogically (Souzis, 2005). If we then assume that the learning process is never confined to the materials used in an online course, but that it is also fruit of the interaction among the members of the group, with the wiki-based ontologies we provide a *scaffolding* (Devedzic, 2004) which will facilitate communication and the construction of meaning among all the actors involved (academics, teachers, tutors, students) and at the same time represent the structure and contents of the discipline.

FUTURE TRENDS AND DEVELOPMENTS

As ontologies in EduOntoWiki will be the result of the active involvement of both a community of practice of academics and actors from different educational fields (teachers, students and trainers), the social/relational aspect which turned out to be increasingly significant in the course of this research, led us to systematically further the study of the relationships within a community, as well as between different communities, in relation to the knowledge construction process supported by the wiki-based software. We want to verify how this environment can ease knowledge construction and formalization as "instance" from different communities of practice interacting together. In fact, the direction that the most promising, current research is taking involves the study of so-called "complex constellations of communities of practice" (Wenger, 2004). This definition has been used to describe the special relationships which unite various communities and render them permeable in such a way that they can reciprocally share knowledge, contextualizing, and enriching it with new meanings, thus favoring creative solutions to complex problems.

On the basis of these premises, we will seek to verify whether a social theory of learning can effectively lead to the overcoming of rigid borders between training/educational systems, work environments and social activities. In this sense, we can try to "free" learning so it is no longer seen to be linked to a specific area or moment of one's life, but actively constructed in the inter-community interactions of a lifelong learning continuum.

What will be investigated in particular are the negotiational interrelations between people who, in various forms, are members of different communities, people who share an active interest in all training environments and who bring valuable examples of "good practice" even if they belong to different work contexts. This aspect, led us to expand a new learning dimension, aimed at stimulating reciprocity, transferring and recontextualization processes, insofar as learning is recognized as a social/relational process, and the multiple contexts where learning takes place that becomes a precious alternative representation (Lave, 1988), effectively expressed by the learning subjects by means of a narrative description (Bruner, 1996) that a rigid codified ontology formalization would, on the contrary, risk penalizing. Narrative is used in education and training contexts to motivate and to illustrate, the reason for this is that the cognitive structures we use to understand the world around us are similar to the cognitive structures we use to understand narratives. It is assumed that the interaction, comparison and reciprocal recognition of the different communities involved will succeed in triggering off a virtuous process of crossfertilization able to transfer skills, processes, and models.

An important challenge highlighted by a close examination of international research on this theme, is that inter-community relations are not easy to manage or formalise since the members can only count on relatively limited shared meaning and practices (Friesen, 2002). Interoperability among communities, which our EduOntoWiki environment wishes to foster, is thus closely linked to a negotiation of meanings, identities and roles. Identity and roles for example, can be formalized using the semantic standard FOAF, (friend of a friend) while other important personal relations are more difficult to express. Maybe the only way is to include the innovative approach of the folksonomies (Petrucco, 2006) and/or the creation of specific "instances" in the ontologies intended mainly as a narration of personal and contextual experiences lodged in a precise space, time and place. It is not by chance that social networking tools, such as LinkedIn, Friendster and Orkut, are now considered a necessary extension of the recent blog phenomenon. In fact, we intend to evaluate whether, and in what way, it is possible that this process of *narrative conceptualization* can lead from the formulation of "descriptive instances" to spontaneous formalization, on behalf of community members, of "normative instances", that is, knowledge models which can be reused in multiple experiential contexts for solving problems.

CONCLUSION

Our research group believe that the potential effects of the Semantic Web for the world of education and training, and in particular for e-learning, will certainly be positive, but only if governed by a strong pedagogical-methodological reference structure which facilitates integration of the new technological-semantic paradigm into the more recent social theories of learning. Given these assumptions, combining the Semantic Web with social software appears to be a natural choice: it can support the creation of semantically enriched content using simple interfaces and by allowing domain experts and novices, teachers and students to collaborate within rich inter-cultural communities, sharing their true life experiences. To conclude, while it is true that the EduOntoWiki project presents considerable challenges both on a technical-scientific and on a theoretic-methodological level as it attempts to integrate the most innovative instances of international research on the Semantic Web and on pedagogic research, we also believe that fast and complex contemporaneous social evolution necessarily requires adequate instruments able to interpret and manage it.

REFERENCES

Bao, J., & Honavar, V. (2004). *Collaborative ontology building with Wiki@nt. A multiagent based ontology building environment.* In *Proceedings of the 3rd International Workshop on Evaluation of ontology-based Tools (EON2004)*, Hiroshima, Japan.

Berners-Lee, T. (1998). *A roadmap to the semantic Web, W3C Consortium*. Retrieved March 13, 2008, from http://www.w3.org/DesignIssues/Semantic.html

Bruner, J. (1996). *The culture of education*. Cambridge, MA: Harvard University Press.

Campanini, S., Castagna, P., & Tazzoli, R. (2004). Platypus wiki: A semantic wiki web. In *Proceedings of the 1st Italian Semantic Web Workshop Semantic Web Applications and Perspectives (SWAP)*, Ancona, Italy.

Cognition and Technology Group at Vanderbilt. (1996). Looking at technology in context: A framework for understanding technology and education research. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 807-840). New York: Simon & Schuster Macmillan.

Cuban, L. (1986). *Teachers and machines: The classroom use o technology since 1920*. New York: Teachers College Press, Columbia University.

Devedzic, V. (2004). Education and the semantic Web. *International Journal of Artificial Intelligence in Education, 14*, 39–65.

Domingue, J. (1998). Tadzebao and WebOnto: Discussing, browsing, and editing ontologies on the Web. In *Proceedings of the 11th Knowledge Acquisition, Modelling and Management Workshop, KAW'98*, Banff, Canada.

Downes, S. (2004). Resource profiles. *Journal of Interactive Media in Education*, 5.

Dzbor, M., Domingue, J., & Motta, E. (2003). Magpie - Towards a semantic Web browser. In *Proceedings 2nd International Semantic Web Conference (ISWC2003), Lecture Notes in Computer Science 2870/2003*. (Springer-Verlag), pp. 690-705.

Friesen, N. (2002). Semantic interoperability and communities of practice. In J. Mason (Ed.), *Global summit of online learning networks: Papers*. Retrieved March 13, 2008, from http://www.educationau.edu.au/globalsummit/papers/nfriesen.htm [cited]

Galliani, L., Costa, R., Amplatz, C., & Varisco, B. M. (1999). *Le tecnologie didattiche*. Lecce, Italy: Pensa Multimedia.

Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition, 5*(2), 199–220. doi:10.1006/knac.1993.1008

Hendler, J. (2001). Agents and the semantic Web. *IEEE Intelligent Systems*, *16*(2), 30–37. doi:10.1109/5254.920597

Hepp, M., Bachlechner, D., & Siorpaes, K. (2005). OntoWiki: Community-driven ontology engineering and ontology usage based on Wikis. In *Proceedings of the 2005*.

Jonassen, D. H. (1995). Supporting communities of learning with technology: A vision for integrating technology with learning in schools. *Educational Technology*, *35*(4), 60–63.

Lave, J. (1988). *Cognition in practice*. UK: Cambridge University Press.

Mathes, A. (2004). *Folksonomies: Cooperative classification and communication through shared metadata*. In *Proceedings of the Computer Mediated Communication, LIS590CMC (Doctoral Seminar)*, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign.

Mayer, R. (2005). *The Cambridge handbook of multimedia learning*. Cambridge University Press.

Petrucco, C. (2003). Le Prospettive Didattiche del Semantic Web, *Atti Didamatica 2003*, TED 27-28 Febbraio 2003, p.168-176.

Petrucco, C. (2006). "Folksonomie" nella rete: costruire categorie alternative, creative ed interculturali, in *TD: tecnologie didattiche: quadrimestrale di tecnologie didattiche*. - 1, (2006), p. 36-48

Schaffert, S. Gruber, A., & Westenthaler, R. (2005). A semantic wiki for collaborative knowledge formation. In *Proceedings of Semantics 2005*, Vienna, Austria.

Souzis, A. (2005). Building a semantic wiki. *IEEE Intelligent Systems*, *20*, 87–91. doi:10.1109/MIS.2005.83

Stutt, A., & Motta, E. (2004). Semantic learning webs. [Special Issue on the Educational Semantic Web]. *Journal of Interactive Media in Education*, 10.

Trentin, G. (2004). *Apprendimento in rete e condivisione delle conoscenze*. Milano: Franco Angeli.

KEY TERMS

Communities of Practice: Communities of practice are groups that form to share what they know and to learn from one another regarding some aspects of their work. People in organizations often realize they could benefit from sharing their knowledge, in-sight, and experiences with others who have similar interests or goals. For the most part, this process of informal gathering and sharing of expertise is a voluntary.

Folksonomies: Folksonomies are bottom-up taxonomies that people create on their own, as opposed to being created and imposed by a group or institution such as by professional librarians using complex and lengthy rule sets (e.g., Dewey decimal system or Library of Congress index). Synonyms include folk categorization, social tagging, and ethnoclassification. They are grass-roots classification systems for data. The value in folksonomies is derived from many people adding their own tags. The more people tagging one object, the better, because it gives alternative ways of searching for and finding information.

LOM (Learning Objects Metadata): We can define metadata as "information about information", and a LOM is a metadata about a learning object that can refer to multimedia or digital educational resources. Sets of metadata are used to identify and meaningfully describe characteristics relevant to these resources, for example, the learning resource type, the intended end user, difficulty level, educational goal, and so forth. The Learning Technology Standards Committee

(LTSC) give rise to the IEEE LOM (Learning Object Metadata) 1484.12.1-2002 standard of educational metadata.

Ontologies: An ontology is a formal representation of knowledge about an area of interest. The part of the world conceptualized or described is called the "knowledge domain." Ontologies provide a vocabulary for representing and communicating knowledge domains and a set of relationships that hold among the terms in that vocabulary.

Semantic Web: The Semantic Web is an extension of the current Web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation. The mix of content on the Web has been shifting from exclusively human-oriented content to more and more data content. The Semantic Web brings to the Web the idea of having data defined and linked in a way that it can be used for more effective discovery, automation, integration, and reuse across various applications. For the Web to reach its full potential, it must evolve into a Semantic Web, providing a universally accessible platform that allows data to be shared and processed by automated tools as well as by people. [W3C]

Social Network: A social network is a set of people or organizations or other social entities connected by a set of social relationships, such as friendships, coworking or information exchange. The connections between them may show specific patterns and can be represented by graphs. Recently many online social networking sites have begun to flourish with millions of users describing themselves in terms of who they are, what music they listen to, what books they read, and so forth, and trying to discover other people with similar interests.

Wiki: A Wiki is a collaboratively-edited Website that uses a software publishing tool. The distinguishing feature of wikis is that they typically allow all users to edit any page, with full freedom to edit, change and delete the work of previous authors. Collaborative knowledge creation is thus a central aspect of a wiki system. Wiki pages are accessible and usable at any time, and the content constantly evolves. The first wiki was created by Ward Cunningham, and the word "wiki" came from a phrase in Hawaiian—"wiki wiki"—which means "quick". It's quick because the process of editing is entwined with the process of reading. Both are done using a standard Web browser. Unlike most Websites, there's no need to edit a file, upload it to a Web server, then reload the original to check it.

This work was previously published in Encyclopedia of Information Communication Technology, edited by A. Cartelli; M. Palma, pp. 195-201, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.22
Modeling Best Practices in Web-Based Academic Development

Diana K. Kelly
San Diego Miramar College, USA

ABSTRACT

This chapter makes a case for the importance of preparing e-teachers by requiring them to have an experience as an e-learner. The chapter begins with a review of the challenges and criticisms of e-learning. Some of the literature indicates that e-learners have been dissatisfied with their learning experiences. Some academics have concerns about the rigour of courses offered through e-learning. The literature of academic development and e-learning is used to link theory with practice. The chapter provides examples of best practice in the preparation of academic staff for e-teaching. Two case studies of lived examples of e-teaching preparation are provided from a North American perspective. Future research directions are outlined, with research questions to be explored regarding the link between the preparation of e-teachers through e-learning and the quality of the e-learning experience for students.

INTRODUCTION

Academic staff in higher education are enthusiastic about getting involved in e-teaching, yet most are getting started with no experience as an e-learner. Experiencing e-learning from the learner's perspective is immensely helpful, if not essential, for effective e-teaching. Ideally, it would be best to experience a very positive and involving model of e-learning, which may be used as a model for one's own e-teaching. This chapter is a presentation of a lived example of academic development through e-learning.

The aim of this chapter is to make a strong case for the preparation of e-teachers through successful completion of a fully online programme to prepare for e-teaching. International examples of e-teaching programmes will be included, including lessons learned from participation in two North American Web-based e-teaching programmes: one generic programme (for anyone from any institution), and one programme offered by a university for new e-teachers.

DOI: 10.4018/978-1-59904-814-7.ch003

BACKGROUND

Whenever new or innovative teaching methods are used, it is normal for sceptics and critics to express concerns about the quality of teaching and learning, and e-learning has attracted some criticism. While some studies have shown “no significant difference” between learning outcomes in face-to-face classrooms and in the e-learning environment (Joy & Garcia, 2000), other studies have shown high attrition rates in e-learning, student frustrations with inexperienced e-teachers, and frustrations of e-teachers with poor student participation and learning outcomes.

While all of these criticisms cannot be directly linked to the quality of the preparation of the e-teachers, some of the frustrations of novice e-teachers show that preparation for e-teaching is a significant issue that does contribute to the quality of the overall teaching and learning experience for students and teacher.

Academic staff who plan to begin e-teaching usually need some professional development to provide an introduction to the new learning and teaching environment. According to a recent study in the United States, two thirds of 320 colleges and universities surveyed require academic staff to complete some training prior to teaching online (Lokken & Womer, 2007). Professional development for e-learning often takes the form of face-to-face workshops, one-on-one assistance and mentoring, and sometimes hybrid or blended e-learning experiences. The focus of some professional development is on the use of the technology, or on the development of materials to put up on a Web site for students. Some academic development programmes are also focused on the use of e-learning technology to enhance student learning.

The main point of this chapter is to consider the potential benefits of a professional development programme that is provided fully online. Some universities currently provide professional development preparation for teaching online through classroom instruction in computer labs or through blended learning formats. The premise of this chapter is that, while these approaches are useful, it may be even more effective for academic staff to have the opportunity to experience e-learning fully at a distance as their students will. This chapter will focus on the role of effective professional development fully through e-learning to prepare teachers for high-quality e-teaching that is focused on student involvement and learning. First the criticisms of e-learning will be explored to determine what needs to be done to improve the quality of e-learning. Second, best practices in professional development will be examined, including possible ways of translating these to the preparation for e-teaching. Third, research on the current practices in academic development for e-teaching will be explored. Finally, recommendations will be made for the improvement of the preparation of e-teachers in the future.

This chapter is not a research-based chapter, but rather a detailed review of the existing literature on the challenges of preparing academic staff for teaching in the online environment, and some of the best practices that are emerging in the field.

In this chapter, the terms *e-teaching*, *e-learning*, *Web-based learning*, and *online learning* refer to any instructional course component delivered using the Internet, whether provided fully at a distance or in a hybrid or blended format. E-teaching refers to the processes used by teachers, and e-learning refers to students learning online.

ISSUES, CONTROVERSIES, PROBLEMS

Several concerns are addressed frequently in the literature of e-learning: the quality and rigour of instruction, including learning outcomes; student persistence; and student satisfaction.

Concerns about the Quality and Rigour of E-Learning

There is a perception, particularly among those who have not experienced e-learning, that it is much less rigourous for learners and teachers than face-to-face classroom-based learning. A recent report from the Sloan Consortium (Allen & Seaman, 2006) notes that although perceptions of quality in e-learning have improved somewhat since 2003, only 62% of chief academic officers surveyed believe that learning outcomes are the same as face-to-face learning, and only 16% believe outcomes are superior in e-learning.

The highly quoted report *The No Significant Difference Phenomenon* (Russell, 1999) was a meta-analysis of research studies that showed evidence that the learning outcomes from e-learning were no different than learning outcomes in traditional courses. However, since then, others, including Phipps (2000) and Joy and Garcia (2000), claim that the original analysis was flawed and that cause and effect cannot be determined because the research did not control for extraneous variables. Joy and Garcia recommend that rather than looking at the use of technology as an issue to be debated, it is more important to focus on learning by considering this question: "What combination of instructional strategies and delivery media will best produce the desired learning outcome for the intended audience?" (p. 38).

In a meta-analysis of many studies of learning effectiveness in e-learning courses offered at a distance, Zhao, Lei, Yan, Lai, and Tan (2005) found that interaction is the key element that contributes to student learning outcomes. More interaction among students and teacher, including both asynchronous and synchronous interactions, was the most important element in many studies of e-learning.

Problems with Student Persistence in E-Learning

Student persistence in online distance learning courses is another concern. Although it is difficult to obtain accurate statistics on dropout rates, higher education officials in the United States estimate that student persistence is generally 10 to 20 percentage points lower in e-learning courses (Barefoot, 2004; Carr, 2000). A more recent study of over 300 colleges in the United States shows a much smaller difference in student retention in semester-length modules: 72% for distance learning and 78% for face-to-face modules (Lokken & Womer, 2007).

While theories on persistence in face-to-face learning emphasise engagement and social cohesion, Gibbs (2004) points out that an attempt to adapt these theories to the online environment have not been successful. However, according to one study, feelings of isolation, anxiety, or confusion can contribute to decisions to drop out of online courses (King, 2002). Several studies have found that students underestimate the workload of e-learning and will drop out when they feel they have fallen too far behind (Aqui, 2005). Jo Tait (2004) of the Open University (United Kingdom) explains that student persistence is difficult to address because there are many factors that may contribute to students' decisions to drop out. However, she also points out the important role of tutors in distance learning, and the need for academic development to teach in ways that enhance student persistence. One example of this comes from an introductory computer module at a community college in Tyler, Texas. The e-teacher, Emilio Ramos, reported that when he started holding regular chats and provided more interactive discussions for his students, his course completion rates jumped from 62% to 90%. Ramos says, "The key to having low attrition and successful completion in the online medium is the ability of instructors to keep the students engaged, and that requires quite a bit of effort from the instructor's point of view" (as cited in Carr, 2000).

Student Criticisms about the Quality of E-Learning

Some of the criticism of e-learning has come from students. In a large-scale study of students who have participated in Web-based distance learning in the United States (Noel-Levitz, Inc., 2006), students responded that the following areas needed improvement in e-learning: the quality of instruction, the responsiveness of e-teachers to students' needs, and timely feedback from e-teachers. Experienced e-learners would agree. After dropping out of an online astronomy module mid-semester, a student said, "It wasn't worth the headache. The instructor wasn't a bad teacher. He just did not have the experience with online courses" (as cited in Carr, 2000). The instructor was teaching online for the first time and had not set up the course materials and labs properly.

Other studies reinforce these findings, including a study of Canadian university students (Stodel, Thompson, & McDonald, 2006). When asked to compare their experiences in face-to-face classes with online classes, students expressed concerns about the quality of the online asynchronous discussions. Some felt that they were too drawn out, going over the same issues too many times. Others were unhappy with the flow of the discussion and felt that students were really just "checking in" rather than paying close attention to what others had already written on the discussion board.

Despite the criticisms about the quality of e-learning and the concerns about the consistency and rigour of e-learning, it is possible to address these concerns by preparing e-teachers more effectively.

SOLUTIONS AND RECOMMENDATIONS

To examine the solutions, it is important to first look at best practices for the preparation of e-teachers. Two case studies of e-teaching preparation will be described in detail as concrete examples of the best practices outlined.

Best Practice to Prepare for E-Teaching

In determining best practices for preparing for e-teaching, it is important to examine four dimensions of the preparation. First, how do novice e-teachers learn to teach online? Second, how can best practices in academic development for face-to-face teaching be translated into the preparation of e-teachers? Third, which methods of professional development are currently used in preparing academic staff for e-teaching? Fourth, what is the focus of the professional development programmes to prepare new e-teachers?

First, How do Novice Teachers Learn to Teach?

Those with no background in teaching and learning often try to reproduce what they have experienced as students. If they have seen excellent lectures, they will try to emulate them. If they have experienced small group work and lively discussions, they will try to create a similar learning environment for their own students. They often tend to use the teaching methods that best suit their own learning style.

How do novice e-teachers learn to teach online? If novice teachers tend to reproduce what they have experienced as students, what happens if they have never experienced e-learning? If they do not have a frame of reference or a prior e-learning experience to draw upon, it is very challenging to begin to teach online. This is confirmed in an article written from the perspective of a first-time online teacher. Using a reflective approach through teaching journals, student feedback, and analysis of online discussions, Yu and Brandenburg (2006) analysed several dimensions of a first-time e-teacher's experience. The issues and frustrations that were raised indicated a lack of experience in e-learning as a learner, and a significant lack of preparation to teach online. In particular, the importance of facilitating student interactions and collaboration was a lesson learned through hard experience. In a case study of another very frustrated novice e-teacher, Choi and Park (2006) outlined very similar issues and concluded,

If the new online instructor had had training regarding the pedagogical issues of online teaching and vicarious experiences through experienced online instructors, she could have been better prepared and had a different impression about online teaching. This implies that training for online instructors should be designed with more focus on the pedagogical issues of online teaching and on vicarious experiences with the actual online teaching rather than on technical issues. (p. 322)

The University of Hull developed a tutor training programme for e-teaching based on a model that started with face-to-face workshops, progressed to 4 weeks of online teaching observation, and ended with 12 weeks of online teaching practice. In their extensive evaluation of this programme, it became clear that one of the major issues was the lack of experience in e-learning as learners. The participants could only imagine what it might be like as an e-learner, and they only had a few weeks of experience as an observer (not learner participant). The programme evaluation also noted the importance of modeling practice in the tutor training programme that matched the group work and interactive discussions that would be expected of tutors when they were teaching (Bennett & Marsh, 2002).

Second, Which best Practices in Academic Development may be used Effectively in Preparing E-Teachers?

The following characteristics of academic development programmes are important to consider for long-term impact and positive changes to teaching practice: a long duration, social construction, a focus on content, an experiential model of learning, and reflection on learning.

Longer duration programmes are more effective than short-term workshops. Several studies have shown that activities for academic development that are longer in duration tend to have a more substantial impact on making changes to teaching practice over the long term (Hinson & LaPrairie, 2005).

The social construction of learning through cohorts is important for long-term impact on teaching practices. Tom Angelo (2001, p. 100) explains, "Faculty [academic] developers intent on change must engage their colleagues in constructing or adapting new, shared, contextually relevant concepts, rather than presenting faculty [academic staff] with imported prefabricated models for adoption."

Academic development programmes linked clearly to the content of teaching are more meaningful for teaching practice. When lecturers have opportunities to apply their learning to teaching in their own discipline, they are more likely to make changes to enhance their teaching. In a successful programme of professional development for online course development at Louisiana State University, participants moved from learning and practicing new skills in using the e-learning platform to applying the skills into their own courses (Hinson & LaPrairie, 2005). Another study of over 1,000 science and mathematics teachers found that professional development activities with a focus on content knowledge and active learning had the greatest positive impact on increases of knowledge and skills in teaching, which changed the teachers' teaching practices. In addition, those activities that were longer in hours of participation and spanned a longer period of time had the greatest positive impact (Garet, Porter, Desimone, Birman, & Yoon, 2001). Most lecturers consider teaching methods to be linked strongly with the discipline, so they are more likely to be accepting of ideas and advice on teaching from those within their own discipline. Those in a department who have a solid background in learning theory and teaching and learning methods can be very effective consultants to their colleagues. According to Maxwell and Kazlauskas (1992, pp. 356-357), "expert consultation by colleagues on specific teaching matters were among the most effective modes of development."

Experiential or situated learning is the notion of experiencing a model of teaching and learning to be used in a real-life situation. If learning is embedded in the context in which it will be used, it will be more meaningful to the participants (Brown, Collins, & Dugiud, 1989). Ideally, the learning experience should provide authentic situations and activities, process models, collaborative constructions of knowledge, and opportunities for reflection (Herrington & Oliver, 1995). Staff developers at Southern Cross University in Australia call their module a "staff immersion" programme that immerses participants in the role of online students, who learn about the potential for online interaction (O'Reilly & Brown, 2001).

Stephen Brookfield (1993, p. 21) explains the importance of becoming learners to learn about teaching: "I argue that regularly experiencing what it feels like to learn something unfamiliar and difficult is the best way to help teachers empathise with the emotions and feelings of their own learners as they begin to traverse new intellectual terrains."

By experiencing a well-designed and well-facilitated e-learning course about teaching online at a distance, new e-teachers understand from their own experience what a good e-learning experience feels like. The University of Southern Queensland drew upon the ideas of situated and experiential learning to develop a situated staff development model for e-teaching (Taylor, 2003). This successful model included awareness building for novice e-teachers experiencing the actual e-learning environment with authentic activities, a small amount of face-to-face training, online reflection, and peer mentoring.

Reflection on the learning experience and possible application to teaching must go hand in hand with experiential learning. Cowan (2003) points out that we learn from experience only if we also reflect upon that experience: "What have I learnt from that which will be useful to me in the future?" (p. 207) is a useful question for stimulating reflection on a learning experience. Cowan calls this "reflection for action," expanding upon Schön's (1988) model of reflective practice.

Third, Which Methods of Professional Development are used to Prepare for E-Teaching?

Whether teaching and learning occurs fully at a distance or in a blended format with some face-to-face meetings, academic staff must be well prepared to teach effectively in this new learning environment. This preparation often includes an orientation to the course management software, such as WebCT™ or Blackboard™, and usually takes the form of a face-to-face course or a series of workshops that include some underpinning learning theories, the use of features of the e-learning platform, and the development of materials to load to the course Web site. There are several examples of this type of workshop designed to prepare those who are new to e-teaching, including the Jump Start programme at Indiana University Purdue ("IUPUI Jump Start Program Prepares Faculty to Teach Online," 2006), the CampusNet online workshop provided by the University of Houston, Texas (Kidney, 2004), and the Xanadu project at the University of Turin, Italy (Trentin, 2006).

While these face-to-face workshops are helpful, they might not provide the same experience as an online e-teaching course. How will the new e-teacher learn what a good discussion looks like? How will teachers understand the experience of a new e-learning student if they have never participated in an e-learning module?

Blended or hybrid e-learning is the format of choice for some university professional development programmes, including the e-moderating course offered by the University of Glamorgan (Fitzgibbon & Jones, 2004). However, blended learning comes with the challenge of finding an appropriate time for the face-to-face sessions.

Some programmes are fully online self-paced tutorials. Prospective e-teachers are expected to work through the materials to learn to teach online. While this provides experience in using the course management software, and often gives prospective e-teachers a look at innovative possibilities for course materials and assessments, the self-paced workshops lack one of the most important aspects of e-teaching: the facilitation of online discussions.

Fully online e-teaching programmes offered at a distance include the e-moderating programme at the Open University, United Kingdom (Salmon, 2006), and the two case studies described later in this chapter. This model provides a comprehensive experience in e-learning for the prospective e-teachers and, if well modeled and well facilitated, provides a positive experience for future e-teachers to draw upon when they begin teaching online.

Fourth, What is the Focus of E-Teaching Programmes?

The focus of e-teaching programmes has been evolving. Some programmes still focus on the use of the technology for e-learning, including how to use various features of the specific e-learning course management software, such as WebCT™, Blackboard™, Moodle™, and so forth. As a part of this focus, novice e-teachers are most interested in learning how to develop materials for a module Web site with a content-driven focus when preparing to teach online. As Dianne Conrad (2004) noted in her study of novice e-teachers, teachers' overall concerns stemmed from their perception of their role as "deliverers of content." They appreciated the e-learning platform as a place to put more content to be accessed by their students. However, they did not seem to be concerned about issues of social interactions among learners, and facilitation and mentoring of learners. The participants in Conrad's study took part in face-to-face workshops and one-on-one mentoring sessions that focused only on the use of the technology for e-learning. While this is a necessary part of professional development for e-teaching, the professional development focus must go beyond a focus on technology and content.

A study of over 500 members of the Multimedia Educational Resource for Learning and Online Teaching (MERLOT) indicated that the focus of interest for e-teachers has shifted from technology skills training to enhancing skills in e-moderating for high-quality online learning (Kim & Bonk, 2006). Gilly Salmon's (2006) well-respected work in this field indicates that this is a crucial component in successful e-teaching. Using Vygotsky's "zone of proximal development" as a model, those who provide e-teaching workshops online can help the academic staff to develop their skills in facilitation by modeling behaviours in asynchronous discussions that will be more productive for their learning development, including questioning techniques that probe for deeper learning (Welk, 2006). Through this modeling, participants will experience the type of facilitation that will help them to be more effective facilitators of online asynchronous discussions.

To summarise, high-quality e-teaching programmes focus on the learning theories and principles that have been proven to be effective in face-to-face teaching, adapting them to the e-learning environment. The "Seven Principles of Effective Teaching," originally developed by Chickering and Gamson, were used by Graham, Cagiltay, Lim, Craner, and Duffy (2001) to provide a useful way of looking at the qualities of teaching that help students learn in the online environment. These seven principles include contact between students and teacher, cooperation among students, active learning, prompt feedback to students, time on task, high expectations, and diverse talents and ways of learning. By applying these same principles to e-learning, e-teachers can fine-tune their teaching practices.

TWO CASE STUDIES: MODELS OF THE FULLY ONLINE E-TEACHING PROGRAMMES

To provide concrete examples of model programmes to prepare e-teachers, two programmes are described (Kelly, 2000, 2002). Both programmes were provided fully online at a distance with no face-to-face meetings. This was an intentional part of the design to give participants the same experience that learners will have when they participate in a fully online programme rather than blended or hybrid learning.

The first case study describes a postgraduate-level certificate programme open to anyone in the world who has teaching experience in education, higher education, or in professional development in the business world. The second case study describes a programme that was designed specifically for Walden University to prepare tutors to work online at a distance with postgraduate students in the PhD in education programme.

The most important similarities between the programmes are the strong grounding in learning theory, a focus on facilitating active learning through asynchronous discussions and collaborative activities, and, most importantly, providing a relevant learning experience in context and allowing reflection on this experience.

University of California at Los Angeles: Online Teaching Programme

In 1999 I decided to update my skills as an academic and learn about online classes. Because my time was limited and my day-to-day schedule was somewhat unpredictable, I was happy to find a fully online certificate programme on teaching online offered by UCLA (University of California at Los Angeles, http://www.uclaextension.edu). The certificate programme consisted of five core modules and one elective module. The school recommends taking two modules at a time, so I started with the first two core modules: Introduction to Online Technologies and Developing Curriculum for Online Programs. Other core modules were Teaching and Learning Models for Online Courses, Internet and Online Teaching Tools, and Practicum in Online Teaching (a capstone course). I chose the module Multimedia Production as my elective module. Modules were offered in 4-week or 6-week periods in four terms each year. Taking two modules each term, it was possible to complete the certificate programme in three terms over 9 months.

Flexibility in Learning

Although there were clear starting and ending dates for each module, the time students put into the actual course work was entirely flexible. Almost all of the collaborative work and online discussions were asynchronous. Each module usually had 10 to 15 participants, and only those who were experienced teachers were accepted into this programme. Some were from very remote areas, and they were happy to participate in this online programme because they had no university within traveling distance.

People have often asked how much time this online programme required of me as an e-learner. My experience was that, as with any course, it depends on the student's level of interest and motivation. Students could spend as much time online and completing assignments as they wanted to, but on average, I probably spent about 10 hours per week on each module, and more when major assignments were due. Those 10 hours per course (20 hours per week for two courses) were spread out over lunchtimes, evenings, weekends, and generally whenever I had a chance to work on assignments.

International Participation

As a fully online programme offered at a distance, we had a very international group of participants from many parts of the world including Hong Kong, Saudi Arabia, Australia, Switzerland, and North America. This resulted in rich discussions with an international perspective. We learned quite a bit about educational issues in other countries and gained some new ideas. Having such broad international participation would not have been feasible if face-to-face sessions had been required.

Mandatory Student Orientation

After enrolling I received a welcoming e-mail from OnlineLearning.net, UCLA's online learning provider. It recommended ordering textbooks soon, and provided links to several places where textbooks could be ordered online. The e-mail also included information about how to download the necessary software from Embanet, the online course management system, which was a simple process.

Prior to the start of the first module, it was required that all participants complete a four-part online self-paced orientation to the Embanet software. There were dire warnings that those who did not complete the entire orientation would be removed from the module. Although this sounded a bit harsh, it soon because obvious that the online orientation to the software was absolutely essential to the experience. Students learned how to use the asynchronous discussion groups, submit assignments, participate in synchronous chat, go to the course resources, get help from Embanet, and so forth. Without this orientation and the easy availability of Embanet's technical-support team (by phone and e-mail), this fully online programme would have been very frustrating, if not impossible.

The Importance of Technical-Support Systems for E-Learners

High-quality e-learning is impossible without good technical support in place, ideally 24 hours per day, 7 days a week. If this is not possible, then it should be available at specified times when usage is highest, particularly on weekends and evenings. It can be incredibly frustrating if the course Web site crashes, especially when an assignment is due. Embanet had a habit of doing this periodically. Fortunately, Embanet had excellent technical support for major problems like a Web site crash or individual student problems such as software incompatibility. Students could e-mail or call the technical-support desk for immediate assistance with any problem.

Providing a Welcoming Learning Environment

As with any face-to-face class, introductions at the beginning of each module help students to become involved more quickly. Online learning is no different. Participants have a desire to make connections with other students, but the e-teacher must facilitate this. This initial interaction on the discussion board also helps e-learners get into the habit of checking in to the course Web site regularly to see if anyone new has added their introduction.

About a week before the module started, we received a welcoming e-mail from the e-teacher. She asked all of us to go to the course Web site and introduce ourselves, including our background, our interest in online learning, and what we were hoping to achieve, as well as any personal details we wanted to share. As a good student, I was happy to follow her directions and thought I would be the first one there, but I was not. Everyone was enthusiastic about getting started. It was interesting to learn that they were not only from higher education, but also from secondary education and from training and development.

Facilitation of Asynchronous Discussions

At the beginning of each module, the e-teacher provided general discussion guidelines, or netiquette rules, which emphasised the importance of participating, contributing, and encouraging. With these guidelines, our discussions were very positive and encouraging. Even when someone in the class was struggling with an issue, many would respond positively to offer advice and encouragement.

The online asynchronous discussions were lively and stimulating, and the teacher was an active participant, sometimes providing answers to questions that were raised, and sometimes raising new stimulating questions. Because the discussions were written and asynchronous (occurring whenever someone felt like submitting a discussion item), they were much more thoughtful than the typical face-to-face discussion. People had time to read another student's thoughts, digest them, and respond thoughtfully. Students also became great resources for the rest of the group. If one person raised a question, often two or three others would respond with answers or online resources. The teacher did not feel obliged to be the only one providing answers. In fact, the teacher was truly a "guide on the side," actively participating and guiding the discussion but encouraging the students to provide the majority of input.

All of the online modules seemed to follow the same pattern in terms of organization, discussions, and assignments, but the quality of the learning experience was really influenced by the tone set by the teacher. Those modules in which the teacher was less involved seemed to be less interesting and less involving for the students. The modules with the greatest interaction and that stimulated more learning were those in which the e-teacher was actively engaged on a daily basis, and showed his or her enthusiasm for the e-learners, the topic, and the discussions through comments that were worded in the most positive way. These modules were so involving that I found myself checking into the course Web site several times a day to see the new postings: at lunchtime in front of my computer and most evenings.

Practical and Relevant Assignments

The assignments for each module were very practical, relating the theoretical readings to the creation of online materials. However, one big difference was that we submitted our assignments to the module Web site so all e-learners in that module could look at the assignments and offer formative comments and suggestions before final submission. The comments were very positive and affirming, and when suggestions were offered it was in the spirit of helpfulness: e-learners helping other e-learners. This was all a part of the process of learning how to provide formative feedback to e-learners, as well as modeling an excellent collaborative process to be used with our own students.

Group Projects

Some of the course projects and assignments were done in groups. It is possible to do group work online if it is well organized and facilitated. One assignment was to create a fictional module Web site around a particular topic. In my small group (members were assigned by the teacher), there was one member in Switzerland, one in New York, one in Texas, and two in California. We decided who would do which piece of the project, and most of our work was done asynchronously through our own group discussion site that the teacher had set up on the course Web site. We also decided to try a synchronous chat just to check in with the group members. Considering the 9-hour time difference between California and Switzerland (and the others in between), we determined a time that would work for all of us. It worked pretty well, but at times it was somewhat confusing because just as I thought of a response to someone's comment, there were three other responses about something else. So, the asynchronous mode generally worked best to pull our project together. We were happy with the course Web site we created as a team, and it was also interesting to see how the other groups developed their course Web sites.

Lessons Learned as an E-Student

From this experience, I learned that e-teaching is completely different from a scheduled lecture or tutorial meeting 3 hours per week. It is much more flexible. As prospective e-teachers, we wanted to know how much time an online module requires of e-teachers. Our e-teachers answered truthfully that the busiest time in online courses is on the weekends because that is when most students have the time to do some concentrated work. So e-teachers plan to be online several times on weekends to respond to questions and problems. In addition, they check in everyday to read student comments and assignments, facilitate discussions, and address questions. If our e-teachers were traveling to a conference, they would often let us know that they would be out of touch for a particular period of time until they had their laptop set up in the hotel room. Because e-learners may also contact the teacher privately through e-mail on the module Web site, the best teachers also felt that it was important to be quick in responding to these individual queries. However, e-teachers usually said that questions that were not of a personal nature should be addressed to the discussion board so that other students may respond, and/or see the teacher's response. There is no question that e-teaching online takes a lot of time and dedication, and a learner-centred approach to e-teacher availability.

As with any learning and teaching method, online learning is not the preferred learning mode for everyone. Some of the people in the online course said that they really missed the face-to-face contact or hearing the voices of the teacher or the other students. Perhaps it is a learning style issue. On the UCLA Web site (and other university Web sites), there is a self-assessment tool for prospective online students to determine how well suited they may be for the e-learning experience. It is important for prospective e-learners to recognize that e-learning also takes much more self-discipline and self-motivation than a face-to-face class. Those who think it will be easier are in for a big surprise. Anyone looking for an easy ride really does not belong in an online programme.

The final module of the UCLA programme was a supervised e-teaching experience with Alfred Rovai, who has written widely on e-teaching practices and was an excellent mentor to us one on one. After receiving the UCLA Certificate in Online Teaching, I was able to immediately apply my learning to create some online self-paced workshops for lecturers interested in learning new teaching strategies. These e-workshops were also designed to provide a test experience as an e-learner, allowing one to see how e-learning works, how it feels, the pitfalls, and the advantages.

WALDEN UNIVERSITY: ORIENTATION FOR NEW FACULTY MENTORS

Background on Walden University

Walden University (http://www.waldenu.edu) is an accredited postgraduate university that started in 1970 based on the learner-centred principles outlined by Harold Hodgkinson, professor at

University of California at Berkeley, in his 1969 article in the journal *Soundings*. The founders of Walden, inspired by Henry David Thoreau, envisioned an institution that would provide the opportunity for adults to earn doctorate degrees as scholar-practitioners so that they might develop into leaders committed to the betterment of society.

Walden University is fully accredited by the Higher Education Commission in the United States, offering master's and doctoral programmes in education, management, nursing, health sciences, psychology, social service, public policy, and engineering. Walden University is based in the United States and has 20,000 online students from 95 countries, including a partnership with the University of Liverpool for three online programmes: the MBA, MS in IT, and MA in information systems management. Walden is part of the large Laureate International Universities network that includes a total of over 240,000 students in 25 universities in 16 countries.

Mentoring Research Students Online

In December 2005, I was invited by Terry O'Banion to join Walden University as a faculty mentor in the College of Education, working part-time at a distance, supervising and mentoring doctoral-level students in the Community College Leadership and Adult Education Leadership programmes. Terry O'Banion is the director of the Community College Leadership programme and was very enthusiastic about Walden University's focus on learning, as outlined in his 1997 book *A Learning College for the 21st Century*. He explained that the doctoral students at Walden must complete three "knowledge area modules" (KAMs) prior to beginning work on their dissertations. In other doctoral programmes in the United States, these might be considered equivalent to the required qualifying exams that are normally completed prior to the dissertation. These KAMs are very lengthy, analysing the breadth, depth, and application of a particular topic relevant to the area of the student's academic work. Each KAM focuses on a different theme: KAM I is Principles of Social Change, KAM II is Principles of Human Development, and KAM III is Principles of Social Systems. Within each KAM, the *breadth* portion is a study of major theorists, the *depth* portion is a study of the current literature that applies the theory to a specific topic, and the *application* portion provides students with the opportunity to apply what has been learned in the *breadth* and *depth* sections through a mini research study or a real-world project aimed at creating a positive social change.

Mandatory Orientation for New Mentors

As a new faculty mentor, I was required to complete a 12-week orientation programme provided by Walden University starting in January 2006. This programme was similar in some ways to the UCLA programme described earlier, with a strong focus on learning theory, facilitation of discussions, and provision of good support to students. However, the major difference in the programme is that we also needed to learn the "Walden way" of KAMs, learning agreements, personal development plans, and the methods for submitting and assessing work. Our online orientation was facilitated by an experienced Walden faculty mentor from the education programme, who had a wealth of experience she was willing to share with us. The group was small with only four of us, but we had well-facilitated discussions around issues of e-teaching and mentoring at Walden University. In order to pass the orientation, we were required to participate in all discussions, complete all assignments and projects (usually one each week), and successfully pass an exam at the end of the orientation course. Those who were unable to complete all of the work were invited to participate in another upcoming orientation, but were not allowed to teach for Walden until successful completion of the full orientation. This fully online orientation was an excellent model for new faculty mentors to experience learning at Walden before starting to teach there.

Lessons Learned from the Walden Orientation Programme

Walden University has a unique structure with its KAMs, and new doctoral students can become very frustrated in the early stages. Without the extensive orientation programme, as a faculty mentor, I would have been equally frustrated. However, by providing a safe environment for new e-teachers to learn about Walden's structures and methods, I was able to provide my early students with the support and advice they needed. Because the programme is fully online for students, it was important for the prospective mentors to gain experience as learners in the same e-learning environment. Of course, some things will be learned only through the experience of working with mentees, but the preparation through the orientation programme provided a firm foundation for us as new mentors.

After examining the best practices in the preparation of e-teachers and two case-study examples of e-teachers prepared through e-learning, it is important to look at the future trends in e-teaching and e-learning.

FUTURE TRENDS

With the rapid growth of e-learning and exponentially growing demand for fully online courses, universities are starting to pay more attention to the need for professional development to prepare e-teachers more effectively. Some universities offer their own professional development programmes in e-teaching, although most are short in duration and few are fully online. It will be important in the future for universities to consider how to best prepare novice e-teachers for effective online teaching and include facilitation and e-moderating to promote deep learning and student success. Will all institutions provide their own fully online e-teaching programmes, or will many academic staff participate in programmes offered by a few institutions that already have excellent programmes in e-teaching? In either case, the need for providing a fully online experience will be met.

CONCLUSION

The purpose of this chapter has been to highlight one important way of supporting academic staff in higher education who are thinking of introducing e-learning as a way of enhancing student learning. This is a critically important part of applied e-learning and e-teaching in higher education. Without good preparation for e-teaching, the quality of e-learning experiences for students will vary widely.

After my two experiences as an e-learner to prepare for e-teaching, I strongly believe in the importance of having a high-quality experience as an e-learner fully at a distance before attempting to be an e-teacher. Universities and colleges that are serious about the quality of their e-learning programmes require their prospective e-teachers to complete a programme or module in online teaching as an e-learner prior to teaching online. If it is not feasible to offer this programme within the institution, it would be worthwhile to support academic staff in participating in high-quality programmes offered by other institutions.

Prospective e-teachers who want to provide a high-quality learning experience for their e-learners should plan to participate in a well-organized, well-facilitated fully online course to see how it feels from the student perspective, whether or not it is required by their own universities. Through this type of immersion in e-learning as professional development, it is likely that the quality of online instruction will continue to improve, resulting in better student e-learning outcomes in the future.

FUTURE RESEARCH DIRECTIONS

Further research must be done in the future to demonstrate the links between e-teacher preparation and student learning and success. Although all research on student success is challenging due to the number of variables that contribute to learning outcomes, it will be necessary to demonstrate the effectiveness of the professional development programmes to prepare e-teachers. Without this evidence, it is difficult to justify a lengthy professional development programme that models best practices in e-learning because these programmes are expensive. This research may also reveal some new ideas for professional development that will enhance e-learning outcomes in the future.

Research in the area of preparation for e-teaching should ideally analyse student learning outcomes, including completion and success rates in e-learning modules and courses. In programmes or courses with high success rates, how are the e-teachers prepared? How many of the e-teachers have participated as e-learners? Literature on e-teaching has often shown, through interviews or surveys of e-teachers, that quite a few felt unprepared when they started e-teaching. It would be interesting to find universities with e-teachers who felt well prepared to determine how many of these e-teachers had an e-learning experience first.

The student perspective is another important area to include in researching this issue. What do students look for in an effective e-teacher? Are the characteristics of e-teachers different in universities that provide preparation for e-teaching through e-learning? How much interaction and involvement in learning do students experience with e-teachers who were e-learners first in comparison to e-teachers who did not have their own e-learning experiences?

Future research in the area of Web-based academic development as preparation for e-teaching must be linked to the literature of academic development, experiential learning, and e-learning. The objective of this research will be to bring together relevant learning theories with e-teaching in order to provide the best possible e-learning experience for our students.

ADDITIONAL READING

The following sources were selected because they focus on the practical aspects of providing professional development for those who are preparing to be e-teachers. They all emphasise the importance of supporting academic staff prior to teaching online. These sources provide details of the ways in which e-teaching is much more than simply putting materials up on a Web site for students. The idea of building learning communities at a distance is a theme that runs through many of these references for further reading.

REFERENCES

Allen, I., & Seaman, J. (2006). *Making the grade: Online education in the United States.* Needham, MA: The Sloan Consortium (Sloan-C). Retrieved July 7, 2007, from http://www.sloan-c.org/publications/survey/pdf/making_the_grade.pdf

Angelo, T. (2001). Doing faculty development as if we value learning most. In D. Lieberman & C. Wehlburg (Eds.), *To improve the academy* (Vol. 19, pp. 97-112). Boston: Anker Publishing Company, Inc.

Aqui, Y. (2005, June 27-30). *Characteristics of the online learner: Experiences, participation level, and achievement.* Paper presented at the Meeting of the National Educational Computing Conference (NECC) of the International Society for Technology in Education, Philadelphia. Retrieved January 9, 2007 from http://center.uoregon.edu/ISTE/uploads/NECC2005/KEY_7030859/Aqui_Aqui_OnlineLearnerCharacteristics-NECC2005_RP.txt

Barefoot, B. (2004, February). Higher education's revolving door: Confronting the problem of student drop out in U.S. colleges and universities. *Open Learning*, *19*(1), 9–18. doi:10.1080/0268051042000177818

Bennett, S., & Marsh, D. (2002, January). Are we expecting online tutors to run before they can walk? *Innovations in Education and Teaching International*, *39*(1), 14–20. doi:10.1080/13558000110097055

Brookfield, S. (1993). Through the lens of learning: How the visceral experience of learning reframes teaching. In D. Boud, R. Cohen, & D. Walker (Eds.), *Using experience for learning* (pp. 21-32). Milton Keynes, United Kingdom: The Open University Press.

Brown, J., Collins, A., & Dugiud, P. (1989). Situated cognition and the culture of learning. *Educational Research*, *18*(1), 32–42.

Carr, S. (2000, February 11). As distance education comes of age, the challenge is keeping the students. *The Chronicle of Higher Education*.

Choi, H., & Park, J. (2006). Difficulties that an online novice instructor faced. *The Quarterly Review of Distance Education*, *7*(3), 317–322.

Conrad, D. (2004, April). University instructors' reflections on their first online teaching experiences. *Journal of Asynchronous Learning Networks*, *8*(2), 31–44.

Cowan, J. (2003). Learning from experience. In P. Kahn & D. Baume (Eds.), *A guide to staff and educational development* (pp. 192-211). London: Routledge.

Fitzgibbon, K., & Jones, N. (2004, March). Jumping the hurdles: Challenges of staff development delivered in a blended learning environment. *Journal of Educational Media*, *29*(1), 25–35. doi:10.1080/1358165042000186253

Garet, M., Porter, A., Desimone, L., Birman, B., & Yoon, K. (2001). What makes professional development effective: Results from a national sample of teachers. *American Educational Research Association Journal*, *38*(4), 915–945.

Gibbs, G. (2004, February). Editorial. *Open Learning*, *19*(1), 3–7. doi:10.1080/0268051042000177809

Graham, C., Cagiltay, K., Lim, B., Craner, J., & Duffy, T. (2001, March/April). Seven principles of effective teaching: A practical lens for evaluating online courses. *The Technology Source*. Retrieved July 4, 2007, from http://www.technologysource.org/article/seven_principles_of_effective_teaching

Herrington, J., & Oliver, R. (1995, December 4-6). *Critical characteristics of situated learning: Implications for instructional design of multimedia*. Paper presented at the Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education (ASCLITE), Melbourne, Australia. Retrieved July 4, 2007, from http://www.ascilite.org.au/conferences/melbourne95/smtu/papers/herrington.pdf

Hinson, J., & LaPrairie, K. (2005). Learning to teach online: Promoting success through professional development. *Community College Journal of Research and Practice*, *29*, 483–493. doi:10.1080/10668920590934198

IUPUI Jump Start program prepares faculty to teach online. (2006, September 15). *Academic Leader.*

Joy, E., & Garcia, F. (2000, June). Measuring learning effectiveness: A new look at no-significant-difference findings. *Journal of Asynchronous Learning Networks*, *4*(1). Retrieved July 5, 2007, from http://www.sloan-c.org/publications/jaln/v4n1/pdf/v4n1_joygarcia.pdf

Kelly, D. (2000, August 18). Adventures on the cyber shop floor. *Times Higher Education Supplement.* Retrieved January 9, 2007, from http://www.thes.co.uk/search/story.aspx?story_id=63580

Kelly, D. (2002, July 1-4). *Being an online learner before becoming an online teacher: Ten lessons from experience.* Paper presented at the Improving University Learning and Teaching (IUT) Conference, Vilnius, Lithuania.

Kidney, G. (2004, June). When the cows come home: A proven path of professional development for faculty pursuing e-learning. *T.H.E. Journal, 31*(11). Retrieved July 3, 2007, from http://thejournal.com/articles/16803

Kim, K., & Bonk, C. (2006). The future of online teaching and learning in higher education: The survey says. *Educause Quarterly, 4*, 22-30. Retrieved January 10, 2007, from http://www.educause.edu/ir/library/pdf/EQM0644.pdf

King, F. (2002). A virtual student: Not an ordinary Joe. *The Internet and Higher Education*, *5*(2), 157–166. doi:10.1016/S1096-7516(02)00085-4

Lokken, F., & Womer, L. (2007). *Trends in e-learning: Tracking the impact of e-learning in higher education.* Washington, DC: Instructional Technology Council. Retrieved July 7, 2007, from http://www.itcnetwork.org/Trends_in_Elearning_2006.pdf

Maxwell, W. E., & Kazlauskas, E. J. (1992). Which faculty development methods really work in community colleges? A review of research. *Community/Junior College Quarterly, 16*, 351-360.

Noel-Levitz, Inc. (2006). *National online learners' priorities report.* Iowa City, IA: Author.

O'Banion, T. (1997). *A learning college for the 21st century.* Washington, DC: American Association of Community Colleges.

O'Reilly, M., & Brown, J. (2001). *Staff development by immersion in interactive learning online.* Lismore, Australia: Southern Cross University. Retrieved July 7, 2007, from http://ausweb.scu.edu.au/aw01/papers/refereed/o_reilly/paper.html

Phipps, R. (2000). *What's the difference?* Washington, DC: The Institute for Higher Education Policy.

Russell, T. (1999). *The no significant difference phenomenon.* Chapel Hill, NC: Office of Instructional Telecommunications, North Carolina State University.

Salmon, G. (2006). *Scaffolding for e-moderator's development: The early years.* Beyond Distance Research Alliance, University of Leicester. Retrieved January 9, 2007, from http://www.atimod.com/docs/atim12dec12doc%20(2).pdf

Schön, D. (1988). *Educating the reflective practitioner.* San Francisco: Jossey-Bass Publishers.

Stodel, E., Thompson, T., & McDonald, C. (2006, December). Learners' perspectives of what is missing from online learning: Interpretations through the community of inquiry framework. *International Review of Research in Open and Distance Learning*, *7*(3), 1–24.

Tait, J. (2004, February). The tutor/facilitator role in student retention. *Open Learning*, *19*(1), 97–109. doi:10.1080/0268051042000177872

Taylor, J. (2003, May). Managing staff development for online education: A situated learning model. *Journal of Higher Education Policy and Management*, *25*(1), 75–87. doi:10.1080/13600800305742

Trentin, G. (2006). The Xanadu project: Training faculty in the use of information and communication technology for university teaching. *Journal of Computer Assisted Learning*, *22*(3), 182–196. doi:10.1111/j.1365-2729.2006.00168.x

Welk, D. (2006). The trainer's application of Vygotsky's "zone of proximal development" to asynchronous online training of faculty facilitators. *Online Journal of Distance Learning, 9*(4). Carrollton, GA: University of West Georgia Distance Education Center.

Yu, C., & Brandenburg, T. (2006). I would have had more success if…Trials and tribulations of a first-time online instructor. *The Journal of Technology Studies, 32*(1), 43-52. Retrieved July 5, 2007, from http://scholar.lib.vt.edu/ejournals/JOTS/v32/v32n1/pdf/yu.pdf

Zhao, Y., Lei, J., Yan, B., Lai, C., & Tan, H. (2005, August). What makes the difference? A practical analysis of research on the effectiveness of distance learning. *Teachers College Record*, *107*(8), 1836–1884. doi:10.1111/j.1467-9620.2005.00544.x

Bennett, S., Priest, A., & Macpherson, C. (1999). Learning about online learning: An approach to staff development for university teachers. *Australian Journal of Educational Technology, 15*(3), 207-221. Retrieved October 20, 2007, from http://www.ascilite.org.au/ajet/ajet15/bennett.html

Developing online instructors requires more than workshops. (2004, November). *Distance Education Report*.

Donovan, M., & Macklin, S. (1999). The Catalyst project: Supporting faculty uses of the Web with the Web. *Cause/Effect Journal, 22*(3). Retrieved October 20, 2007, from http://www.educause.edu/ir/library/html/cem/cem99/cem9934.html

Driscoll, M. (1998). *Web-based training*. San Francisco: Jossey Bass.

Faculty training, on demand. (2004, December 15). *Distance Education Report*.

Gibbons, H., & Wentworth, G. (2001). Andrological and pedagogical differences for online instructors. *Online Journal of Distance Learning Administration*, *4*(3).

Hagel, J., & Armstrong, A. (1997). *Net gain: Expanding markets through virtual communities.* Boston: Harvard Business School Press.

Jones, G. (1997). *Cyber schools: An educational renaissance*. Englewood, CO: Jones Digital Century, Inc.

Kahn, B. (Ed.). (1997). *Web-based instruction.* Englewood Cliffs, NJ: Educational Technology

Kandlbinder, P. (2001, July 2-5). Peeking under the covers: Understanding the foundations of online academic staff development. In L. Richardson & J. Lidstone (Eds.), *Flexible learning for a flexible society* (pp. 372-378). Proceedings of ASET-HERDSA 2000 Conference, Toowoomba, Australia. Retrieved October 20, 2007, from http://www.ascilite.org.au/aset-archives/confs/aset-herdsa2000/procs/kandlbinder2.html

Kosak, L., Manning, D., Dobson, E., Rogerson, L., Cotnam, S., Colaric, S., et al. (2005, June 27-30). *Prepared to teach online? Perspectives of faculty in the North Carolina University system.* Paper presented at the Meeting of the National Educational Computing Conference (NECC) of the International Society for Technology in Education, Philadelphia. Retrieved October 20, 2007, from http://center.uoregon.edu/ISTE/uploads/NECC2005/KEY_7248775/Dobson_necc_paper_RP.pdf

Laurillard, D. (2002). Rethinking teaching for the knowledge society. *Educause Review, 37*(1), 16-25. Retrieved October 20, 2007, from http://www.educause.edu/ir/library/pdf/FFPIU017.pdf

Organisation for Economic Co-operation and Development (OECD). (2005). *E-Learning in tertiary education*. Paris: Author. Retrieved October 20, 2007, from http://www.oecd.org/dataoecd/55/25/35961132.pdf

Paloff, R., & Pratt, K. (1999). *Building learning communities in cyberspace.* San Francisco: Jossey Bass.

Porter, L. (1997). *Creating the virtual classroom.* New York: John Wiley & Sons.

Rovai, A. (2007). Facilitating online discussions effectively. *The Internet and Higher Education*, *10*, 77–88. doi:10.1016/j.iheduc.2006.10.001

Russell, D. (2006, July 5-7). *Online professional development for innovative educators.* Paper presented at the Meeting of the National Educational Computing Conference (NECC) of the International Society for Technology in Education, San Diego, CA. Retrieved October 20, 2007, from http://www.iste.org/Content/NavigationMenu/Research/NECC_Research_Paper_Archives/NECC_2006/Russell_Donna_NECC06.pdf

Salter, G., & Hansen, S. (2000, July 2-5). Facilitating Web-based staff development in higher education. In L. Richardson & J. Lidstone (Eds.), *Flexible Learning for a Flexible Society: Proceedings of ASET-HERDSA 2000 Conference*, Toowoomba, Australia (pp. 612-617). ASET & HERDSA. Retrieved October 20, 2007, from http://www.ascilite.org.au/aset-archives/confs/aset-herdsa2000/procs/salter1.html

Sherer, P., Shea, T., & Kristensen, E. (2003). Online communities of practice: A catalyst for faculty development. *Innovative Higher Education, 27*(3), 183–194. doi:10.1023/A:1022355226924

Wilson, G., & Stacey, E. (2004). Online interaction impacts on learning: Teaching the teachers to teach online. *Australasian Journal of Educational Technology, 20*(1), 33-48. Retrieved October 20, 2007, from http://www.ascilite.org.au/ajet/ajet20/wilson.html

Zilberman, D. (2002) Training online for teaching online. *Educause Resources: Effective Practices.* Retrieved January 10, 2007, from http://www.educause.edu/Browse/705?ITEM_ID=90

This work was previously published in Applied E-Learning and E-Teaching in Higher Education, edited by R. Donnelly; F. McSweeney, pp. 35-55, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.23
The Use of Weblogs in Language Education

Thomas Raith
The University of Education Heidelberg, Germany

ABSTRACT

This chapter explores in how far Web 2.0, Weblogs in particular, has changed foreign language learning. It argues that Weblogs, along with Web 2.0, have created new genres for which users need new forms of literacy. A qualitative study on the relationship between the online audience of Web 2.0 and learners' writing processes is presented and the findings are discussed. The study supports the assumption that learners are aware of the social interaction taking place through weblogs and that this awareness of audience influences the writing process. The author's intention is to point out that Web 2.0 has created new communities of language practice and that foreign language learning is happening in these discourse communities through social interaction. The challenge in foreign language education is to integrate these communities of practice into the foreign language classroom.

DOI: 10.4018/978-1-60566-190-2.ch015

INTRODUCTION

From the very beginning, the Internet was a community that offered many possibilities for networking, linking people worldwide and for publishing information for the online community. In recent years the term "social software" has come to describe a new phenomenon within the online world. The social software application that has gained the most attention in recent years are weblogs. Originally, weblogs were mainly created to link together pages on the Web that the weblog author considered interesting or noteworthy. These lists of links included the weblog author's comments on the content of the linked websites. A community of weblog-owners networked around a certain topic, linking and exchanging information. One famous example was the weblog of Stanford college students Filo and Yang, who created a link-catalogue in 1994 that eventually developed into the Internet portal Yahoo! (Möller, 2005).

In the early days of weblogs, the users needed knowledge in programming and had to host blogs on their own servers. Since the introduction of blog-hosters in 1999, weblogs have become more accessible and easier to start for every user on the net: publishing a post on a weblog nowadays is as easy as writing an e-mail. As a result, since 1999 the number of weblogs on the Internet has increased dramatically. In April 2007, Technorati tracked 70 Million weblogs worldwide and estimated that 120,000 new blogs were being created each day ("The State of the Live Web", 2007). Although the numbers only show 15,5 Million active weblogs, the blogosphere is constantly growing. Blogs have challenged traditional journalism through fast and multi-perspective coverage of news which was not possible before. Already by the end of 2004, weblogs were so popular that the Merriam-Webster dictionary chose it as its "Word of the Year for 2004" (Richardson, 2006). Weblogs are only one tool in the growing Web 2.0 family which has changed the way people communicate and network. In how far do these changes affect education and foreign language learning? This chapter discusses this topic in relation to weblogs as Web 2.0 tools. The basic question will be whether weblogs transform learning in the foreign language writing classroom.

WEB 2.0 AND THE TRANSFORMATION OF LANGUAGE LEARNING

O'Reilly (2005) sees, among others, two key features that distinguish Web 2.0 from Web 1.0: the platform-based usage of the Internet and harnessing collective intelligence of Internet users. The Web is the platform on which users work collaboratively and on which they store and exchange data. Rather than installing and using software on the PC, services are used online to create blogs, documents and wikis. These features can be seen as the basic principle of all social software devices which link users for collaboration and social interaction. However, the phenomenon of collaborative projects, like Wikipedia, and the rapid growth of the blogosphere, to name only two, is not only a consequence of new Web 2.0 technology. As Alby (2007) points out, these phenomena go hand in hand with faster Internet connections via broadband and flat rates that are affordable for the masses.

How far, then, has Web 2.0 transformed language learning? Warschauer (2004) describes three stages of CALL which have emerged since the 1970s and have represented the development of technology, on the one hand, and the formation of language acquisition approaches, on the other hand. The first stage, structural CALL, uses drill and practice activities focussing on correct language use. Structural CALL followed an audio-lingual approach to language learning, and the technology was mainframe computers. The 1980s and 1990s were marked by the upcoming communicative approach to language learning. At the same time, the introduction of PCs offered the technology for computer-assisted communicative exercises. The third stage, which Warschauer (2004) calls integrative CALL, has a socio-cognitive approach to language teaching and uses computers for authentic discourse. In this last stage, the computer functions as a tool that connects learners for interaction. In earlier stages of CALL, computers were seen as a tool to support the language learning process. Integrative CALL is different, because it doesn't only use technology to create space for isolated language learning activities, but it uses technology that is made for communication. With Web 2.0 this shift in CALL becomes even more obvious. Communication in the virtual world has become more than simply using a different tool to transfer the same information as with more traditional tools. Web 2.0 has created new genres, new identities, and users need new forms of literacy to interpret information. Therefore, as Warschauer (2004) points out, new teaching methods are required. Shetzer and Warschauer (2000) define electronic literacy as a threefold competence of communication, construction and research. These three aspects are still applicable for Web 2.0 tools, but they are more combined, for example, in blogs, which are simultaneously used for communication, construction and research. Bloggers construct content, are part of a discourse community, and blogs serve as an important source of information.

The approach of multiliteracies, in development by the London Group since 1996, takes the idea of electronic literacy one step further, by including intercultural communicative competence as an essential goal of electronic literacy. This concept sees hypertext as a nonlinear, multimodal environment, where information is "variously coded in animation, symbols, print text, photos, movie clips, or three-dimensional and manoeuvrable graphics" (Luke, 2000, p. 72). These hypertexts are embedded in local cultural diversifications which must be understood by the reader for successful communication. In Web 2.0, which networks people from all over the world, intercultural communicative competence as an integral part of multiliteracies is of great importance. Consequently, foreign language learners must be prepared for these multimodal environments, as speaking the language without understanding the multimodal contexts in which it is embedded, is not sufficient for global communication.

The nationwide longterm-study, JIM, regularly collects data about recreational and media behaviour of young people from 12 to 19 years of age in Germany. In 2007, almost every household in Germany had computer and Internet access, and 67% of teenagers had their own computer; almost every teenager was using computers. The daily use of the Internet increased from 51% in 2004 to 77% in 2007. One third of the participants state that they contribute actively and regularly to some kind of interaction on Web 2.0 (Abfalterer, 2007). These figures emphasize that young people in Germany are already regularly participating in online communication in Web 2.0. These young people are part of global social networking and need corresponding literacy to communicate successfully. Therefore, technology need not assist language learning foremost, but with Web 2.0, technology created new communities in which learners can practice and for which they must be prepared. This requires more than language learning as a linguistic skill, but language learning in the sense of multiliteracies.

THE POTENTIAL OF WEBLOGS FOR THE CLASSROOM

Weblogs are a good example for what is meant by multiliteracies in the context of Web 2.0. Readers of blogs need more than the ability to understand the language in order to really make sense of the content of many blog posts. They also need the skill of understanding the related discourse with other blogs or linked references to truly comprehend the context of the texts. Furthermore, the competence of understanding the writer's cultural context is important. However, many of these aspects are also true for other literary texts — what, then, has changed? The difference is that Web 2.0 is a Read/Write Web, and the reader can easily interact with the writer or can also become an author. Blog readers can write comments or react by posting responses. They become part of a discourse community and interact in a complex multimodal setting. Using weblogs in the language classroom does not simply mean having a nice tool with which to practice writing, but rather is an opportunity to prepare students for communities of practice connected to the use of weblogs or other Web 2.0 tools in real life.

Weblogs have been used in many classrooms in different ways, and a variety of classifications can be found as to how weblogs can help in acquiring a language. Campbell (2003) mentions *tutor blog*, *learner blog*, and the *class blog* as different approaches for using weblogs in the foreign language classroom. Whereas Campbell's classification distinguishes the writers and their purpose for using a weblog, Richardson (2006) analyses the use of weblogs in the classroom in terms of pedagogy. According to Richardson, weblogs are used in schools for a wide range of purposes: a weblog can be put to use as a *class portal,* an *online filing cabinet,* an *e-portfolio,* a *collaborative space,* for *knowledge management and articulation* and as a *school website*.

Richardson points out that one key feature that distinguishes weblogs and the Read/Write Web from more traditional media is that of the potential audience. Here we can see a significant shift compared to traditional media because collaboration with an audience is not bound to the classroom anymore. By means of weblogs, the classroom can be extended to dimensions previously not possible. Ward (2004) points out different benefits of using weblogs in the writing classroom. A genuine audience is one important aspect that can motivate students in the writing process. He quotes Kitzmann (2003), who writes that "the [online] audience is not only anticipated but expected, and thus influences and structures the very manner in which the writer articulates, composes, and distributes the self-document" (p. 1). Thus, the audience encourages writers to present and express themselves.

Furthermore, weblogs not only provide the audience and therefore change the way learners see their products, but they also change the way content is being constructed. Weblogs often belong to a network of writers functioning as a collaborative blog, in which the authors edit each other's texts. Consequently, texts refer and react to other authors, which means that new content is constructed through collaboration. These new ways of constructing content demand of the learners new literacies, as discussed above. Some of the basic criteria of these multiliteracies is that writing is enbedded in an interactive dialogue between the writer and the audience: "The differences between blogging in this manner and writing as we traditionally think of it are clear: Writing stops; blogging continues. Writing is inside; blogging is outside. Writing is monologue; blogging is conversation. Writing is thesis; blogging is synthesis" (Richardson, 2006, p. 31). The contradiction between writing and blogging, which the author points out, might lead one to the conclusion that blogging is not even writing. In the following sections it will be argued that blogging should not be seen as a contradiction to writing, but rather as a certain form of writing, namely writing as social interaction.

Wrede (2003) puts the aspect of discourse in weblogs this way: "[W]eblogs are usually a form of writing in public and with the intention to offer opportunities for communication. A weblog is a constant invitation for conversation – directly and indirectly" (p. 2). In fact, a weblog writer often has a number of different audiences simultaneously: the group of people the writer is collaborating with, the audience the product is presented to, and the wider audience of the Internet. Each audience cannot only just read the text, but write a comment or even an article. Thus, discourse can happen on different levels and can reach a degree of authenticity which would not be achieved without extending the classroom through online networking of that kind.

As we have seen, weblogs can be used in the foreign language classroom for interactive language learning following a sociocultural paradigm. However, the major change weblogs have brought for language teaching goes beyond that. Weblogs have, in company with other Web 2.0 tools, created new genres and new communities of practice which demand new literacies. Therefore, they have not just added some methods to foreign language teaching, but they have transformed the goals by creating new contexts of communication.

TEACHING WRITING WITH WEBLOGS

New communities of practice with new genres demand specific ways of teaching and learning literacies which help the learner to interact in these contexts. With weblogs in particular, new communities of discourse writing have emerged and the question is what kind of approach for teaching writing could support learners in acquiring the respective literacies.

As Hyland (2002) points out, three different approaches to researching and teaching writing can be identified. The first approach can be described to see texts as autonomous objects, referring to structuralism. The focus in this approach is on the correct arrangement of elements, and the idea of language learning is based on "an autonomous mechanism which depends neither on particular writers or readers, but on setting out ideas using correct forms" (Hyland, 2002, p. 6).

The second approach focuses on the writer and the process of creating texts. Learning writing is a process which can be encouraged by providing writers "with the space to make their own meanings through an encouraging, positive, and cooperative environment with minimal interference" (Hyland, 2002, p.23). Since weblogs provide this open space for writer-oriented creativity, they can be used in language learning for such writing processes. However, more traditional media, such as paper journals, can provide this space also, thus it is not this aspect of weblogs which makes them an exclusive and new tool for teaching the writing process.

It is because weblogs fulfill the requirements of the third approach which traditional media cannot easily satisfy, to provide a tool for writing as social interaction, that they can be considered novel and unique. This third model considers that a writer always has a certain purpose and audience in mind when writing a text. Either the audience is directly addressed through the text (e.g. in a letter) or the audience is invoked, meaning that it is meant to read a certain text although it is not addressed directly (e.g. a novel). A text is always about sharing or negotiating meaning with an audience; if there were no audience, there would be no reason to write a text. The writer is influenced by the addressed or invoked audience, which means that there is an interaction between the writer and the reader. Even though this interaction might not be too obvious in many cases, it is an important factor by which the writing process is influenced.

Connected to the notion of audience is the idea of social construction. The writer is a member of a community, and writing is understood as discourse in this certain community. The way we think and communicate is seen as "language constructs generated by knowledge communities and used by them to maintain coherence" (Hyland, 2002, p. 41). Thus, each part of writing happens in a context of a social community, aiming to construct meaning within this community. Writers "position themselves and their ideas in relation to other ideas and texts in their communities and this helps them both to legitimate their membership and establish their individual identities through discourse" (Hyland, 2002, p. 41). This social interaction characterizes the writing processes in weblogs, as described above. The audience for weblog writers is obvious and the blogging community a real, existing community. Therefore, compared to other writing tools traditionally used in class, weblogs have the potential to extend the audience beyond the classroom and to create new writing communities.

Grabe and Kaplan point out that "audience is essential to the creation of text and the generation of meaning" (Grabe & Kaplan, 1996, p. 207). They mention five parameters by which the writer of a text is influenced with regard to the reader: the number of persons who are expected to read the text, the extent to which readers are known or unknown, the difference of status, shared background knowledge and shared knowledge of the topic at hand. Through these aspects, the identity of the discourse community is defined and the patterns of discourse established. It is very likely that, for instance, the discourse between students in a project at university will differ from pupils at a primary school who work on a project. However, within these communities shared patterns of communication help to communicate and negotiate meaning.

Hedge (2000) sees the matter of audience as an important aspect to foster good writing. Real-life audience for her is a precondition for developing real-life writing tasks. In the context of the task-based language learning classroom these real-life writing tasks are of great importance, since they promote meaningful communication.

Teaching writing as interaction in a discourse community can be one way of implementing the sociocultural approach to the foreign language classroom. Swain (2000) mentions that research suggests that comprehensible input alone cannot provide opportunities for language acquisition. Moreover, the role of interaction with its components of input and output in collaborative dialogue constitute language learning.

Lantolf (2000) points out that the central concept of sociocultural theory is the *mediation* of higher forms of mental activity. In second language learning this mediation takes place with others through social interaction, with oneself through private speech, or by means of artefacts like tasks and technology. Sociocultural theorists do not draw a clear distinction between "use" of a second language and "knowledge" of a second language, as in their view use creates knowledge (Ellis, 2003). A central means of mediation is verbal interaction by creating situations in which novices can negotiate meaning and thus participate in their own learning. The expert can function as providing support in order to help the learners reach the next level or understand a certain language structure they need for interaction. This scaffolding is important for reaching the next potential level of development, which Vygotsky (1978) called "the zone of proximal development."

Teaching and learning writing using a sociocultural approach means providing learners with opportunities to engage in collaborative discourse communities with the goal of social interaction. The assumption is that weblogs provide such environments for second or foreign language learners. Moreover, the blogosphere is an authentic community of social interaction through discourse writing. Bonk and King (1998) developed a collaborative writing taxonomy for such electronic writing environments and attempt to describe how electronic writing tools could be used in the context of a sociocultural approach. However, Bonk and King (1998) note that in terms of research, many questions remain unexplored. One of these questions is, "How do different interaction structures and collaboration formats impact student writing?" (Bonk & King, 1998, p. 6). The assumption is that weblogs inherit a purpose for interaction and therefore the impact on student writing should be one that supports connective writing. Some of the findings of the following qualitative study, researching the influence of the online audience on the students' writing process, imply that the collaboration format of weblogs does influence students' writing.

A STUDY ON THE INFLUENCE OF AUDIENCE ON WRITING IN WEBLOGS

Research Design

The study consists of 29 single cases that had the same task to accomplish: writing a reading journal. These single cases were compared to each other with the focus on differences between the writers of weblogs and the ones with paper reading journals. Since different single cases were compared in different sections, the research can be seen as a comparative study (Flick, 2003).

The study explores how an online audience influences students in their writing of a reading journal. The assumption is that a real online audience in the context of a discourse community has certain effects on students' writing. Therefore, in terms of meaningful communication, the outcome of weblog reading journals written for an immediate audience in a discourse community should differ from reading journals written on paper for an abstract audience. The expected outcome of the study was to find out in how far students realize the interactive character of the Web 2.0 and whether this changes their writing in terms of connective writing. The research questions of this study are as follows:

- To what extent do weblogs support meaningful writing?
- What differences are there in students' writing between weblogs and paper journals?
- To what extent does the online audience of a discourse community influence students in their writing of a reading journal?

The participants consisted of 29 students of a grade 9 secondary girl's school in Germany who had English as a foreign language in their fifth year. According to the Common European Framework the goal set by the state curriculum for this grade is to reach level A2 which means they should be basic users of English as a foreign language on a way stage level (Council of Europe, 2001). The test scores on a class level showed that most students met the criteria of A2, some students are slightly beneath A2 and few even scored B1 (independent user on a threshold level). The students could choose whether they wanted to write a paper journal or create a weblog on the Internet. 10 students decided to write a weblog and 19 students wrote a paper journal. It was for the first time that the students did a reading journal. None of the students had worked with a weblog before.

The participants were assigned the task of writing a reading journal about the book *If You Come Softly* by Jaqueline Woodson. The reading process was supported in class with pre- while- and post-reading tasks. They were instructed to write their thoughts and impressions about the chapters they had read by using texts, photos, drawings, articles or poems. Each student had to write at least one post about each chapter during a period of six weeks. The students could choose what they wanted to write about and how they wanted to comment on issues they considered relevant. The blogging software used for the project was developer-hosted blogs on blogger.com. The advantage of blogger.com is that creating a blog is easy and free, without downloading software and hosting by the user. Furthermore, password protected communities can be established. But there are other providers offering similar features, like wordpress.com.

DATA COLLECTION AND ANALYSIS

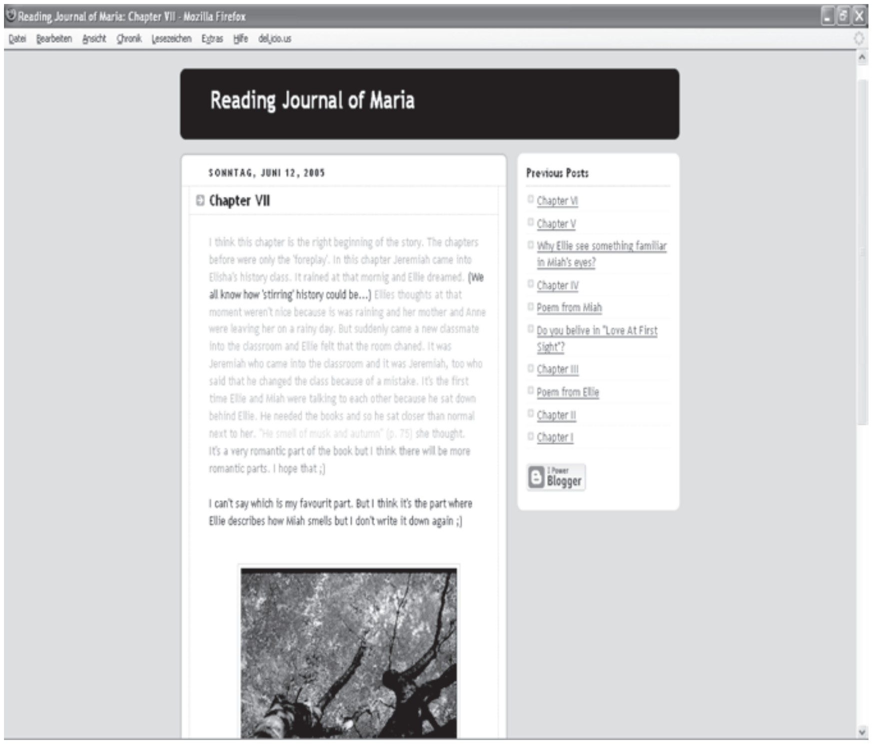
The data consisted of the students' reading journals (weblogs and on paper), two questionnaires, and a guided interview. The first questionnaire was done before the students started to write their reading journals; the second questionnaire and subsequent interviews, after they had finished their journals. The data was analysed using Glaser's (1998) grounded theory.

First Questionnaire

In the first questionnaire the type of questions was mostly open, following the qualitative paradigm. In response to the question regarding the students' underlying motivation for choosing either a weblog on the one hand or a pen-and-paper journal on the other, three different categories emerged: (a) Personal preferences, (b) Computer Issues and (c) Audience.

Category (a) summarizes students' general statement, e.g. "I decided to work with a weblog because I have never done it before." Categories (b) and (c) stand for two single topics that appeared in most answers. Either students saw their decisions in the context of computer issues (e.g. "I decided not to work with a weblog because we don't have Internet access at home") or in the context of the audience connected to weblogs (e.g. "I didn't choose weblogs because I don't want to publish anything" or "I chose the weblog because I like the idea that everyone could read my texts"). Of

Figure 1. Screenshot of a student's weblog



particular interest for the research project is category (c), because it reveals that many students were aware of the potential online audience, regardless of whether they considered the audience as motivating or threatening. Another remarkable fact is that in the answers none of the students associated pen and paper journals with an audience. In general, the findings of the first questionnaire suggest that the students had a general awareness of audience that they associated with weblogs but not with paper journals.

Reading Journals

The data analysis of the reading journals referred to the first research question. The main concern when analysing the data of the reading journal was to identify any differences between the texts written in weblogs and the ones written on paper. Through the open coding process, following the *grounded theory*, the reading journal texts were analysed to find distinctive features. The findings suggest that in general the students' texts can be classified in three different categories:

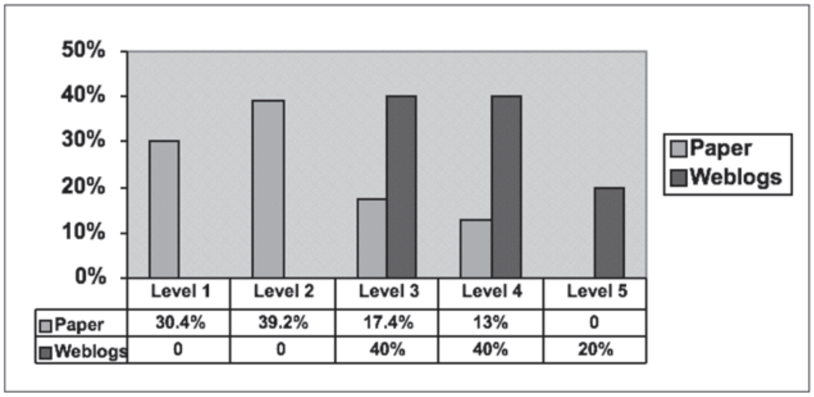
Category 1: Summaries of the Chapters

This category represents students who only wrote summaries of each chapter. The main goal of the writer is to summarize the most important things that happened in the story or the respective chapter.

Category 2: Summaries of the Chapters with Added Opinion

This category represents students who wrote summaries of each chapter and added their personal opinion to the summary. The main goal of the writer is to summarize the most important things that happened and then to add the personal opinion. The amount of text devoted to the expression

Figure 2. Findings of reading journal analysis



of personal opinion is in relation to the amount devoted to summaries much smaller.

Category 3: Personal Comment with Summary Included

Students who wrote personal comments on each chapter fall into this category. These comments consist of utterances of personal opinion and feelings. They can contain summaries of the chapters as well, but in the light of the personal perception of the writer.

Using the above categories, the data was analysed a second time with the intention of developing a concept which could help describe the distinctive features in a more detailed way. In particular, the issue of writing from a personal perspective versus writing summaries should be at the centre of the analysis. For that purpose descriptors were developed which describe the differences between the students' reading journals more precisely.

In the last phase of the coding process each reading journal was classified into one of the described levels. Figure 2 summarizes the findings.

The numbers show the percentage of each comparison group for every level. The findings show that the difference between paper journal students and weblog students is significant. The students with weblogs all reached at least levels 3 to 5. On the other hand, only 30 percent of the students with paper reading journals reached these levels. Additionally, none of the paper reading journal students reached level 5 and only 13 percent reached level 4. However, 60 percent of the students with weblogs were in levels 4 and 5.

For clarification, it has to be added that these figures do not indicate anything about the accuracy of the students' texts. Nevertheless, the summaries show clearly identifiable trends with regard to content; students who wrote weblogs expressed to a much higher degree their own opinion, attitudes and personal thoughts, as defined in the descriptor. The language of the "weblog students" was not necessarily better in terms of language accuracy. However, it was more personal than the language of the students with paper journals. One can say that the students with weblogs talked more about themselves, they wanted to communicate a message. To find out why they wanted to communicate on a more personal level, we will have to look at the data collected in the second questionnaire and the interviews.

Second Questionnaire

The students completed the second questionnaire after they had written their reading journals. They

Table 1. Levels of student interaction

Level 1	Writes only short summaries of the readings. No personal reflections or expression of attitudes, no utterance of opinion. Example: Selina (paper journal): Chapter 10 Jeremiah remembers his childhood and his father's new girlfriend Lois Ann. It makes him sad, especially that her parents could hurt each other. Jeremiah is thinking about Ellie, he showed his feelings. And he thinks about his life and it makes him sad, because so much happened.
Level 2	Writes mainly summaries of the readings with some personal reflections. In relation to the summaries the personal reflections is very brief and lacks expression of attitudes. Personal opinion is expressed, but only in short, isolated sentences. Example: Theresa (paper journal): Chapter 10 Loneliness is in the air. Jeremiah is sitting in his mother's room looking at pictures of her and his Dad. Thinking about how it was in the past when the family was still together. Now his parents are separated, because his dad left them. He went to another woman. Lois Ann. If my parents were separated a world would crash down for me. Jeremiah is also thinking about Ellie. He's thinking about going to kiss her. Soon. It's sweet to read this, but how does he know if Ellie likes kissing him? I think he'll find out …
Level 3	Writes summaries of the readings and personal reflections. The amount of personal utterances is significant but still less than the summaries. Attitudes and personal opinion are mentioned frequently and sometimes in detail. Example: Lisa B. (paper journal): Chapter 10 This chapter is about Jeremiah. Jeremiah explains how he felt when he saw his dad with his new wife Lois Ann. He always saw the picture from his mother when she married his father. He thought back at the marriage and that his parents thought their love will be forever. Jeremiah said that he sometimes want a brother or a sister. He also thought about Elisha and her smile. Jeremiah looked at the house and noticed how empty it was and that the house echoed when he was speaking. I know that it is hard when your dad or mama had an new wife or husband but you must except it whatever happens, but not every person will except it because it's hard. I think Jeremiah is in this situation but I also think he except it, because he loves his daddy deep in his heart what ever happens.
Level 4	The amount of summary compared to the amount of personal reflections is about equal. Attitudes and personal thoughts are expressed frequently and in detail. Personal opinion is expressed by developing arguments consisting of several sentences. Alena (weblog): Chapter 10 This Chapter is very hard. There are a lots of feelings I can't really discribe just understand. Jeremiah is in his mothers room, there are photos and he look at them. There are old photos but the most important photo is the picture with his mum in a wedding dress. He cry, he imagine whats happend wrong. What happend when he was little and can't understand. A long time ago the father left his mum but Jeremiah couldn't understand he was just 12 or 13, he smiled because he couldn't understand that it will be a hard time. But now he understand all what's happened. I think he hate Ann Lois, I think he hate his father too. But did Jeremiah hate him really? I think he want to hate him but he can't, he left his mother but not himself. There is a thing, a little thing which nobody see. On the next day his parents would be 17 years married. I think he tought at this and how it could be when his parents where together now. I think when two people separate each other, they haven't got a future because when the love goes the love can't come again. Jeremiah dream, he dream that his parents will be together one day, without Ann Lois. But maybe there is a light:) I call the light Ellie, maybe she can show him the way in the happyness. I think he love Ellie really, he has longing at her because he want to tell her all what's happened and makes him sad.

were asked about their experiences with writing the journals and in particular their motivation and what kind of audience they had in mind while writing.

The analysis of the second questionnaire reveals, among other things, two important categories related to the relationship between the writing process and the potential audience.

The first category (A) refers to the kind of audience writers had in mind, whereas the second category (B) is concerned with the influence the audience had on the writers. Each category has been divided into three subcategories that can be described as follows. A table displaying the students' answers and a brief analysis of the answers follows the description of each category.

Table 1. Levels of student interaction

Level 5	The amount of summary compared to the amount of personal reflections is at least equal. The summaries are written in the context of a personal perspective. Attitudes and personal opinion are expressed frequently and in detail. The personal opinion is expressed by developing arguments including several sentences. Maria (weblog): Chapter 10 It's a very sad chapter and I … I don't know … I was shocked. I stared for a few minutes at the last word and thought nothing. I still don't know what I should think. Oh yes … I should describe what made me so shocked. It was because of Jeremiah. He was in his mother's room and looked at the pictures wich stood on her dresser. There was a picture from his mother in a wedding dress and she smiled and looked happy. Very happy. When he looked at this picture he thought about the relationship between his mom and his dad. They were nearly seventeen years undivorced. It was a long time but they only had one child – Jeremiah. He felt very lonely but he wouldn't like a sister or a brother. "He wanted more than that – somone deep. Somebody who could know him -know all of him- the crazy things he dreamed on stormy nights, when he woke with tears in his eyes and pulled the covers tight around him" (p. 100). Then he thought about Ellie. Ellie was there in his head and didn't go away. I think they need each other. Both need someone who's there for the other person and who knows all about the other person. And if it's only for a short time. (now I think so because the book isn't very long and we are at chapter X already and not much happened. Oh … I deviate from the description. Yes. He thought about Ellie and how much he needs her. It was a depressing situation: the empty house, his oppressive situation with his mom and his dad, his pain about all the discrimination. It was too much and at the end he cried. The sweet part in this chapter: "I'm going to kiss you soon, Jeremiah had found himself thinking. I don't know when or where or how, but soon I'm going to kiss you" (p. 101). I love this part because it's so sweet and…I don't know an other word for this sentence. But I liked the part that I put in my thoughts (the other blue sentence), too.

Category A (Kind of audience)

- **A1 The teacher:** Students who mentioned that they were aware of the teacher as audience.
- **A2 The teacher, fellow students and friends:** Students who mentioned that they were aware of the teacher, fellow students and friends as potential readers.
- **A3 The online community:** Students who mentioned that they were aware of the online community as potential readers.

The figures in Table 2 show a clear difference between the students who wrote paper journals and those with weblogs. The ones who wrote paper journals had either only the teacher, or the teacher and classmates, or friends in mind while writing. The students with weblogs were up to 90% aware of the audience in the online community. Although the weblog students knew that the teacher would read their blogs, none of them mentioned the teacher as reader. This implies that the students associated weblogs strongly with the online community connected to them.

Table 2. Findings category A

Audience Category	Total	Paper Journal		Weblog	
Category A1	8	8	44,4%	0	0,0%
Category A2	11	10	55,6%	1	10,0%
Category A3	9	0	0,0%	9	90,0%

Table 3. Findings category B

Influence Category	Total	Paper Journal		Weblog	
Category B1	14	12	66,7%	2	20,0%
Category B2	11	6	33,3%	5	50,0%
Category B3	3	0	0,0%	3	30,0%

Category B (Influence of Audience)

- **Category B1: No Influence, no comment.** Students who made it clear that they think the audience they had in mind did not influence their writing at all. Students who didn't make any comment are included as well.
- **Category B2: Making it interesting and understandable.** Students who mentioned that they tried to write accurately, so that others will be able to understand their texts. Furthermore, many said in the same context they wanted to make the texts interesting, because they had in mind that someone would read the texts.
- **Category B3: Writing personally.** Students who answered that they tried to write on a personal level. Some mentioned that this was meant to express attitudes, others wrote that they wanted to tell their opinion.

Table 3 indicates a tendency towards a greater influence of the audience on writers of weblogs than writers of paper journals. 66,7% of the students with paper journals either negated the influence of audience on their writing or did not mention any influence (category B1), while only 20% of the weblog writers were classified in this category. In category 2, more weblog writers (50%) than paper journal writers (33,3%) mentioned that, because of the audience, they wanted to make the journal more interesting or understandable. In category 3 the difference is even clearer. 30% of the weblog writers think that the audience makes them write more personally, but none of the paper journal writers.

To summarize the analysis, we can say that 80% of the weblog writers see an influence of the audience on their writing, but only 33,3% of the paper journal writers can see any influence of the audience on their writing process.

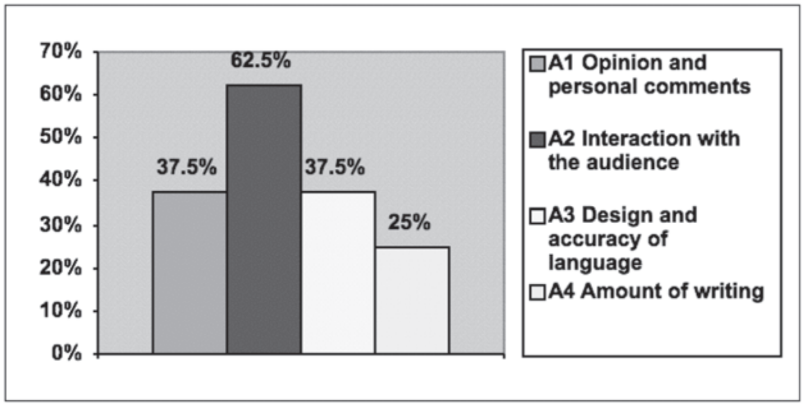
Guided Interviews

In the guided interviews the students were confronted with, amongst others, the observation that the weblog students wrote more personal comments and showed a higher degree of reflective writing. The students were asked to comment on these findings and come up with reasons for the differences. The open coding process was done separately for each comparison group. Therefore, the categories for the weblog students and the paper journal students are different.

Students with Weblogs (Category A)

- **Category A1: Opinion and personal reflections.** Because they knew that other people would read their weblogs, they wrote more about their opinion and showed personal reflections. These were typical answers of students who can be categorized in category A1. They said they wanted to make their weblog personal and make the reader understand what they think about certain parts of the book.
- **Category A2: Interaction with the audience.** Category A2 is in a way similar to category A1 with regard to personal communication. However, it focuses on answers in which students talk about interaction with the audience. This possibility of communicating and interacting with the audience caused them to write in a personal manner and to negotiate meaning.
- **Category A3: Design and accuracy of language.** Some students described how their consciousness of audience motivated them to pay more attention to form, i.e. either to formally correct language and/or to the appearance of the weblogs themselves. The two aspects are put together, since they both deal with form rather than with content.
- **Category A4: Amount of writing.** In addition to other influences that the weblog

Figure 3. Findings guided interviews weblogs



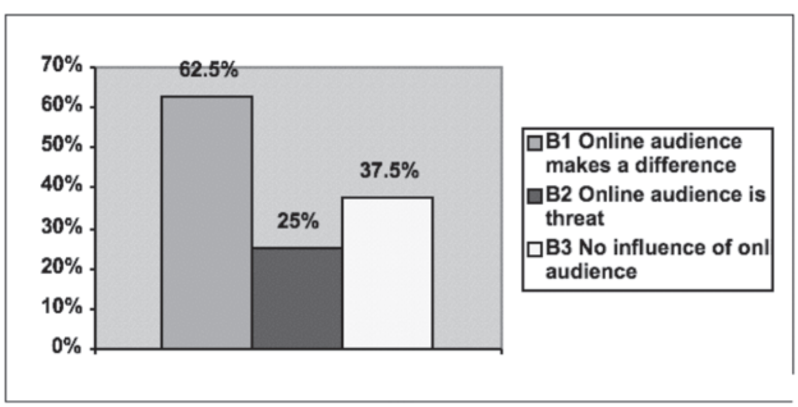
authors' awareness of the audience had on their weblogs, the weblog authors also tended to produce a larger amount of text. Answers in this category implied that the students thought they wrote more because they were aware of the fact that someone was actually going to be reading their weblogs.

The findings of the interviews with weblog students (Figure 3) indicate that most students in this comparison group had an awareness of the online audience and were convinced that these potential readers influenced their style of writing. Moreover, 62,5% mentioned that they meant to interact with the audience when writing posts.

Students with Paper Journals (Category B)

- **Category B1: Online audience makes a difference.** Although they did not experience an online audience for themselves when writing a reading journal, these students mentioned in the interviews that they think an online audience makes a difference concerning the content of writing. They based this assumption on observations they made on the weblogs of their classmates.

Figure 4. Findings guided interviews paper journals



- **Category B2: Online audience is a threat.** Some students see the online audience as a threat. They did not want anybody to read their texts; therefore, they didn't use weblogs themselves. Although most of them did not specify reasons for that fear, they would feel uncomfortable with an online audience in mind.
- **Category B3: No influence of online audience.** Category B3 is comprised of students who mentioned that they don't see an influence of the online community on the writing process. They asserted that there is no difference between weblogs and paper journals with regard to audience influence.

Figure 4 shows that 62.5% of the interviewed students with paper journals think an online audience influences content, although they did not experience an online audience themselves. 25% of the interviewed students with paper journals perceived the online audience as a threat. 37.5% of the paper journal students found the online audience had no influence on the writing process. One of the most interesting findings of the guided interviews is the fact that a majority of the paper journal students saw the online audience as a main reason for more reflective and personal writing. These findings confirm the shared notion of the weblog students, the majority of whom showed at least some cognizance of an online audience.

CONCLUSIONS FROM THE STUDY

Awareness of Audience

In summary, we can say that the data show that the students' writing process reflects an awareness of the online audience. Both, weblog writers and paper journal writers mentioned the online audience as an influence either on the decision-making process for or against weblogs, or they saw that the online audience influenced the writing process of the weblog students. This shows that the audience, as is typical for social software applications like weblogs, is something that students are acutely conscious of in the writing process. They are aware of the audience and it influences them in their writing process. Thus, we can speak of a real and immediate audience since it was not constructed or made up by the teacher or through an artificial textbook task ("Imagine you are writing a letter to a friend"). Moreover, it exists independently of the task put to the students. They were not told to imagine an audience or someone who would read their entries: they were automatically aware of the audience by virtue of their familiarity with the weblog medium. These findings support the idea that students associate weblogs directly with an audience as Richardson (2006) and others have pointed out.

Meaningful Communication

Before having a look at the data here, clarification of the term "meaningful communication" is in order. For this purpose, Littlewood's (2000) definition can be helpful. He describes a continuum from non-communicative learning to authentic communication with three categories (pre-communicative language practice, communicative language practice, structured communication) in between. The closer an activity moves towards authentic communication, the more a focus on meaning can be identified. He defines authentic communication as "Using language to communicate in situations where the meanings are unpredictable, e.g. in creative role-play, more complex problem-solving and discussion" (Littlewood, 2000, p. 5).

Which findings here can be associated with this definition of meaningful communication? In this respect, the findings of the reading journal are of interest. Students who wrote weblogs expressed their personal opinions and attitudes to a considerably higher degree than those students writing paper journals. These findings show that students with weblogs shared their opinion and personal attitudes on the book to a greater extent than paper journal students. Since sharing opinion and personal attitudes means communication and negotiation of meaning, students with a higher degree of these characteristics of writing can be classified in Littlewood's categories on a level close to "authentic communication." On the other hand, students who wrote mostly summaries of the chapters just followed a rather pre-communicative language practice, because they did not try to negotiate meaning or to communicate a message, they simply reproduced content. Since students with paper journals could be classified to a much higher level than weblog students into the category of "mostly summary writing," they do not fulfil the criteria of authentic communication to the same degree as weblog students do. The students themselves assumed that the difference is rooted in the online audience that is associated with weblogs. These results support the assumption that weblogs are a new text genre, one of connective writing, by which author and audience communicate with each other.

Community Discourse

Hyland (2002) along with Grabe and Kaplan (1996) see a "discourse community" as an essential aspect of authentic writing. Authentic writing always happens in the context of a social community, aiming to construct meaning within this community. The findings of the research here indicate an awareness of such a social community. The weblog writers' awareness of audience and the higher level of focus on meaning of the weblog students' writing compared to the paper journal students' writing suggest that the weblog students saw themselves as part of a social community in which they wanted to negotiate meaning. In particular in the interviews most weblog students mentioned that they intended to interact with the audience. Hence, they saw themselves as part of a discourse community. We can say that weblog students show a high awareness of a social community they want to interact with.

FUTURE TRENDS

Interaction and collaboration in Web 2.0 are becoming increasingly important in a globalized world and a new kind of social networking through weblogs is one key feature of this change. Users who want to participate in these social networks need the skills to understand multimodal texts. The concept of *multiliteracies*, which combines intercultural communicative competence with electronic literacy, helps to describe the skills learners of a foreign language will need to reach that goal. In the light of these changes, a sociocultural approach in the foreign language classroom will gain greater importance. For the use of weblogs in foreign language education this means that *networking* and interaction can happen within a class or beyond the classroom in collaboration with other classes worldwide. In tele-collaborative projects weblogs can be used for publishing texts, exchanging ideas and perspectives on certain topics, or in a literature project. The sociocultural dimension of language acquisition is a vital characteristic of such projects. However, even though weblogs have the inherent potential to facilitate the kind of interaction that supports the language learning process, this process doesn't automatically come about simply by using the medium. It is important also to consider the aspect of content: if students have nothing to say, it doesn't matter in which medium they have nothing to say. Learners need meaningful, authentic tasks that encourage them to produce meaningful, authentic output. Nunan (2004) points out that the classroom itself always has a pedagogical dimension, but that the goal of task-based language learning is to prepare learners for real-world tasks. Therefore, the basic question is what learners need to do with language and how we can prepare them for these situations.

As Van den Branden (2006) says, "Tasks are supposed to elicit the kinds of communicative behaviour (such as the negotiation of meaning) that naturally arises from performing real-life language tasks, because they are believed to foster language acquisition" (p. 9). Further research has to be undertaken to explore what kind of tasks support social interaction with weblogs and how the students' development of *multiliteracies* can be encouraged by certain settings.

CONCLUSION

As this chapter has shown, Web 2.0 has created new dimensions of communication. Some key aspects of this transformation are social networking, interactive user-generated content and global collaboration. This change has led to new communities in Web 2.0 and new kinds of genres are developing. This in turn requires a different literacy of the learner. Weblogs, for example, represent different text modes: creating user-generated content, interaction between the author and readers, discourse between different authors and information on a certain topic. Furthermore, weblogs are contextualized in a certain cultural setting which the reader needs to be aware of to understand the implication of texts. These different modes require *multiliteracies*. The study has shown that students were aware of the discourse community connected to weblogs. Their texts showed a higher amount of connective writing, meaning they had an audience in mind to interact with. This supports the assumption that weblogs have created a new genre and that users are aware of the multimodal levels connected to blogging. In summary we can say that Web 2.0 has transformed writing, in particular writing in weblogs. Moreover, Web 2.0 has changed language learning because speakers of a foreign language already use Web 2.0 to communicate meaning and generate content in new genres, by using the foreign language as a lingua franca. Therefore, the community of language practice is already existent. The challenge of foreign language education will be in how far teachers realize these changes and prepare learners for these new environments of language practice. There is a variety of possibilities to work with weblogs in foreign language education and create opportunities for authentic language practice.

REFERENCES

Abfalterer, E. (2007). *Foren, Wikis, Weblogs und Chats im Unterricht.* Boizenburg: Werner Hülsbusch.

Alby, T. (2007). *Web 2.0 Konzepte, Anwendungen, Technologien*. München: Habser.

Bonk, C. J., & King, K. S. (1998). *Electronic collaborators.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Branden, Van den K. (2006). *Task-based language education: From theory to practice*. Cambridge: Cambridge University Press.

Campbell, A. (2003). Weblogs for use with ESL Classes. *The Internet TESL Journal*, Retrieved September 12, 2007, from http://iteslj.org/Techniques/Campbell-Weblogs.html

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford: University Press.

Flick, U., von Kardoff, E., & Steinke, I. (Eds.). (2003). *Qualitative Forschung.* Reinbeck bei Hamburg: Rowohlt.

Glaser, B. (1998): *Grounded theory: Strategien qualitativer Forschung.* Bern: Verlag Hans Huber.

Grabe, W., & Kaplan, R. W. (1996). *Theory and practice of writing.* London: Longman.

Hedge, T. (2000). *Teaching and learning in the language classroom*. Oxford: Oxford University Press.

Hyland, K. (2002). *Teaching and researching writing*. London: Pearson Education.

Kern, R., & Warschauer, M. (2000). Theory and practice of network-based language teaching. In M. Warschauer & R. Kern (Eds.), *Network-based language teaching: Concepts and practice* (pp. 1-19). New York: Cambridge University Press.

Kitzmann, A. (2003). That different place: Documenting the self within online environments. *Biography*, *26*(1), 48–65. doi:10.1353/bio.2003.0026

Lantolf, J. P. (Ed.). (2000). *Sociocultural theory and second language learning*. Oxford: Oxford University Press.

Littlewood, W. (2000). Task-based learning of grammar. *Teaching and Learning Update*, *1*, 40–57.

Luke, C. (2000). Cyberschooling and technological change: Multiliteracies for new times. In M. Kalantizis & B. Cope (Eds.), *Multiliteracies: Literacy learning and the design of social futures* (pp. 69-91). New York: Routledge.

Möller, E. (2005). *Die heimliche Medienrevolution*. Hannover: Heise.

Nunan, D. (2004). *Task-based language teaching*. Cambridge: Cambridge Univ. Press

O'Reilly, T. (2005). *What is Web 2.0: Design patterns and business models for the next generation of software*. Retrieved January 20, 2008, from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Richardson, W. (2006). *Blogs, Wikis, podcasts, and other powerful Web tools for classrooms.* Thousand Oaks, CA: Corwin Press.

Shetzer, H., & Warschauer, M. (2000). An electronic literacy approach to network-based language teaching. In M. Warschauer & R. Kern (Eds.), *Network-based language teaching: Concepts and practice* (pp. 171-185). New York: Cambridge University Press.

Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J. P. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 97-114). Oxford: Oxford University Press.

The State of the Live Web. (2001). Retrieved June 18, 2008, from http://www.sifry.com/alerts/archives/000493.html

Vygotsky, L. (1978). Interaction between Learning and Development. In M. Cole, *Mind in society* (pp. 79-91). Cambridge, MA: Harvard University Press.

Ward, J. M. (2004). Blog Assisted Language Learning (BALL): Push the button publishing for pupils. *TEFL Web Journal*, *3*(1), 1–16.

Warschauer, M. (2004). Technological change and the future of CALL. In S. Fotos & C. Brown (Eds.), *New perspectives on CALL for second and foreign language classrooms* (pp. 15-25). Mahwah, NJ: Lawrence Erlbaum Associates.

Warschauer, M. (2005). Sociocultural perspectives on CALL. In J. Egbert & G. M.Petrie (Eds.), *CALL Research Perspectives* (pp. 41-51). Mahwah, NJ: Lawrence Earlbaum.

Wrede, O. (2003). Weblogs and discourse: Weblogs as transformational technology for higher education and academic research. *Blogtalk conference paper Vienna*, May 23rd 2003. Retrieved September 12, 2007, from http://wrede.interfacedesign.org/articles/weblogs_and_discourse.html#24

KEY TERMS AND DEFINITIONS

Blogosphere: The term relates to the perceived network that joins all weblogs on the Internet together in one community.

Collective Intelligence: A form of intelligence that emerges from a community of individuals who collaborate together. It is an approach to working on products such as texts, documents, codes, decisions with no centralized hierarchy. One central idea is that the collective product of a community is more than just the sum of the individual parts.

Community of Practice (CoP): A group of individuals who engage in and contribute to the practices of their communities through active participation and therefore share a common identity. The term *community of practice* was created by Etienne Wenger and Jean Lave in 1991, who positioned learning in the context of social interaction. One substantial part of knowledge acquisition in *communities of practice* is the construction of knowledge through participation in a community.

Discourse Community: This term connects the notion of discourse (typically relating to numerous forms of communication) with a group of users, usually on a specific subject or area of interest. A discourse community might be used to describe a particular group where members meet to discuss topics of specific interest to them.

Electronic Literacy: The ability to read and write in an electronic medium and to find, organize and make use of information in the context of a hypertext environment. Electronic literacy combines texts and other media, has a focus on collaboration and includes the use of online sources.

Multiliteracies: The term deals with the complexity of language in two major aspects: first, the multimodality of texts through the increasing importance of the written word as part of visual, audio and spatial patterns, and second the cultural and linguistic diversity through global connectedness.

Sociocultural Approach to Language Learning: This approach derives from sociocultural theory that sees learning as the mediation of higher forms of mental activity through interaction. A central means of mediation is verbal interaction by creating situations in which novices can negotiate meaning and thus participate in their own learning. The expert can function by providing support in order to help the learners reach the next level or understand a certain language structure they need for interaction.

This work was previously published in Handbook of Research on Web 2.0 and Second Language Learning, edited by M. Thomas, pp. 274-291, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.24
The Effects of Web-Enabled Self-Regulated Learning and Problem-Based Learning with Initiation on Students' Computing Skills

Pei-Di Shen
Ming Chuan University, Taiwan

ABSTRACT

Vocational degree earners represent a major portion of the work force in Taiwan. However, the reality of computing education in Taiwan's vocational schools is that it's not so practically oriented, revealing much room for improvement. In this context, we conducted a quasi-experiment to examine the effects of applying web-based self-regulated learning, web-based problem-based learning with initiation, and their combination to enhance students' computing skills. Four classes in successive years, with a total of 177 third-year students, were divided into 2 (SRL vs. non-SRL) × 2 (PBL with initiation vs. PBL without initiation) experimental groups. Results were generally positive. Results revealed that the effects of web-enabled SRL, web-enabled PBL with initiation, and their combinations on students' computing skills have significant differences. We hope that the online teaching method applied in this study is also useful for those teachers engaged in e-learning, specifically, in vocational schools. [Article copies are available for purchase from InfoSci-on-Demand.com]

INTRODUCTION

The vocational education system in Taiwan constantly evolves to meet needs such as: the new demands for highly skilled manpower, continued progress of modern technology, worldwide economic development, changing industrial structure, and social/cultural changes. However, vocational education in Taiwan is highly competitive in that it must attract sufficient student enrollment in the face of continually decreasing birth rate and rapidly increasing number of schools. Students' technical skills and the number of professional certifications earned are the main criteria when judging teachers' teaching performance and students' learning effects. However, students in these schools tend to have lower levels of academic achievement. They have low interest and negative attitude toward their learning (Chen & Tien, 2005), spend more time on part-time jobs, do not adequately get involved in their schoolwork, and don't care so much about their grades (Shen, Lee, & Tsai, 2007a). Teaching in such a context, particularly teaching courses in application software – with the target on earning certificates – is a great challenge to most educators.

Web-based instruction seems to be an ideal learning environment because students can access an almost unlimited amount of information and apply it in multiple ways (Kauffman, 2004). However, implementing e-learning for students with low self-regulatory skills inevitably runs high risks. It is a big challenge for teachers to help college students, who are often addicted to the Internet, engage in an online course in an environment with filled with millions of interesting websites, free online games, and online messenger. This addiction to the Internet and the lack of on-the-spot teacher monitoring in web-based instruction makes it even more difficult for students to concentrate on online learning. Moreover, teachers generally feel that students' lack of time management skills is the greatest problem and obstacle to learning in virtual environments. However, the students do not perceive lack of time management as a problem (Löfström & Nevgi, 2007). In this context, it is very important to develop students' skills of self-regulated learning (SRL) to manage their learning in web-based learning environments (Winnips, 2000). Therefore, SRL was applied in this study to help students develop regular learning habits.

Courses in application software traditionally emphasize memorization by applying short, disjointed, lack-of-context examples. Even the professors in National Open University in Taiwan who teach on-the-job students Microsoft Office through television also tend to use short, inappropriate examples in their curricula. The lack-of-context examples in textbooks and used by lecturers may result in uncompetitive employees. There is a gap between what is learned in school and what is required in the workplace (Wu, 2000). In this regard, the computing education in vocational schools in Taiwan can hardly be deemed as effective. In order to increase students' learning motivation and to develop practical skills, problem-based learning (PBL) is considered to be one of the most appropriate solutions. PBL uses real-world, simulated, contextualized problems of practice to motivate, focus and initiate content learning and skill development (Boud & Feletti, 1991; Bruer, 1993; Williams, 1993). It is believed that PBL would help less academically inclined students to develop practical computing skills.

However, students with low academic achievement usually lack the ability to seek essential information and solve the problems they face, particularly in a web-based course without the teacher's on-the-spot assistance and monitoring. Thus, in a PBL environment, these students have to climb a stiff learning curve and overcome much resistance that might pose challenges to both instructors and students, particularly in the initial stage. This stiff learning curve may become a bottleneck and limit the potential effects of PBL. In this regard, we believe that a teacher should provide assistance to his/her students as they adapt to PBL. For example, establishing students' background knowledge and developing required skills is especially important before asking students to solve simulated problems. This could help students be more confident and more involved in the PBL environment and expand the effects of PBL.

As more and more institutions of higher education provide online courses, the question arises whether they can be as effective as those offered in the traditional classroom format (Shelley, Swartz & Cole, 2007). However, few studies have discussed effective online instructional methods for vocational students (Shen, Lee, & Tsai, 2007a). Moreover, there has been relatively little empirical research on students' SRL with such complex technology-based learning environments (Azevedo & Cromley, 2004). Appropriate contextualization is decisive in making educational software and teaching websites effective; otherwise, the potential of even the best program will remain largely unexploited (Bottino & Robotti, 2007). Therefore, we redesigned a course that focuses on the development of students' skills of web page programming and website planning to integrate innovative teaching methods and learning technologies to help students learn. In this study, we conducted a series of quasi-experiments to examine the effects of web-enabled SRL, PBL with initiation, and their combinations on vocational students' skills in using application software.

SELF-REGULATED LEARNING

Zimmerman and Schunk (1989) define SRL in terms of self-generated thoughts, feelings, and actions, which are systematically oriented towards attainment of students' own goals. SRL is also defined as a learner's intentional efforts to manage and direct complex learning activities and is composed of three primary components, namely: cognitive strategy use, meta-cognitive processing, and motivational beliefs (Kauffman, 2004). In a SRL environment, students take charge of their own learning by choosing and setting goals, using individual strategies in order to monitor, regulate and control the different aspects influencing the learning process and evaluating their actions. Eventually, they become less dependent on others and on the contextual features in a learning situation (Järvelä, Näykki, Laru, & Luokkanen, 2007). Characteristics attributed to self-regulated persons coincide with those attributed to high-performance, high-capacity students, as opposed to those with low performance (or learning disabilities), who show a deficit in these variables (Reyero & Tourón, 2003; Roces & González Torres, 1998; Zimmerman, 1998).

As for the effects of SRL on using computer software, Bielaczyc, Pirolli and Brown (1995) incorporate self-explanation and self-regulation strategies in the attainment of the cognitive skill of computer programming. They find that their treatment group, which incorporates the self-regulation strategies of self-monitoring and clarifying comprehension failures in conjunction with self-explanation strategies outperform a control group that did not have the benefit of instruction in these strategies. Their study implies that in addition to knowledge acquisition strategies, students benefit from the incorporation of strategies which allow them to plan, monitor, and evaluate their understanding and strategy use (Bielaczyc, Pirolli & Brown, 1995). In a similar vein, this study provides us an insight that SRL is appropriate to be applied in computer software education.

Previous studies have established that self-regulation skills can help foster learning from any instructional method (see Ertmer, Newby, & MacDougall, 1996; Lindner & Harris, 1993; Weinstein, 1989; Zimmerman, 1990). With regard to the effects of SRL in the online learning environment, it is indicated that successful students in an online course generally used self-regulated learning strategies and the effect of self-regulation on students' success was statistically significant (Yukselturk & Bulut, 2007). In Shen, Lee, and Tsai's (2007b) study, it is also revealed that the intervention of web-enabled SRL contributes to students' learning.

However, only a few psychological studies investigated the research questions of how learners should be supported when learning to use a software product and how instructions should be designed to achieve effective software training. Various fields of researchers should make contributions to this topic, which will not only allow for the formulation of recommendations and guidelines for instructional approaches to software training, but also provide valuable knowledge concerning theoretical and educational issues beyond software training (Bannert & Reimann, 2000). Therefore, the authors not only explore the potential effects of web-enabled SRL on students' computing skills, but also provide an instructional design for web-enabled SRL in this study.

PROBLEM-BASED LEARNING WITH INITIATION

Problem-based learning is a teaching method that may engage students in authentic learning activities by using challenging problems in practice as a starting point, stimulus, and focus for learning (Barrows, 1985, 1986). PBL may help students achieve learning goals such as professional reasoning, integration of scientific, academic and professional knowledge, and lifelong learning skills (Dunlap, 2005). Many researchers have reported PBL's positive impact on knowledge and skill acquisition and transfer, problem solving, attitudes and opinions about courses and programs, measures of performance, and self-directed learning (Albanese & Mitchell, 1993; Berkson, 1993; Colliver, 2000; Davies, 2000; Norman & Schmidt, 1992; Vernon & Blake, 1993).

PBL is a flexible approach, demonstrated to work well with both small teams and large groups. However, there might be disagreement whether PBL will be as effective or even possible for online learning. In this regard, Chanlin and Chan (2004) examined the effects of PBL in a web-based approach. Results revealed that students in the PBL treatment group perform better than those from the control group. Furthermore, in Shen, Lee, and Tsai's (2007a) study, it is indicated that the intervention of web-enabled PBL contributes to students' development of computing skills. However, low-academic-achievement students usually lack the ability to seek the essential information and solve the problems they face.

Technological problem-solving skills provide students with the opportunity and facility to fulfill the various requirements of the technological design process, regardless of learning environment type. For example, students participating in web page programming lessons initially require some form of teacher exposition of the necessary skills to enhance their competent use of Microsoft FrontPage software. Once understood and learned, this knowledge serves to support student-initiated design and problem-solving activities within the same web page programming learning environment. Similarly, more traditional teaching supports students' procedural abilities, in terms of hand and machine skills, until students are adept at applying these previously acquired skills during their technological design activities (Walmsley, 2003). Technology teachers need to facilitate with students the autonomous development of their own cognitive and metacognitive strategies when solving technological problems. It is argued that these types of learning environments require a teacher-facilitated student-centered pedagogy (Deluca, 1992; Johnson, 1996). Therefore, it is believed that the effects of web-enabled PBL with initiation on students' skills with application software are positive, and higher than those without.

METHODS

Participants

The participants in this study were 177 vocational students during a period of two semester from four class sections each taking a compulsory course titled 'Web Page Programming and Website Planning' in a university of science and technology in Taiwan. None of them majors in information or computer technology. Students at this university were expected to spend much more time and effort in mastering a variety of technological skills as compared to those in comprehensive universities in Taiwan.

Course Setting

The course under study is a semester-long, 2 credit-hour class, targeted at third-year college students in the department of business administration. This course focuses on the development of students' skills of web page programming and website planning. Computing education is emphasized for students of all levels and disciplines. Even students in the department of business administration or nursing still have to develop the required computing skills.

Students in this online course have to solve a series of tasks by applying several application software packages, including the tools of website building, web page programming, graphic design, and file transfer software. They are required to collaborate to build up business-quality websites. Therefore, there will be real products, that is websites, to be designed and completed by the end of the semester.

EXPERIMENTAL DESIGN AND PROCEDURE

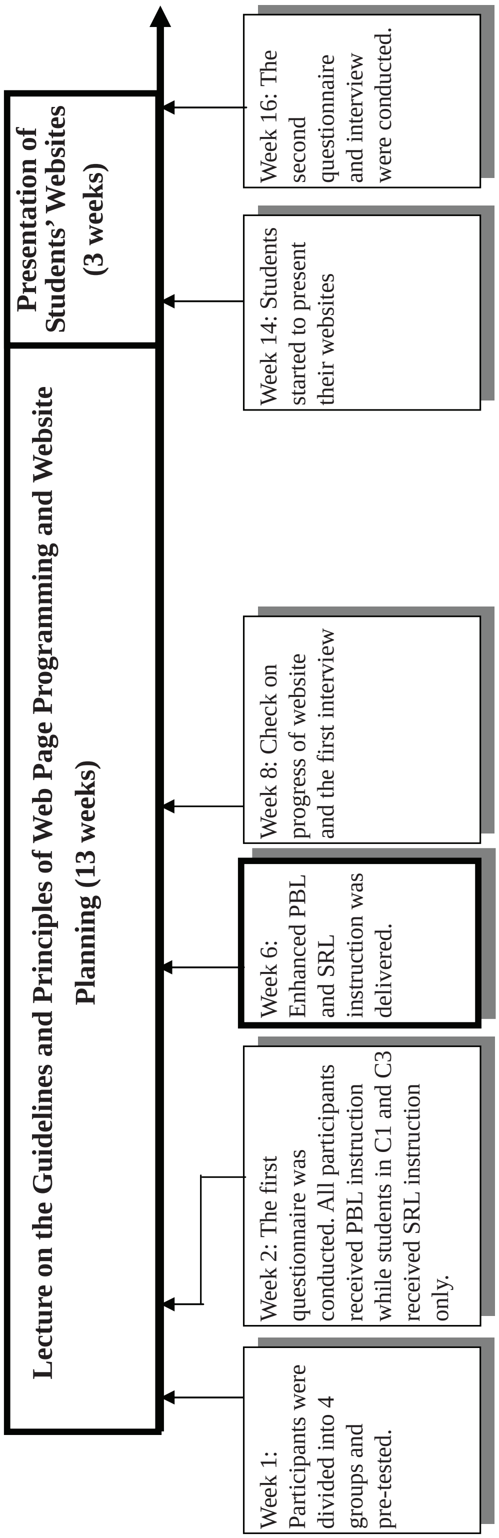
For this study, we redesigned our course and adopted innovative teaching methodologies and technologies to help students develop their practical computing skills. We explore whether students in the course 'Web Page Programming and Website Planning' enhance their skills of application software via e-learning. Based on reflection on our earlier research, we re-designed the course and conducted a series of quasi-experiments to examine the effects of web-enabled SRL, web-enabled PBL with initiation, and their combinations on students' computing skills.

The experimental design is a 2 (SRL vs. non-SRL) × 2 (PBL with initiation vs. PBL without initiation) factorial pretest-post test design (see Figure 1). Four classes were selected from two successive years for this quasi-experiment. Participants are randomly assigned to one of the four experimental conditions. The SRL and PBL with initiation group (C1, n=56), non-SRL and PBL with initiation group (C2, n=44), SRL and PBL without initiation group (C3, n=44) are experimental groups, while non-SRL and PBL without initiation group (C4, n=33) is the control

Figure 1. Expected effects of variation in instructional methods

	SRL n	on-SRL
PBL with Initiation	The most significant effect (C1 Group)	Medium effect (C2 Group)
PBL without Initiation	Medium effect (C3 Group)	Minor effect (C4 Group)

Figure 2. The schedule of the course and presentations



group. Moreover, C1 and C2 were conducted in the first year, while C3 and C4 were conducted in the second one. Before the experiment, we first measured students' skills of web page programming and their involvement in this course as a pretest at the beginning of the course. However, there are no significant differences among the four groups. Then, treatments and interventions were applied to students in the four groups according to the different combinations of teaching methods.

Students were divided into teams, each consisting of 5 to 6 members. Each team had to build up a quality website that conformed to the guidelines or principles in the literature and textbook. The schedule of teaching and students' presentations is depicted in Figure 2.

Treatment of Web-Enabled SRL

Students in C1 and C3 received instruction in an after-school course teaching them SRL strategies. The SRL group was gathered in a classroom and a two-hour lecture was delivered discussing how to manage study time and regulate their learning. The content of this SRL course was composed of the four processes addressed by Zimmerman, Bonner and Kovach (1996), that is, self-evaluation and monitoring, goal-setting and strategy planning, strategy implementation and monitoring, and monitoring of the outcome of strategy. Students were taught how to implement these four processes to become more self-regulated learners.

Reflection on our previous experiences suggests the rigid and extra requirements of SRL may result in low achievers' resistance and antipathy to the treatment. Many vocational students have been taught via spoon-feeding since they were elementary school students. It is difficult to require them to take the responsibility for their own learning. Teachers should adopt appropriate and acceptable SRL strategies for students. Keeping records and self-evaluation were the two most-often-mentioned learning strategies (Ross, 1999). Moreover, high-achieving students utilized the strategies of reviewing notes and keeping records and monitoring most often (Zimmerman & Martinez-Pons, 1986). Therefore, in this study we emphasized keeping records, reviewing notes, and self-evaluation when implementing SRL among low-achieving students. SRL students were required to regularly prepare and read the textbook before classes, and to review what they had learned after school. They were also required to record their learning behavior every week. The data was recorded on the course website instead of in their notebooks in order to prevent falsification of records.

Moreover, students who have gotten used to spoon-feeding teaching method would not change their learning habits and adapt to the innovative teaching methods immediately. Thus, the instructor gave enhanced SRL instruction to improve students' SRL and enhance their confidence by illustrating the positive effects of SRL for students from the previous semester of another SRL experiment. Through this link between outcomes of learning the use, or non-use of appropriate strategies, students can learn to believe that desirable learning outcomes are the products of the strategy used. This can encourage students to persist in leaning and use the appropriate strategies (Shin, 1998).

Treatment of Web-Enabled PBL with Initiation

A teacher creates interesting, challenging, and authentic problem situations. Students played the role of Net Programming Develop Engineer in a simulated company. Students in the PBL group were required to build a prototype of their website in the initial stage of this course. After that, the teacher lectured on the guidelines or mistakes and bugs of web pages in each class, then the students started to discuss about how to implement what they learned through online chat room or forum. They had to reconsider and revise their web pages according to the new knowledge they just learned. The knowledge and skills they learned were immediately applied in their products.

In this course, the teacher scaffolded student learning in a systematic manner. In this regard, Barrows' (1992) "whiteboards" model was applied to help students solve problems and learn from the problem-solving processes. The whiteboard served as a focus of concentration for negotiation of the problem and as a forum for students to co-construct knowledge. All students in this study received a one-hour lecture about PBL. The content of this PBL lecture consisted of the four-column chart put forth by Barrows, that is:

1. **Facts:** It holds information that the students gathered from the problem statement such as what the problem is and where it occurred.
2. **Ideas:** This column serves to keep track of their evolving hypotheses about solutions, such as reducing the search time in a website.
3. **Learning** Issues: The students place their questions for further study into this column. In this example, students identify issues related to the technology for designing the structure of the website.
4. **Action Plan:** This column is used to keep track of plans for resolving the problem or obtaining additional information.

The whiteboards model provided a systematic approach to solve problems. Applying this model made the problem-solving processes more feasible for low achievers. Therefore, Barrows' whiteboards model was applied in this study to help students solve problems in a consistent and systematic manner, and learn from the problem-solving processes.

As mentioned above, vocational students have been used to spoon-feeding teaching methods for many years. They usually lack the ability to seek the essential information and solve the problems they face, particularly in a web-based course without the teacher's on-the-spot assistance. In this round of experiment, the teacher initiated his students into the field of web page programming and website planning. He first established students' background knowledge and developed required skills in the initial stage of the course. After students climbed the stiff learning curve and encountered bottlenecks, students were required to gather information and solve problems with Barrows' whiteboards model by themselves.

Evaluation

Before the experiment and intervention in this study, we first measured students' skills of web page programming and their involvement in this course as a baseline. Students completed three Microsoft FrontPage documents as a pretest, and the scores show a uniformly low skill level. Microsoft FrontPage was chosen for the pretest because almost every student in Taiwan learns FrontPage before he/she learns other web page programming software.

Moreover, Zaichkowsky's (1985) Personal Involvement Inventory (PII) was used in this study to measure students' psychological states regarding personal relevance or importance of this course. PII measures three constructs: interests, needs, and values. In the second week, students completed the first questionnaire as a pretest. The difference in students' involvement in this course at this beginning stage among the four classes was not statistically significant. Therefore, it was considered that the students had equal skills of web page programming and involvement when they began this course. In addition, none of them had any experience in taking a web-based course.

Students began to present their websites in the 14th week of the semester. The teacher graded according to Website Evaluation Criteria (Tan & Tung, 2003) and Web Usability (Schaffer & Sorflaten, 1999). All students in one team would get the same grade. Finally, the enhancement of computing skills is the result of one's grade on the website minus one's pretest grade. We tested the differences in students' enhancement of the skills of web page programming under different conditions.

RESULTS

We took grades on students' performance with their websites as a measure of their computing skills. The 'independent samples t-test' was used to compare improvement of grades between PBL with initiation and PBL without initiation groups. As shown in Table 1, students' computing skills in the PBL with initiation group (82.9425) was significantly higher than that in the PBL without initiation group (73.3117). Therefore, it is believed that the effects of initiation on students' skills in web page programming and website planning were positive, and higher than that without, in a web-enabled PBL environment.

Results from Table 2 show that students' computing skills in the SRL group (81.0800) was significantly higher than that in the non-SRL group (75.7305). That is, the effects of web-based SRL on students' skills in web page programming and website planning were positive, and higher than those without SRL intervention.

Finally, data from Table 3 shows that combination of SRL and PBL with initiation intervention a group results in the highest grades among the four groups. The improvement of skills in web page programming and website planning in C1 is significantly higher than C3 and C4, and also higher than C2, though insignificantly. Therefore, we conclude that the effects of web-enabled SRL and PBL with initiation intervention on students' skills of web page programming and website planning are positive, and higher than for those who do not receive web-enabled SRL or/and PBL with initiation.

Table 1. Independent samples t-test: The improvement of grades

Groups	n	Mean	S. D.	F	t-value	df	*p*
PBL with initiation	100	82.9425	10.77606	15.536	5.080	175	.000***
PBL without initiation	77	73.3117	14.45070				

*Note: ***p < 0.001, **p < 0.05, *p < 0.1.*

Table 2. Independent samples t-test: The improvement of grades

Groups	n	Mean	S. D.	F	t-value	df	*p*
SRL	100	81.0800	9.72123	14.497	2.688	175	.008**
non-SRL	77	75.7305	16.54384				

*Note: ***p < 0.001, **p < 0.05, *p < 0.1.*

Table 3. One-way ANOVA: The improvement of grades

Dependent Variable		(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.
Improvement of Grades	Scheffe	1	2	2.89042	2.45515	.709
			3	7.12338(*)	2.45515	.041
			4	15.94156(*)	2.67449	.000
		2	1	-2.89042	2.45515	.709
			3	4.23295	2.59828	.450
			4	13.05114(*)	2.80646	.000
		3	1	-7.12338(*)	2.45515	.041
			2	-4.23295	2.59828	.450
			4	8.81818(*)	2.80646	.022
		4	1	-15.94156(*)	2.67449	.000
			2	-13.05114(*)	2.80646	.000
			3	-8.81818(*)	2.80646	.022

* *The mean difference is significant at the .05 level.*

DISCUSSION

The Effects of Web-Enabled PBL with Initiation

In this study, the instruction of PBL with initiation was found to play a positive role in enhancing students' computing skills. As the data in Table 1 shows, there was a very significant difference between the PBL with initiation and PBL without initiation classes on the skills of web page programming and website planning ($p = 0.000$). This demonstrates that PBL with initiation is good for computer software education in general, and e-learning in particular. It was suggested that traditional lecturers should shift or adapt to problem-based learning and then align constructively in online teaching (Talay-Ongan, 2003). The findings of the present study are similar to those that appeared in previous studies, which revealed that students in PBL treatment groups performed better than those from the control group in a web-based learning environment (Chanlin & Chan, 2004; Shen, Lee, & Tsai, 2007a).

In Taiwan, vocational students have been taught via spoon-feeding since they were elementary school students. They usually lack the skills to gather information and inquire knowledge to solve problems by themselves. If a teacher implements PBL in his class without initiation, it is very difficult for students to adapt to this innovative teaching method; hence it may lead to limited effects. Therefore, we suggest that teachers design their courses systemically and initiate their students into the field of study. Teachers should provide the essential skills and knowledge for students in the initial stage to help them climb the learning curve, and achieve the potential effectiveness of PBL.

The Effects of Web-Enabled SRL

The data shown in Table 2 also support that the difference in students' skills of web page programming and website planning between web-enabled SRL and non-SRL groups was statistically significant ($p = 0.008$). The importance of self-regulated learning and students' abilities to successfully direct their own learning efforts in Internet-supported environments is emphasized in the literature (Winnips, 2000; Cennamo, Ross & Rogers, 2002; Azevedo, Cromley, Winters & Moos, 2004). Thus, strategies must be put into practice to prepare students for the rigors of learning at a distance and increase the probability of retention and success (Chang, 2005).

In general, students in the vocational system tend to have low confidence and motivation in learning (Su, 2005), have low interest and negative attitude toward their learning (Chen & Tien, 2005), spend more time on part-time jobs, do not adequately get involved in their schoolwork, and don't care so much about their grades (Shen, Lee, & Tsai, 2007a). In this specific context of low achievers, teachers may be taking a high risk in implementing e-learning. However, this study supported that web-enabled SRL helped low achievers learn better in an online course, which may lower the risk.

The Effects of Combination of Web-Enabled SRL and PBL with Initiation

With respect to the effects of the combination of SRL and PBL with initiation, we found preliminary support from the results in Table 3. The results show that the effects of combined training in SRL and PBL with initiation on enhancing students' computing skills are positive and higher than for those who did not receive SRL or/and PBL with initiation, although the difference between C1 and C2 is not statistically significant. These results are also similar to those that appeared in Paris and Paris's (2001), Perels, Gürtler and Schmitz's (2005), and Kramarski and Gutman's (2006) studies. To conclude, this study suggests that teachers could apply PBL and SRL simultaneously rather than singly in the modules to strengthen the interaction of SRL and PBL to promote students' learning.

Based on our findings, there is a warning signal for teachers who plan to implement e-learning, particularly in vocational schools. For those teachers who wish to stick to traditional methods of teaching, or adopt PBL without initiation, this may not be a fruitful approach. Students who experienced neither the SRL nor initiation treatments, in C4, had the lowest grades among the four groups, significantly (see Table 3). It is suggested that teachers should redesign their courses, adopt new instructional methods and technologies, and initiate their students into the field of study to fully exploit the benefits of web-based learning environments.

CONCLUSION

SRL and PBL have been applied successfully for teaching in different academic fields for decades. These two instructional methods could make further contributions to students' learning through the Internet. In this study, we re-designed the treatment of web-enabled PBL to include initiation to further the understanding of its appropriate design and potential effects. Moreover, SRL and PBL with initiation were simultaneously applied as web-based pedagogies to help students develop their practical skills in web page programming and website planning. Results were generally positive, showing enhanced student computing skills.

Providing online courses in an environment where Internet addiction looms is a challenge to both instructors and students. Teachers and students may suffer from ineffectiveness resulting from use of traditional instructional methods without appropriate re-design, particularly in online courses. This study provides an intervention in the context of vocational education addressing how to help students adapt and develop practical computing skills. In addition, it also provides specific reference about how to assist students to regulate their learning and further improve their grades. This study may provide valuable insights and shed light on new and effective practices for schools (particularly vocational schools), scholars and teachers preparing for or presently engaged in implementing e-learning.

REFERENCES

Albanese, M. A., & Mitchell, S. (1993). Problem-based learning: A review of literature on its outcomes and implementation issues. *Academic Medicine*, *68*(1), 52–81.

Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology*, *96*(3), 523-535.

Azevedo, R., Cromley, J. G., Winters, F. I., & Moos, D. C. (2004). *Designing adaptive scaffolds in hypermedia to facilitate students' self-regulated learning.* Paper to be presented at the Annual Conference of the American Educational Research Association, San Diego, CA.

Bannert, M., & Reimann, P. (2000). Guest Editorial: approaches to the design of software training. *Journal of Computer Assisted Learning*, *16*(4), 281-283.

Barrows, H. S. (1985). *How to Design A Problem-Based Curriculum for the Preclinical Years.* New York: Springer.

Barrows, H. S. (1986). A taxonomy of problem-based learning methods. *Medical Education*, *20*(6), 481-486.

Barrows, H. S. (1992). *The Tutorial Process.* Springfield: Southern Illinois University School of Medicine.

Berkson, L. (1993). Problem-based learning: Have the expectations been met? *Academic Medicine*, *68*(1), 79-88.

Bielaczyc, K, Pirolli P., & Brown A. (1995). Training in self-explanation and self-regulation strategies: investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction*, *13*(2), 221-252.

Bottino, R. M., & Robotti, E. (2007). Transforming classroom teaching & learning through technology: Analysis of a case study. *Educational Technology & Society, 10*(4), 174-186.

Boud, D., & Feletti, G. (1991). *The challenge of problem based learning.* London: Kogan Page.

Bruer, J. T. (1993). *Schools for thought: A science of learning in the classroom*. Cambridge, MA: MIT Press.

Cennamo, K. S., Ross, J. D., & Rogers, C. S. (2002). Evolution of a web-enhanced course: Incorporating strategies for self-regulation. *Educause Quarterly*, *25*(1), 28-33.

Chang, M. M. (2005). Applying self-regulated learning strategies in a web-based instruction: An investigation of motivation perception. *Computer Assisted Language Learning*, *18*(3), 217-230.

Chanlin, L. J., & Chan, K. C. (2004). Assessment of PBL design approach in a dietetic web-based instruction. *Journal of Educational Computing Research*, *31*(4), 437-452.

Chen, C. H., & Tien, C. J. (2005). *Market Segmentation Analysis for Taking Skill Test by Students in an Institute of Technology*. Retrieved March 28, 2008, from http://www.voced.edu.au/td/tnc_85.574

Colliver J. A. (2000). Effectiveness of problem-based learning curricula: Research and theory. *Academic Medicine*, *75*(3), 259-266.

Davies, P. (2000). Approaches to evidence-based teaching. *Medical Teacher, 22*(1), 14-21.

Deluca, W. V. (1992). Survey of technology education problem-solving activities. *The Technology Teacher, 51*(5), 26-30.

De Corte, E. (1996). Changing views of computer supported learning environments for the acquisition of knowledge and thinking skills. In S. Vosniadou, E. De Corte, R. Glaser, & H. Mandl, (Eds.), *International Perspectives on the Designing of Technology-Supported Learning Environments* (pp. 129-145), Mahwah, NJ: Lawrence Erlbaum.

Dunlap, J. C. (2005). Changes in students' use of lifelong learning skill during a problem-based learning project. *Performance Improvement Quarterly*, *18*(1), 5-33.

Ertmer, P. A., Newby, T. J., & MacDougall, M. (1996). Students' approaches to learning from case-based instruction: The role of reflective self-regulation. *American Educational Research Journal*, *33*(3), 719-752.

Järvelä, S., Näykki, P., Laru, J., & Luokkanen, T. (2007). Structuring and Regulating Collaborative Learning in Higher Education with Wireless Networks and Mobile Tools. *Educational Technology & Society*, *10*(4), 71-79.

Johnson, S. D. (1996). Technology education as the focus of research. *The Technology Teacher*, *55*(8), 47-49.

Kauffman, D. F. (2004). Self-regulated learning in web-based environments: instructional tools designed to facilitate cognitive strategy use, metacognitive processing, and motivational beliefs. *Journal of Educational Computing Research*, *30*(1), 139-161.

Kramarski, B., & Gutman, M. (2006). How can self-regulated learning be supported in mathematical E-learning environments? *Journal of Computer Assisted Learning*, *22*(1), 24-33.

Lindner, R. W., & Harris, B. (1993). Teaching self-regulated learning strategies. In M. R. Simonson & K. Abu-Omar, (Eds.), *Proceedings of selected research and development presentations at the annual conference of the Association for Educational Communications and Technology*. (pp. 641-654). IA: Instructional Resources Center, Iowa State University.

Löfström, E., & Nevgi, A. (2007). From strategic planning to meaningful learning: diverse perspectives on the development of web-based teaching and learning in higher education. *British Journal of Educational Technology*, *38*(2), 312-324.

Lowe, D., & Hall, W. (1999). *Hypermedia & the Web: An Engineering Approach*, New York: John Wiley & Sons.

Nielsen, J. (1995). *Multimedia and Hypertext: The Internet and Beyond*. San Diego: Academic Press.

Norman, G., & Schmidt, H. (1992). The psychological basis of problem-based learning. *Academic Medicine*, *67*(9), 557-565.

Paris, S. G., & Paris, A. H. (2001). Classroom applications of research in self-regulated learning. *Educational Psychologist, 36*(2), 89-101.

Perels, F., Gürtler, T., & Schmitz, B. (2005). Training of self-regulatory and problem-solving competence. *Learning and Instruction, 15*(2), 123-139.

Psaromiligkos, Y. (2003). Re-evaluating the effectiveness of a web-based learning system: A comparative case study. *Journal of Educational Multimedia and Hypermedia, 12*(1), 5-20.

Reyero, M., & Tourón, J. (2003). *El desarrollo del talento: la aceleración como estrategia educativa* [The development of talent: acceleration as an educational strategy]. A Coruña: Netbiblo.

Roces, C., & González Torres, M. C. (1998). Capacidad de autorregulación del aprendizaje [Ability to self-regulate learning]. In J.A. González Pienda & J.C. Núñez (Eds.), *Dificultadesde aprendizaje escolar.* (pp. 239-259). Madrid: Pirámide/Psicología.

Ross, J. D. (1999). *Regulating Hypermedia: Self-Regulation Learning Strategies in a Hypermedia Environment.* Va: Virginia Polytechnic Institute and State University.

Schaffer, E., & Sorflaten, J. (1999). Web usability illustrated: Breathing easier with your usable e-commerce site. *The Journal of Economic Commerce, 11*(4), 158-166.

Shelley, D. J., Swartz, L. B., Cole, M. T. (2007). A comparative analysis of online and traditional undergraduate business law classes. *International Journal of Information and Communication Technology Education, 3*(1), 10-21.

Shen, P. D., Lee, T. H., & Tsai, C. W. (2007a). Applying web-enabled problem-based learning and self-regulated learning to enhance computing skills of Taiwan's vocational students: A quasi-experimental study of a short-term module. *Electronic Journal of e-Learning*, 5(2), 147-156.

Shen, P. D., Lee, T. H., & Tsai, C. W. (2007b). Facilitating students to pass certificate tests via blended e-learning with self-regulated learning: A quasi-experimental approach. *WSEAS Proceedings on Multimedia, Internet & Video Technologies*, Beijing, China.

Singh, H. (2003). Building effective blended learning programs. *Educational Technology, 43*(6), 51-54.

Su, M. H. M. (2005). A study of EFL technological and vocational college students' language learning strategies and their self-perceived English proficiency. *Electronic Journal of Foreign Language Teaching, 2*(1), 44-56.

Talay-Ongan, A. (2003). Online teaching as a reflective tool in constructive alignment. In P.L. Jeffery, (Eds.) *Proceedings of International Education Research Conference AARE – NZARE,* Auckland, New Zealand, Australian Association for Research in Education.

Tan, F. B. & Tung, L. L. (2003). Exploring website evaluation criteria using the repertory grid technique: A web designers' perspective. *Proceedings of the Second Annual Workshop on HCI Research in MIS*, 65-69. Retrieved March 28, 2008, from http://sigs.aisnet.org/SIGHCI/Research/ICIS2003/HCI03_10.pdf

Vernon, D., & Blake, R. (1993). Does problem-based learning work? A meta-analysis of evaluative research. *Academic Medicine, 68*(7), 550-563.

Walmsley, B. (2003). Partnership-Centered Learning: The Case For Pedagogic Balance In Technology Education. *Journal of Technology Education, 14*(2). Retrieved March 28, 2008, from http://scholar.lib.vt.edu/ejournals/JTE/v14n2/walmsley.html

Weinstein, C. (1989). Teacher education students' preconceptions of teaching. *Journal of Teacher Education, 40*(2), 53-60.

Williams, S. M. (1993). Putting case-based learning into context: Examples from legal, business, and medical education. *Journal of Learning Sciences*, *2*(4), 367-427.

Winnips, K. (2000). *Scaffolding-by-design: A model for WWW-based learner support*. Enschede: University of Twente Press.

Wu, T. Y. (2000). *Integrative curriculum planning in technological and vocational education in Taiwan, Republic of China*. San Diego, California: Paper presented at the Annual Conference of the Association for Career and Technical Education.

Yukselturk, E., & Bulut, S. (2007). Predictors for student success in an online course. *Educational Technology & Society*, *10*(2), 71-83.

Zaichkowsky, J. L. (1985). Measuring the Involvement Construct. *Journal of Consumer Research*, *12*(3), 341-352.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, *25*(1), 3-17.

Zimmerman, B. J. (1998). Developing self-regulation cycles of academic regulation: An analysis of exemplary instructional model. In D.H. Schunk & B.J. Zimmerman (Eds.), *Self-regulated learning: From teaching to self-reflective practice,* (pp. 1-19). New York: Guilford.

Zimmerman, B. J., Bonner, S., & Kovach, R. (1996). *Developing self-regulated learners: Beyond achievement to self-efficacy*. Washington, DC: American Psychological Association.

Zimmerman, B. J., & Martinez-Pons, M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal*, *23*(4), 614-628.

Zimmerman, B. J., & Schunk, D. H. (1989). *Self-Regulated Learning and Academic Achievement: Theory, Research, And Practice*. New York: Springer-Veriag.

This work was previously published in International Journal of Information and Communication Technology Education, Vol. 5, Issue 3, edited by L. A. Tomei, pp. 1-13, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 5.25
Hypertinence, Serendipity or Elicitation of Passion for Knowledge?
Some Critical Elements of Online Learning by Using Web 2.0 Resources

Simona Marchi
University "Sapienza" of Rome, Italy

ABSTRACT

Learning processes depend on the socio-technical and regulatory contexts in which professional practices and daily usage take place. These processes develop out of and through various systems of activities, consisting of subjects, artefacts, rules, knowledge, and roles. What happens when the rules governing these systems, the roles, artefacts and knowledge change? For instance, what happens to learning processes when the settings in which practices take place are virtual, when they occur, say, in a blog, or in a community, or on a social networking platform? In this chapter the author intends to examine in detail the specific features of learning processes taking place in these new online environments.

INTRODUCTION

In 2005, Tim O'Reilly published an article on the changes developing in the relationship between the web and its users. Essentially, O'Reilly stated that negotiation between the web and its users was heralding a new era for the Internet (known as Web 2.0), in which users are allowed to interact directly with the web to create user-generated content. This takes us beyond the one-way communication phase (from web to users) and gives users a role of direct involvement in communication. In a very short time, the phenomenon of Web 2.0 applications has exploded.

These applications have encouraged a large-scale shift from 'individual publishing' to 'personal managing', an activity that enables any user not just to create and share content or partly personalize interfaces, but also to design and manage an interface completely or even develop one's own platform to manage a range of resources and/or interfaces. Without underestimating the playful aspect of the use of Web 2.0 resources, it is also worth focusing on the value of online practices from the viewpoint of individual and collective learning, with which the playful aspect is undoubtedly connected.

The rapid changes in recent years have led web users to activate and constantly reformulate their knowledge and practices in relation to a variety of contexts, objects and diverse actors - including in their on/off-line relationships - based on the possibility of generating and managing this 'relational content' in terms of personal/collective managing.

This content, in turn, may take the form of technologies or new media. Web 2.0 resources are technologies-in-practice (Orlikowski, 2000), which structure users' practices through their repeated use and the way they are operated. They are also technologies-in-technology, i.e. technologies-in-practice with a potentially large number of technological actants and hence the possibility of numerous levels of interaction - technological, semiotic, material and reflective - within the same context of practices. And these resources can also be practices-in-technology, i.e. the generation of new technological resources through the experience of their usage.

Some of the most important innovations include tagging, blogging and social networking, and it should be noted that folksonomies, as opposed to traditional taxonomies, have developed within these innovations for the management of information and knowledge. Bottom-up processes concerning the organization of knowledge, in reality, involve the opposition of top-down classification and indexation versus bottom-up methods of organizing and indexing knowledge.

But how to find our way around? The web environment is definitely a hyper environment (hyper as in hyperlink, but also in the sense of enormous, exaggerated). The volume of resources that can be found through the web and the number of possible actions are far beyond our capacity to use them. At the same time, however, the web can also be somewhat inaccurate, scattered, uncoordinated, and often not very pertinent. How can we conceive of learning processes that are visible, recognisable, and based on the use of Web 2.0 resources?

Information Technology (IT) systems, conceived with the aim of creating the conditions and processes to support the development, processing and sharing of information, have not always produced the hoped-for results. Some studies show that we can by no means take it as given that IT investment improves the processes of creating and exchanging knowledge within an organization (Bernard, 1996; Davenport & Prusak, 1998); others point out that people are often reluctant to use new technologies (Markus & Keil, 1994).

We do not believe that Web 2.0 resources alone, given their technological features, can overcome these barriers. However there are grounds to believe, as we shall see in this chapter, that a relationship exists between the socio-technical features of the use of a given technology and the degree to which subjects use, activate, participate in and share that technology. This relationship is spurious in that it is mediated by a number of motivational aspects. These aspects, however, are in turn connected to a series of other factors, including features specific to these same participation practices.

The concepts of Hypertinence (De Kerchkove, 1998), Serendipity (Merton, 1949; Merton & Barber 1958) and Passion (Barbalet, 1996; Solomon, 1993; Fabbri, 1998), as we shall see, help us to specify the forces underlying participation in communities, social networks, blogs, etc. These aspects are not mutually exclusive. Rather, they interact and enable us to understand the tensions within the processes whereby individuals or groups participate in online learning practices.

Therefore, in this chapter, we will regard Web 2.0 resources as a set of practices - discursive, organizational, classificatory, connective, etc. - performed within and throughout certain environments. This way of viewing Web 2.0 allows us to consider its nature as a situated social practice and to take a relational view of the technology involved.

Starting from some experiments in using Web 2.0 resources to foster learning processes in groups of students and adults undergoing training, we intend to highlight the collective and co-structured nature of the learning that takes place in socio-technical and regulatory contexts based on the centrality of interaction. We also, and above all, intend to analyze the reflective processes whereby new rules, new roles and new knowledge (and, not last, new power relations and new identities) are developed, organized and internalized by the 'new virtual techno-subjects'.

The chapter develops from three concepts deemed to be essential to understand the specific features and potential of online learning practices; these concepts are also the three objectives that should be borne in mind in designing learning experiences using Web 2.0 resources. The chapter continues by focusing on a number of significant and critical elements that should be carefully considered and which emerged through experiences that enabled us to 'touch on' the above-mentioned three objectives. Finally some ideas about future trends and conclusive remarks are developed.

BACKGROUND

Web 2.0 resources make it possible to build participation networks based on networks of relations whose most significant aspect, in learning terms, is not just the content presented online, but also, and above all, the ways in which individual subjects connect with each other. If the medium is the message (McLuhan & Quentin, 1967), then the social dimension of connectivity becomes the message of the internet medium. Similarly, social networks become the message of Web 2.0 resources and users become the content of this message. Content that is not the 'object' but, rather, a network itself referring to a system of professional and friendship networks that make the individual subject 'communicable' to the outside world. The creation of expanded communities based on shared interests, in which various Web 2.0 instruments interact, is explained by De Kerchkove using the concept of 'hypertinence' (De Kerchkove, 1998).

Hypertinence is expressed through the exaltation of multiple resources combined with the appropriateness of their content, resulting from interactions that make answers increasingly pertinent to questions (e.g., using certain types of search engines that rely on tags), or through the interaction with other users displaying value of experience, systematic organization, webbed organization, bridge links, and index labels.

As a result of the contributions made by users when reporting and reviewing their knowledge resources, hypertinence represents a further step beyond the results of search engines which, although able to re-create a network of quotations, are merely the outcome of quantitative calculations.

If we consider the concept of hypertinence in isolation, the connecting practices may be 'simple clicks', i.e. they may not lead to the production of an extensive connective intelligence or be collaborative or reflective, and therefore may not be very significant to online learning. This concept needs to be considered along with two others.

The concept of 'serendipity' (Merton, 1949; Merton & Barber 1958) has a fascinating story that began in the second half of the 18^{th} century. It is hard to find a single definition for this word, which can be summed up as meaning to accidentally discover something precious while looking for something else entirely, or to find somewhere you were looking for but in a completely unexpected way or place. The word 'serendipity' was originally coined by Horace Walpole (1722) and derives from the title of an Asian tale, *Travels* *and Adventures of Three Princes of Serendip.* Merton first came across the term in about 1945, again in a serendipitous manner: while looking in his dictionary for a word starting with the same letters, his eye fell by chance on serendipity. He was immediately struck by the way its definition chimed with the sociological importance of the unintentional consequences of intentional actions in the life of society in general, and of unexpected stages in the growth of knowledge.

The 'model of serendipity in scientific research' (Merton, 1949) indicates a process whereby new ideas take hold in the minds of researchers and the conditions are created for the definition of new scientific concepts. It consists of observing an 'unexpected', 'anomalous' and 'strategic' event that provides an opportunity to develop a new theory or expand on an existing one. The three adjectives in italics are central to the definition of the serendipity model (*ibidem*). The first two (unexpected and anomalous) describe the unexpected way we come across new knowledge: an unforeseen observation produces unexpected knowledge since it prompts the researcher to find a meaning for the observation. But this is not enough in itself to complete the serendipity model. The importance of the third attribute (strategic), describing the event encountered in an unexpected way, concerns, more than the unexpected event itself, the act of recognizing the potential research value attributed to it by the researcher. It is therefore more closely linked to the researcher's assessment of the research content. Serendipity should not be taken as referring only to the special abilities of people who are able to find the unexpected. Rather, it should also, and above all, refer to the experience itself: the action performed when encountering and giving meaning to the unexpected.

This brings us to the importance of paying due attention, in sociological and cultural terms, to the circumstances in which serendipity is most likely to occur. At several points in his analysis Merton focuses on what he defines as 'serendipitous cognitive micro-environments', i.e. culturally variegated environments in which the opportunity for prolonged socio-cognitive interactions by individual talents in different disciplines creates the conditions for serendipity to occur. Such environments can even be created intentionally; indeed in this respect Merton spoke of 'institutionalized serendipity'.

This concept and its history, strongly resemble the discovery and learning processes that take place online. More specifically, the strategic dimension and the concept of 'serendipitous micro-environments' reflect an idea of using and combining Web 2.0 resources in such a way as to create environments and participation/connection systems with the characteristics described above, i.e. that are able to foster serendipitous experiences. This concept enables us to expand our capacity to understand new learning opportunities in relation to practices that become significant through online experience itself. The strategic dimension also tells us that these experiences should be conceived in such a way that subjects can attribute a potential value to them with respect to their goals and individual learning, so that they can be recognized as meaningful experiences.

Another concept that is dealt with more and more frequently - albeit in a non-uniform and often contradictory manner - in the reference literature on participation in the online production of knowledge, particularly from open sources, is 'passion'. We shall not cover this concept exhaustively in this chapter, but shall mention the key points that enable us to understand its importance in determining forms of participation in online practices.

Including the concept of passion in studies on online learning practices would seem to go without saying. However, we must consider that organizational studies, that analyze knowledge production processes in groups and organizations, and semiotics itself have both kept the rational dimensions of action separate from its non-rational aspects, often omitting the latter.

Passion here is viewed as a form of emotion (Barbalet, 1996) - debatable as this may be - that cannot be reduced solely to the dimension of feelings or sentiment (Solomon, 1993) and is closely connected to certain situations or experiences. Passion can be considered as a judgement system that is not just directly connected to the individual's values and beliefs vis-à-vis the world but is also, and above all, an active way of structuring and giving meaning to an experience (*ibidem*). Consequently, passion cannot be distinguished from the contexts, relations, and objects with respect to which these judgement systems are activated. The point is to separate the question of passion from its usual opposition to reason and to connect the concept of passion to that of action.

Passion can therefore be conceived as a view of action held by those who are affected and transformed by this action (Fabbri, 1998). But at the same time it is not entirely rational. The irrationality is often connected with the urgency with which the judgement - a mix of concepts, desires, attitudes and beliefs - is formed. This way of conceiving emotions, and passions in particular, enables us to overcome the usually negative vision of emotions found in organizational studies (Thompson, 1967). Spontaneous, free participation in certain online practices, such as the co-generation of content in open source environments, or participation in blogging or social networking practices, is driven by passion for the very activity of participating and producing new knowledge, new discourse, or objects and by the fact of being somehow transformed by this participation.

Time dedicated, the determination to write and take part in discursive practices, as well as sacrifices made in unthought-of hours and during holidays are indices of the desire to assert that something has value and is precious and important (Polanyi, 1958). The discourse on passion, if referred to online learning practices, is linked to narrative practices. Narrativity, which is prevalent in online learning practices, is a way of causing shifts of significance by specifically combining not just words, phrases or propositions, but special 'actants' that then become characters, actors, and so on (Fabbri, 1998). It is therefore a way of creating objects of value. From this perspective, narrativity is the act of configuring meaning through actions and passions.

The performance of Web 2.0 resources, being a set of situated practices, is built on their use, through the relations in which they are inscribed and through the relations inscribed in these resources. Often, this performance can lead to the contexts and contents generated being abandoned when the experience that made them significant comes to an end. The web is full of abandoned identities and places, stories left unfinished, contents forgotten for years but never destroyed by the users who own those spaces. Given the ever growing production of and participation in social media, we can describe the web as being full, to paraphrase the title of a publication by Zygmunt Bauman, of 'wasted identities and places' (Bauman, 2003), meaning both e-waste and uninhabited areas. In the web, performance and distinctiveness are such that there is a huge recycling of identity, *habitus* and context, occurring with the massive production of permanently visible 'waste'.

Active participation, therefore, is closely related to the way the environments of online practices are configured, or, rather, the experience of online practices, which must be able to produce voice (Hirschman, 1970) and above all maintain it in the presence of a high and constant risk of exit. But producing and maintaining voice is by no means simple and cannot be left solely to chance or given over entirely to technology.

Active user participation in online learning practices is closely related to the possibility/capacity to create environments made of human and non-human actors conceived, taking the centrality of the relation as a starting point, in a sort of 'heterogeneous engineering' (Law, 1987) in which the following conditions are satisfied:

- Pertinence with respect to certain learning experience issues and possibility of creating an extensive connective awareness of how we conceive of and classify these issues (*objective pertinence and connective awareness*).
- Presence of wide degrees of freedom and constant possibility - in the presence of an idea of action, study or research - of making significant discoveries by serendipitous routes (*serendipity and projective transformations*).
- Possibility of producing, through relations, objects deemed to be precious and valuable and which activate prospects for action by individuals and the group, who are in turn transformed by these prospects (*effectiveness and individual/collective values*).

As situated social practices (Suchman, Blomberg & Trigg, 1999), Web 2.0 resources can be expressed at one and the same time as the set of relations through which knowledge is produced and takes shape, and as 'mediators' in the activities of users. Therefore, these resources can also be expressed as practices that create meaning and learning situations rising above individual activities and single subjects and technologies. In this respect the concept of mediation assumes a generative meaning since all the 'objects' and 'materials' that are mobilized in heterogeneous engineering are repositioned - as a result of the relations within which they are included, and for the time in which they are included and associated to achieve a given objective (Gherardi, 2008).

SOME CRITICAL ELEMENTS OF ONLINE LEARNING PRACTICES

If we want to understand experiences of learning in virtual space, it is necessary to reconsider how we can conceptualize and describe what is happening in Web 2.0, in order to reformulate, inform and enrich new concepts of our theoretical framework.

'Edit content, research, update object, make new friends, create, expand the network, login, share information, tag contents' and so on, are some of the key elements of activity systems (Engëstrom, Miettinen & Punamäki, 1999) in internet time (Wellman & Haythornthwaite, 2002). The concept of activity incorporates the cultural, motivational, contextual and organizational aspects of action. This set of circumstances gives rise to 'situated activity systems' (Goffman, 1967, 1969; Suchman, 1987) characterized by being composed of sets of interactions between subjects and objects/artefacts which play a part in configuring specific situations.

Wikis, folksonomies, blog, social networks, are not only technological artifacts, neither just pre-existed places/environments for users practices (Burrows & Ellison, 2004). Contexts are therefore emergent effects of actions and interactions between the elements (human and non-human) of which they are composed. The 'non-human' elements are also considered as being co-generated through sets of relations (Latour, 2005). Web 2.0 assumes the characteristics of multidimensional emerging 'enacted environment' (Weick, 1969) brought into and kept in existence, continually up-dated and modified/modifiable through subjects' practices.

The peculiarity and sociological interest of practices reside not so much in their variability but in their repetition and recursiveness (Gherardi & Bruni, 2007), and in their function of social reproduction and identity construction. Artistry, creativity, as well as the capacity and will to innovate, change and invent are nested and develop within this recursiveness. In the case of online learning studies, the point is to bring out the variability and creativity associated with practices that manifest themselves as elementary and repetitive, or technically designed to reproduce place/context. In actual fact Web 2.0 environments change constantly in terms of their content, structure of networks, the outcome of their expressive forms and their effect on participants, even though they are constructed and maintained through practices that reproduce the resources themselves. Also if we can work online collaboratively with persons who have different geographical locations (Korica, Maurer & Schinagl, 2006), in online learning practices we can refer to a situated practice which is only potentially global, the world within everyday reach (Schutz, 1971). We refer, that is, to an activity that both depends on and constitutes the context and is expressed in a system of interactions that is only potentially broad in scope.

Creativity/artistry consequently refers, as we have mentioned, to the link between 'expert action' and 'creation' that occurs each time we find ourselves acting in situations of uncertainty, instability, unique circumstances, or conflicts of values (Schön, 1983). This link is often not fluid or 'natural'; it is fraught with tensions, in which significant situations of high creativity or crisis occur. Artistry is required in all situations where there is not 'just one right way' or a single assessment criterion to be applied in doing things. It is required, therefore, in all situations that appear as dilemmas. In Web 2.0 practices we have to choose continually, we have to construct continuously our interpretations of objects, contents, narratives, relationship, identities.

In the case of online learning practices, the artistry/creativity concept manifests itself in expert yet innovative actions expressing the subject's presence and/or participation, in a situation of constant 'case uniqueness'. Each 'expert action'- from the object/text published, to the comment, to the choice of aesthetic features of the context, and so on - is expressed within a state of tension between reproduction, technical rationality and creativity, as well as innovation of form and content. In this way, it becomes the object of a reflective process of expressing a form of presence. It thus becomes artistry. For instance: the collective construction of ideas and concepts using tags and conceptual maps. Tagging not only makes 'lateral searching easier' (Maness, 2006), but can also generate movements in non-linear directions and improve several structures of connection.

In order to develop 'serendipitous cognitive micro-environments' (Merton, 1949), we can to construct a Web 2.0 laboratory of practices by using different Web 2.0 resources at the same time. This kind of laboratory has different cultural, social and organizational dimensions, is a 'place in action', safe, accountable to organizations, institutions and individuals, in which participants co-construct objects, contents, meanings, relationships, re-design actions.

In the laboratory, the meaning of community, referring to Goffman's dramaturgic model, is closer to what Goffman calls 'équipe' rather than community of practices (Wenger, 1998). In this respect the 'actor' never works alone in the staging process but collaborates with a set of persons who are 'accomplices' in ensuring that the situation appears the way they want/expect it to. The équipe acts as a 'secret society' (an 'invisible society' in the use of Web 2.0 resources) in which all subjects play a part in keeping the secret as to who/how/when/why they collaborate with or on to keep a particular state of affairs and/or situation in existence and up-to-date. The traditional 'place' in which the secret is kept is 'backstage', the private dimension of social action: offline.

Even given this traditional concept of the distinction between public-private/backstage-frontstage with which we also associate the activity of masking and unmasking, Web 2.0 resources can also be viewed as places where anti-dramaturgic social trends develop in which the confine between backstage and frontstage becomes fluid. The place in which the 'secret' is kept can therefore be represented by the public, 'frontstage' or online dimension. This is based on a principle of 'sincerity' rooted not in the distinction between true-false or person-actor, but in the impression and/or content communicated by participants through their actions. In this respect the participant is 'sincere' (or a liar) by definition, since they always show and communicate what they intend on any given occasion to show and communicate, including their silence and 'apparent absence'. In this reversal the bonds creating an *équipe*, in the narrow sense, are weakened, which means that in the face of an 'invisible society' the community becomes a 'visible society' whose members are, in their different ways, accomplices in the same game of presence/absence.

This types of communities are characterized by interaction. From the technical point of view, in Web 2.0 resources public interaction is very often asynchronous and occurs through the screen and through a system of links, tags and references within and outside the same setting which enable users to read, produce (verbally-visually-acoustically), navigate or be silent. This is flanked by another type of private interaction that may also be synchronous (e.g. through chat rooms) and takes place through private messages (PVT, emails), the backstage or frontstage nature of which depends on the content, objectives and forms of the interaction itself.

What emerges, then, is a change in the rules of the 'staging process', or rather of the 'presence-absence' in which the optical effect produced by online learning practices is that of generating apparently dichotomous situations of presence or absence in which technically either the former or the latter predominates. The movement between presence and absence, which should on the other hand be considered in its entirety, is connected with other types of interaction linked primarily to new systems of legitimization or recognition activated in online practices, which in turn are mediated by the rules of the places in which they develop. This means that a community's recognition of the initial status it has conferred on a participant is replaced by a constant redefinition of participation and forms of interaction based on the rules governing the 'approval/acceptance' that participants attain in the context of one or more communities (which correspond to social circles that can be very wide in scope and size).

The tension between 'approval/acceptance' and 'spontaneity' within a community helps to determine forms of participation and therefore of presence-absence, as well as the processes involved in constructing participants' online identity.

Network practices - often voluntary, associative and free - are practices that produce value and objects of value within a community. These can be understood as: 'terms, gestures, behaviours and objects that give an idea of what is happening, what is circulating and what has a density, objectivity and solidity that makes it necessary to alter attitudes, shake up pre-conceptions, change opinions', and adjust practices (Latour, 2007, p. 227).

The value produced in a laboratory conceived for online learning does not reside solely in the activity of producing knowledge, but rather in distributing, sharing and co-generating that knowledge. The various forms this value can take depend on participation and objectives, the fact that it is free, and the forms taken by the gift in interaction practices. The gift (Mauss, 1925), a key feature of online practices, is expressed in terms of objects, content and relations based on recognition of the other and on trust, listening, receptiveness and reciprocity.

Reputation in online learning activities is more than building a 'fanbase' (Beer, 2006). It is built through the play of the opposing forces of inclusion and distinctiveness, and through an awareness and recognition by others of the contribution a subject has brought to the discussion, is surely one of the most important issues at stake and one of the most valuable objects in online social practices. The way we appraise and appreciate the value of the contributions made by the various participants is the basis upon which reputations are built online.

Participants in online learning practices use different types of narratives (Boland & Tenkasi, 1995), stories (Denning, 2001), myths (Gherardi, 2003), that may sometimes include creative forms of narration such as fiction. By fiction we do not mean something invented or imaginary, as opposed to facts and events that have actually happened. We are referring to something that is an integral part of the processes of inclusion and exclusion in online learning practices, participation in discussion and reflection on the meaning of experience. It is connected with the representation of participants' thinking and experience in the form of various sorts of narrative, images, sounds and video (Clandinin & Connelly, 2000).

So participants' narratives are considered 'as if' they were cross-sections of social practice in the narrow sense. Social practices that make it possible to produce an 'alignment' effect and 'provisional order' by constructing an individual meaning mediated with respect to a shared objective - represented by the reason for which subjects are invited to register and take part in an online learning environment -, and also with respect to multiple factors internal to the network practice. In this 'as if' mediated within a set of relations that are constantly questioned and remodulated, reconfirmed, reframed and reviewed, we find two aspects expressed. One is artistry, which is important since it is rooted in participants' experiences as well as being performed through a system of relations also expressed through the use of Web 2.0 resources. And the other is an effect of balance and alignment that is always provisional and is the main object of knowledge production on the web.

Rhythm is a dimension of context that characterizes Web 2.0 resources as practices of social interaction. The rhythm and the configuration of contexts come from the frequency of the actions and interactions. It is closely related to the number of participants and their active presence, and to the ways of managing/moderating the learning process itself.

Building a rhythm that makes presence sustainable is essential to create an alignment between actants that paves the way for processes of co-constructing inclusion and/or exclusion mechanisms and/or knowledge on the web.

This balance is extremely difficult to achieve and above all maintain. The 'daily reproduction of practices is the effect - always precarious and open to negotiation - of three processes: stabilization, institutionalization, and habitualization' (Gherardi, 2008, p. 40). On the web this effect is highly problematic since the activity of knotworking itself leaves learning processes always open to continuous redefinitions This leads to a redefinition of the context and so of the practices themselves. Unsustainable rhythms and non-agreed moderations often correspond to various kinds of 'anti-programme' (Latour, 1987), manifested in different ways, which make collective work difficult and sometimes impossible.

FUTURE TRENDS

Reflection/reflectivity refers here to the idea of developing a new epistemology of online practices (to paraphrase Schön, 1983). It is not possible to resolve the concept of reflection/reflectivity in just a few lines but we shall consider here the accepted meaning most applicable to this chapter. Reflective practice consists of a set of methods that make it possible to construct and/or reconstruct the meaning of actions, events and selection processes which occur in relation to certain situations that are currently taking place or have already occurred.

These methods are intended to foster a 'reflective conversation with the situation' (*ibidem*). The objectives can be numerous and diverse: to bring out tacit knowledge, to encourage individual and organizational learning, to produce knowledge, to build a community of practices, to identify the meaning of decisions made, practices realized and effects obtained, to foster organizational improvement, to encourage an awareness of the motives for participation/non-participation in certain processes, and so on. In the case of online learning practices we find ourselves mainly in an asynchronous situation in which events appear on a screen and seem to have already occurred. The most recent trend is an interest in more creative forms of reflective practice. In other words: *how we can use reflective practice/s to build and sustain more innovative learning, work teams and organisations*? Creative reflective practice/s help us to re-frame and to see things differently. To embrace eclectic pluralism and methodological innovation. The challenge for the reflective learning community, we suggest, is to think about how the critical is embraced within the creative (Ghaye, 2008).

In this sense, we can use the concept of expansive learning (Engëstrom, 1987), which refers to processes activated by subjects in situations that appear to be highly contradictory, or new and unknown. These processes are activated only rarely and concern, in particular, the individual and collective activity of distancing oneself from a specific action setting text in order to build a bigger picture with a view to expanding the setting itself, and in order also to build a broader range of criteria, a wider object of activity. In this, subjects can free themselves from the limits of the setting in which they are acting and in which new criteria for choice and action can be constructed. This is a matter of learning something that is not yet there, in a place where the subject is learning while building a new identity (*ibidem*).

On the basis of the considerations made thus far, we can state that online learning practices are also reflective learning practices if the learning experience:

- Is conceived as a set of situated social practices that occur in relation to objectives that are redefined from one situation to the next.
- Is based on forces of inclusion and at the same time distinction between participating subjects. These forces are in part institutional and in part negotiated through interactions and a redefinition of the ways of being present/taking part.
- Makes it possible for a socially produced and therefore constantly changing reality to emerge, in which participants are immersed and participate by co-determining that same reality.
- Establishes an epistemic yet pragmatic relationship with everyday practice inside and/or outside the community.
- Activates processes of co-constructing knowledge and ways of appreciating and attributing value to the contribution made by each participant.
- Enables individuals and groups to express and organize their knowledge, opinions and objects in an appreciative manner.
- Retroacts on the practices themselves and on modes of participation.

However, there are many difficulties:

- Exclusion mechanisms can be very powerful and may act in non-uniform manner.
- Practice stabilization processes are subject to constant redefining and result in temporary stages of stabilization.
- The rapid obsolescence of knowledge, and experience itself, on the web makes it difficult to identify ways of appreciating and attributing value to the participation and contribution of each subject.
- Reputation is highly important and regulates *participants*' choices in terms of practices, identities and styles of connection. These aspects are variable and may generate tensions, all the more so when relations include heterogeneous subjects and objects.
- Recalcitrant or abstracted/unaligned objects, or anti-programmes, are always present in each online learning practice and may undermine relations or create the conditions for a frequent reformulation of objectives and common values.
- The high degree of variability in using practices of Web 2.0 resources since decisions,

practices and knowledge involve multiple fields of action.

- The tension between silence and the ways of appraising and appreciating the value of individual contributions is the strand along which new forms of online power are played out; these may create a spiralling increase in the rhythm of online practices or else discourage participation.
- The overall effect of online practices might fail to satisfy the expectations of individual participants, who could then decide to abandon the community.
- As the objectives of the community become increasingly stringent and occupy most of the discursive space, the sense of 'citizenship' in certain contexts could decrease, and this could undermine democratic participation. When this happens, any hardening of the coordinators' positions in defining the rules of participating and formulating comments could drive some participants to leave.

CONCLUSIVE REMARKS

Although online learning practices based on the use of Web 2.0 resources can become situated social practices of reflective learning, this does not always happen. The point is to conceive such practices in a way that takes various factors into consideration: promoting the use of Web 2.0 resources is always an intervention so it always produces an effect in the relations in which the practice takes place. It often plays a part in creating, *ex novo*, a context of relations; this changes continually and interacts constantly and unpredictably with participants' practices and choices and, therefore, directly with the technology.

Practices using Web 2.0 resources clearly highlight all these aspects, starting with the procedures for building an online identity. The first outcome of this technology is to produce a user who, while free, is co-produced through interaction with software. But this simple procedure enables us to clarify that the distinction between *users* is generated through a relationship between human and non-human actors (technological/institutional). Participation in a Web 2.0 online learning environment, as we have seen, also requires a more or less tacit agreement by participants, an agreement that is constantly reviewed and renegotiated on the basis of the realignments emerging from interactions.

If we want to satisfy the following three conditions:

1. Objective pertinence and connective awareness
2. Serendipity and projective transformations
3. Effectiveness and individual/collective values

the central issue is not a technological one but is a relational one.

The reason so many complex e-learning systems have failed to produce the hoped-for results is that they often delegate mediation processes entirely to technology and pre-packaged content. The mediation processes are inscribed in the software itself and not co-generated, a situation that negates the epistemic function of network practices and only favours the pragmatic function. Most systems that use digital platforms for distance learning should be defined as e-teaching, rather than e-learning, experiences. What is missing is the relationship. These systems could be defined as 'practices without communities' or 'practices without mediation'. There is a risk that the focus placed on improving the usability of e-learning systems will not resolve but merely hide the problem.

Mediation brings out the complexity of the relations between the different institutions that define the rules of the game in specific activity environments made up of different cultures, access rules and accountability processes. Degrees of freedom, social exchange, improving individual and collective competences, making tacit, no-formal and informal competence accountable, and co-constructing processes of reciprocal recognition and reputation are some of emergent elements of mediations in Web 2.0.

In this way participants' mutual trust is created and constantly tested and their desire to identify with the undertaking and its outcome is expressed. In this respect, knotworking (Engeström, 2008) and constantly appreciating and attributing value to the product of co-generation become essential to the redefinition of relations between participants and the contribution/motivation of each in relation to the common project/objective. The rules of co-existence in an online learning environment, from netiquette to the ethical protocol, also play a role in creating an online community and specifying the rules of access/recognition/reputation and individual and collective contributions.

Reflective and appreciative approach in using Web 2.0 resources can help us. There are three main components of reflective appreciative intelligence: the ability to reframe things in different way, appreciate the positive, and see how the future unfolds from the present. For this to happen, participants need to be persistent, have self-belief, have a tolerance for uncertainty and have irrepressible resilience (Ghaye, 2008).The key question is: *what are our successes and how can we amplify them to build and sustain a better future from valued aspects of the positive present*?

Facilitators have a central role in constructing hybrid environments, bigger pictures, situations where new criteria apply and where participants learn something only potentially present in the setting/environment where they everyday act, precisely while they are co-constructing a new activity. Learning that takes place through online practices configured as hybrid laboratories is both situated, in that it occurs in the context of the settings and activities mentioned above, and at the same time 'expanding', in that it emerges and is activated from a setting that is potentially true/achievable (Marchi, 2008).

REFERENCES

Barbalet, J.M. (1996). Social emotions: Confidence, trust and loyalty. *The International Journal of Sociology and Social Policy*, *16*(9-10), 75-96.

Bauman, Z. (2003). *Wasted lives*. Cambridge: Polity Press.

Beer, D. (2006). The Pop-Pickers Have Picked Decentralised Media: The Fall of Top of the Pops and the Rise of the Second Media Age. *Sociological Research Online*, *11*(3), from http://www.socresonline.org.uk/11/3/beer.html

Bernard, R. (1996). *The corporate intranet*. Chichester: Wiley.

Boland, R., & Tenkasi, R. (1995). Perspective making and perspective taking in communities of knowing. *Organization Science*, *6*(4), 350-372.

Burrows, R., & Ellison, N. (2004). Sorting Places Out? Towards a Social Politics of Neighbourhood Informatisation. *Information, Communication and Society*, *7*(3), 321-336.

Clandinin, D.J., & Connelly, F.M. (2000). *Narrative inquiry: Experience and story in qualitative research*. San Francisco: Jossey-Bass.

Davenport, T.H., & Prusak, L. (1998). *Working knowledge. How organization manage what they know*. Boston: Harvard Business School Press.

De Kerkchove, D. (1998). *Connected intelligence*. London: Kogan Page.

Denning, S. (2001). *The springboard. How storytelling ignites action in knowledge-era organizations*. Boston: Butterworth Heinemann.

Engeström, Y. (1987). *Learning by expanding. An activity-theoretical approach to developmental research*. Helsinki: Orienta-Konsultit.

Engeström, Y. (2008). *From teams to knots*. Cambridge: Cambridge University Press.

Engeström, Y., Miettinen, R., & Punamäki, R.L. (1999). *Perspectives in activity theory*. Cambridge: Cambridge University Press.

Fabbri, P. (1998). *La svolta semiotica*. Roma-Bari: La Terza.

Ghaye, T. (2008). *An Introduction to Participatory and Appreciative Action Research*. Gluocester: New Vista Publications.

Gherardi, S. (2003). Knowing as desiring: mythic knowledge journey in community of practitioners. *Journal of workplace learning*, *15*(7-8), 352-358.

Gherardi, S. (2008). *Apprendimento tecnologico e tecnologie dell'apprendimento*. Bologna: il Mulino.

Gherardi, S., & Bruni, A. (2007). *Studiare le pratiche lavorative*. Bologna: il Mulino.

Goffman, E. (1967). *Interaction ritual*. Garden City: Doubleday.

Goffman, E. (1969). *Strategic interaction*. Philadelphia: University of Philadelphia Press.

Hirschman, A.O. (1970). *Exit, voice, loyalty: Responses to decline in firms, organizations, states*. Cambridge: Harvard University Press.

Korica, P., Maurer, H., & Schinagl, W. (2006). *The Growing Importance of e-Communities on the Web*. Paper presented at the IADIS International Conference on Web Based Communities, San Sebastian, Spain.

Latour, B. (1987). *Science in action, How to follow scientists and engineers through society*. Cambridge: Harvard University Press.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford: Oxford University Press.

Latour, B. (2007). *La fabbrica del diritto*. Enna: Città Aperta (orig. version Le fabrique du droit, 2002).

Law, J. (1987). Technology and heterogeneous engineering: The case of the portuguese expansion. In W.E. Bijker, et al., (Eds.), *The social construction of technical systems: New directions in the sociology and history of technology* (pp. 111-134). Cambridge: MIT Press.

Maness, J.M. (2006). Library 2.0 Theory: Web 2.0 and its implications for Libraries. *Webology*, *3*(2), from http://www.webology.ir/2006/v3n2/a25.html

Marchi, S. (2008). L'approccio logico riflessivo applicato alle pratiche di blogging. In M.M. Mapelli & R. Lo Jacono, (Eds.), *Nuovi modelli di ricerca universitaria. Pratiche collaborative in rete* (pp. 91-99). Milano: Mimesis.

Markus, M.L., & Keil, M. (1994). If we build it they will come: Designing information systems that users want to use. *Sloan Management Review*, *35*(4), 11-15.

Mauss, M. (1925). *Essai sur le don. Forme et raison de l'échance dans les societies archaÏques*. Retrieved September 28, 2008, from http://classiques.uqac.ca/classiques/mauss_marcel/socio_et_anthropo/2_essai_sur_le_don/essai_sur_le_don.html.

McLuhan, M., & Quentin, F. (1967). *The medium is the message*. New York: Bantam Books.

Merton, R.K. (1949). *Social theory and social structure*. Glencoe, Illinois: The Free Press.

Merton, R.K., & Barber, E. (1958). *The travels and adventures of serendipity*. Princeton: Princeton University Press.

O'Reilly, T. (2005), *What Is Web 2.0. Design patterns and business models for the next generation of software*. Retrieved September 28, 2008, from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html?page=1

Orlikowsky, W.J. (2000). Using technology and constituting structures: A practice lens for studying technology in organizations. *Organization Science*, *11*(4), 404-428.

Polany, M. (1958). *Personal knowledge: Towards a post-critical philosophy*. Chicago: University of Chicago Press.

Schön, D. (1983). *The reflective practitioner. How professionals think in action*. New York: Basic Books.

Schutz, A. (1971). *Collected Papers*. The Hague: Nijhoff.

Solomon, R. (1993). *The passions*. Indianapolis: Hackett.

Suchman, L. A. (1987). *Plans and situated actions. The problem of human-machine communications*. Cambridge: Cambridge University Press.

Suchman, L. A., Blomberg, J., Orr, J.E., & Trigg R. (1999). Reconstructing technologies as a social practice. *American Behavioural Scientist*, *43*(3), 392-408.

Thompson, J.D. (1967). *Organizations in action. Social science bases of administrative theory*. New York: McGraw-Hill.

Weick, K.E. (1969). *The social psychology of organizing*. Massachusetts: Addison-Welsey.

Wellman, B., & Haythornthwaite, C. (Ed.). (2002). *The Internet in Everyday Life*. Oxford: Blackwell.

Wenger, E. (1998). *Communities of practice: learning, meaning, and identity*. Cambridge: Cambridge University Press.

KEY TERMS AND DEFINITIONS

Artistry: Individual and collective ability that emerge in situations of uncertainty, instability, or conflict of values.

Hypertinence: Interactions that make answers pertinent to questions.

Passion: Active way of structuring and giving meaning to an experience.

Serendipity: Process whereby new ideas take hold in minds of actors, on things which they were not in quest of.

Reflective Learning: Committed and value-laden action to try to improve particular situations or contexts.

Reputation: Way to appraise and to appreciate the value of the contributions made by individuals.

Situated Practice: Practices performed with other subjects in a context composed of objects and rules.

This work was previously published in Handbook of Research on E-Learning Applications for Career and Technical Education: Technologies for Vocational Training, edited by V. C.X. Wang, pp. 166-179, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 5.26
Blending Virtual Campuses Managing Differences Through Web 2.0 Experiences in Transnational Cooperation Projects

Yuri Kazepov
University of Urbino "Carlo Bo," Italy

Giovanni Torrisi
University of Urbino "Carlo Bo," Italy

ABSTRACT

Starting from the increasingly widespread need to develop effective teaching in complex transnational settings, this chapter presents an innovative blended model with Web 2.0 collaborative learning strategies built in. The model balances pedagogical, technical and content related issues into an ad hoc institutionally designed 60 ECTS (European Credit Transfer System) curriculum of the European Masters in Comparative Urban Studies (E-Urbs). The chapter aims at disentangling the different dimensions involved in the curriculum delivery, highlighting the pros and cons of all dimensions of the model adopted. In doing so the chapter is divided into three sections. The first section addresses the challenges that effective teaching in complex transnational settings has to face, in particular it highlights the crucial need of managing differences. In the E-Urbs project we had 24 students from 14 countries, 5 continents, 6 disciplinary backgrounds, 32 scholars from 9 partner institutions in 8 countries. The second section deals with the way in which challenges and differences have been addressed and describes the dimensions of the blended model the authors adopted, arguing that a sound virtual campus arrangement should address the pedagogical, technical and content related dimensions in a balanced way considering the institutional setting within which they are embedded. The third section addresses the way in which the blended approach has been enriched through a Web 2.0 perspective, promoting p2p (peer-to-peer)

DOI: 10.4018/978-1-60566-358-6.ch014

collaboration in the generation of knowledge. The main argument is that an increasingly fluid society generates and treats information differently and learning agencies should not only acknowledge these differences but should address them with balanced learning models which take advantage of the new 2.0 paradigms. The authors argue that the result of a balanced blended Web 2.0 approach helps to transform the challenges into a resource for each of the stakeholders involved (e.g., students, scholars, partners, institutions) providing an added value in each dimension of the learning process (pedagogical, technical, content related and institutional).

INTRODUCTION

This chapter proposes an innovative blended model in which Web 2.0 collaborative learning strategies have been coupled with a blended approach in order to cope with the difficulties normally faced by online courses such as declining attention over time and the potentially increasing social distance among participants. The model has been developed and implemented during the first year of E-Urbs, a European Master in Comparative Urban Studies[1], funded within the virtual campus stream[2].

The distinctiveness of the model is to provide a balanced system in which different aspects are calibrated in order to provide a 2.0 blended learning environment, based on a very strong tutoring activity. Pedagogical, technical and content related issues have been balanced to construct and sustain an *ad hoc* institutionally designed 60 ECTS curriculum. In the chapter, each dimension has been separately analysed, highlighting the potential problems arising from an "unbalanced" distribution of weights and priorities.

THE CHALLENGE OF EFFECTIVE TEACHING IN TRANSNATIONAL SETTINGS

During the last decade, E-Learning in its various facets has considerably grown (Fletcher, 2004; Waterfield, 2002). Industry, universities and professional schools have experienced the advantages and difficulties derived from this kind of learning arrangement. Some argue that this expansion of E-Learning models is due to the reduction in costs and infrastructure (Munro & Munro, 2004) in particular when compared to traditional face-to-face (f2f) arrangements. Others relate it to the possibilities that the new technology, including Web 2.0 and learning-object style of learning, gives to didactic innovation.

Universities are among the institutions that have used and experienced the most different E-Learning models, exploiting the benefits that virtual arrangements have on campus life (Bacsich, 2004). Increasingly, technology based solutions, including E-Learning and Web 2.0, are considered an answer to the Bologna process[3] and the Europeanisation of higher education systems. The latter ask for new means by which students can experience innovative ways of studying and learning together in a truly European learning community. The European Commission considers this – according to the resolution of the European Council in Lisbon in March 2000 – as a necessary step in order to foster grow and competitiveness in a knowledge-based society (Kok, 2004).

The "virtual campus" becomes, therefore, a new organisational solution for enriching the transnational offer of universities where, thanks to the use of a LMS (learning management system), learning activities are completed either partially or completely online, with the distant/online assistance of the professor and tutors. Different from other kind of arrangements, the virtuality of these campuses facilitates the creation and development of distant learning communities and research programmes. Thanks to technology, designing integrated European curricula becomes easier and virtual campuses might be one of the main instruments for it.

Virtual campuses, however, cannot be directly compared to traditional campuses. They are of a completely new kind, because they offer not only new possibilities of creating and disseminating knowledge for students, professors and tutors, but because they are qualitatively different. Among their main potential strengths we have their ability to unravel and process differences, whether they are cultural, linguistic or disciplinary. This is one of the reasons why they are regarded with much interest when dealing with the knowledge-based society that values information and differences. This is a society where multiculturalism is a fact and thus comparative studies become a necessity. We are living in a society where information is ubiquitous, goods move incessantly, services are global and people continuously migrate. In the 21st century, society is more multicultural than ever. This is not because there are more cultural differences than in the past, but because these differences are endlessly confronted, incessantly related, instantaneously addressed and put in question. As Bauman (2000; 2005; 2008) would say, we are in a fluid-tachycardic society. In this fluid life, words change their meaning and competitiveness and speed become a priority: speed in gathering information, in dealing with complexity and managing knowledge. Universities and other knowledge producing agencies have to cope with this situation, implementing in the learning processes instant communication and new ways of teaching. Comparative studies, where differences are to be studied, valued and jointly analysed become also a fundamental means for competitiveness.

In this context virtual campuses and comparative studies might be seen as unavoidable milestones. But can virtual campuses be considered a possible (good?) answer in dealing with the complexity of a knowledge-based (Kok, 2004) multicultural society? Are virtual campuses well equipped in order to cope with the implicit difficulties that diversity and the need of comparative studies carry with themselves? Which are the main characteristics of the challenges that a virtual campus would need to manage? How is it possible to create and maintain a sense of learning community? And what are the main characteristics that a virtual campus like arrangement has to have in this context?

We will try to answer these questions on the basis of our experiences with the European Master in Comparative Urban Studies (E-Urbs), since:

- It is a virtual campus, as defined earlier;
- It deals with a multicultural and multi-linguistic cohort of students;
- It is organised by 9 different institutions, with different teaching and organisational traditions;
- It uses a comparative approach;
- It makes use of a Web 2.0 approach;
- It fosters a blended approach that relates online, p2p and f2f learning processes one another.

More specifically, E-Urbs strategically deploys the long lasting experience of the partners in comparative urban studies and in online teaching with the aim of institutionalising a truly European curriculum of 60 ECTS. Extensive online teaching (and tutoring) is aimed at encouraging and facilitating both physical and virtual mobility between higher education systems and institutions across Europe. This fosters the underlying principles of the Bologna process and the institutionalisation of a European Higher Education area through the use of an innovative E-Learning environment for higher education (Land of Learning). Moreover, it aims at monitoring the effects on the interactive process and the use of learning objects.

In short, the E-Urbs Master reproduces and intensifies learning processes within a multicultural frame and uses new technologies in order to address differences. From this point of view, E-Urbs has been a laboratory in which the aforementioned challenges have been explicitly faced, in particular the following ones:

Cultural Differences

Twenty four students from 14 countries and 5 continents participated in various activities of the Master[4]. Differences in customs and culture were considerable. Festivities and bank holidays were dissimilar both for the students and the professors involved, as well as their eating habits. Some uneasiness and prejudice was present at the very beginning in the interaction among some students attending the course, especially in relation to sensitive subjects as race, religion and social belonging. E-Urbs became from this point of view a unique laboratory for going multicultural in a technologically mediated (and relatively protected) environment.

Linguistic and Age Differences

A virtual campus involving people from all over the world needs to cope with the language and age differentiation. Our decision was to use English as a working language. All people had at least a basic knowledge of English. This, however, did not iron out the differences in linguistic competence among the students as none of them had English as their mother tongue. The issue became even more problematic when dealing with the writing of the intermediate and final papers where language skills were a crucial asset. As far as the age of the students is concerned, we can distinguish between two different groups. The first one includes students that just finished their Bachelor degree (average age 24). They were looking for a specialisation curriculum eventually to access a Ph.D. programme. The second age group (average age 30) includes mainly professionals. Some of them were working in public or private institutions and dealing with urban issues already, while others wanted to change their professional career.

The Heterogeneity of Institutional Settings

The Master has been organised by 9 different higher education institutions in 8 countries with different institutional constraints, such as, different admission requirements for students, tuition fees, administration habits, procedures that must be fulfilled in order to receive the diploma, a different understanding of the European Credit Transfer System and the Diploma Supplement. National bureaucracies make it difficult to implement common procedures. Often simple and unchallenged issues become insurmountable problems. Can scholars of one university teach in other university of the consortium and can their teaching load be considered as part of their duties? Would this guarantee the fulfilment of the accreditation criteria necessary to establish a Master course? Is it possible to issue a joint degree or a degree that would be recognised by all partner institutions involved? These basic questions heavily influenced the development of the Master course.

The Background of Disciplinary Traditions

Comparative urban studies use a multidisciplinary approach, as many other Master courses nowadays. Students taking part in the Master were coming from different disciplinary backgrounds, from psychology to planning, from sociology to architecture. This situation – despite the high motivation and interest in urban studies of all students – constituted a major challenge, since there was the need to share a basic knowledge on which to build a more specialised understanding of urban-issues.

Physical and Virtual Distance

Virtual-campus-like arrangements make it possible to organise multidisciplinary courses at a distance. This does not mean that distance issues are completely nullified. Keeping the class active and united with a sense of community, while people are scattered around the world has been a major challenge. If keeping a class interested is a challenge by itself in normal learning environments, when the same culture is shared, the language is easily understood and disciplinary background are somehow similar, it becomes a major challenge where these conditions are not given. Motivation seems to be the key issue in both cases, but the absence of a physical gathering and interaction adds further difficulties. As the literature shows, (e.g., Schifter, 2000) interest in forum, chats and traditional virtual methodology tend to weaken over time and the learning community tends to shrink as time passes by. New methodologies were needed in order to cope with these issues.

Psychological Sense of Community

Many of the differences listed earlier and in particular the physical distance among students, made it difficult to create and sustain a psychological sense of community. Living in different contexts, people tend to tie in with their physical environment more than with a virtual course. Nevertheless, a sense of community is a necessary condition in order to guarantee a successful learning context and traditional solutions for this problem are not directly applicable in a virtual campus.

Communication Technologies

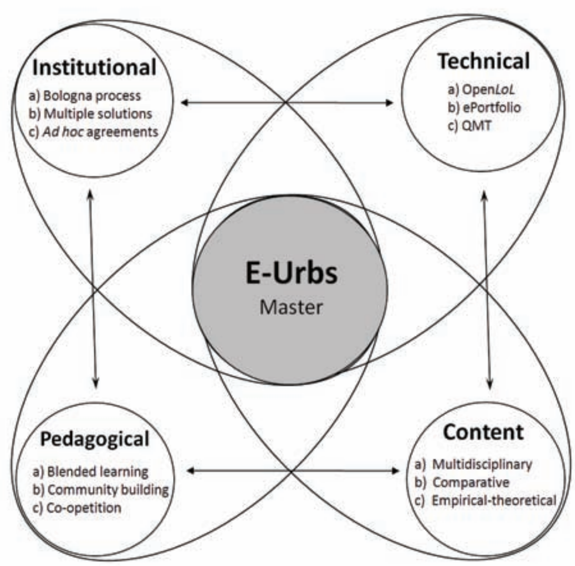
Last but not least, the diversities listed earlier challenge communication technologies. In fact, not all the students had the same IT skills and network performances differed (and still differ) among countries. When the online platform had to be accessed from Cameroon, India or the Netherlands there was the need to guarantee light and user friendly solutions. Moreover, the differences listed earlier required innovative online solutions in order to be addressed adequately and our staff members, students and professors become the beta testers of a new learning software[5] and methodology. For instance we developed co-opetitive (Brandenburger & Nalebuff, 1996; Lihui & Xiaowen, 2005) approaches to enhance distance learning efficacy. As the methodology was innovative and untested, it needed to be put in practice, used and assessed.

Most of the challenges described earlier are common to most virtual-campus-like arrangements and mainly relate to the need of managing differences in a knowledge-based society. In fact, new communication technologies provide us with the means to face the challenges that multicultural and diverse societies pose us in a way that would not have been possible before. Contemporary societies are extremely complex and fluid (Bauman, 2000; 2005; 2008) but they also show an incredible increase in possible ways to address this complexity.

BLENDING LEARNING: AN EFFECTIVE WAY TO MANAGE DIFFERENCES

Reducing the complexity, while keeping the added value of the existing differences, became a crucial objective in designing our learning model. In this second part of the chapter, we will address the challenges described earlier and illustrate how they have been dealt with, balancing the four dimensions of the model: 1) *pedagogical*; 2) *technical*; *3) institutional* and 4) *content*. Their balanced mix, shifts the meaning of the challenges from being critical issues to becoming resources for a mutual enrichment and a "*co-opetivive*" blended approach. In fact, in E-Urbs we addressed and analyzed the effects of specific pedagogical solu-

Figure 1. The E-Urbs Master learning model



tions in teaching and in the didactical organisation as well as the way in which technical solutions could be applied to content considering the countries' existing institutional frames in dealing with diversity.

Our assumption is that the four dimensions we identified need to relate to each other in a flexible yet integrated and balanced way in order to be useful. Keeping this in mind, in the following we will address each dimension, showing how the E-Urbs MA integrated them in an articulated learning model (see Figure 1).

The Pedagogical Approach

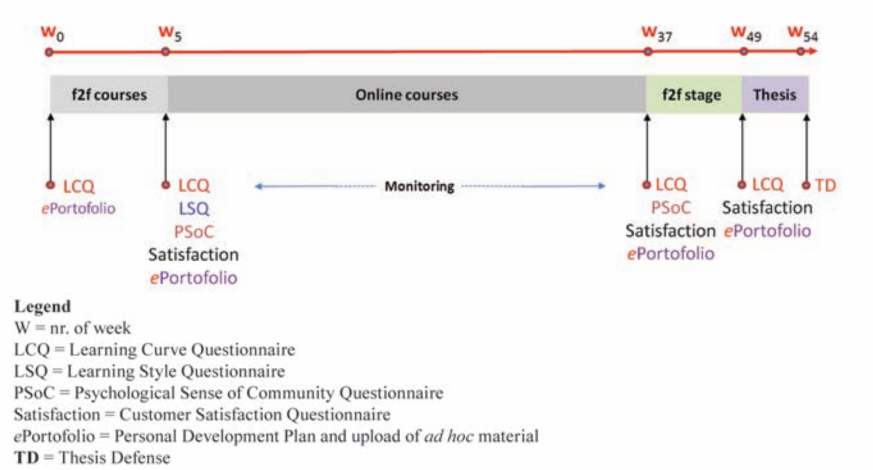
The pedagogical approach we adopted in the design of the E-Urbs Master aimed at integrating three interrelated interaction strategies and tools in order to contrast the problems common in online courses such as the declining attention and the increasing social distance among participants. The three strategies and tools were: a) a blended approach; b) the creation of a spontaneous learning community; c) a co-opetitive behavioural frame.

A Blended Approach

The blended approach (for a critical overview see Voos, 2003) we adopted in the learning path design, integrates f2f and online learning activities and is based on the assumption that both greatly benefit and complement each other. The reasons for this choice are related to three main needs: 1) some disciplines are more complex when they are taught online than others; 2) a truly European curricula might require too much physical mobility; 3) some travelling and transnational exchange was considered vital.

As a solution to these needs, we organised the learning path into three main periods (see Figure 2): 1) a first period (w1-w4) of intensive f2f teaching for a total of 15 ECTS, during which professors, tutors and students met and worked on basic concepts and methods of comparative urban analysis. In this way all stakeholders involved not only developed a common knowledge base, but

Figure 2. E-Urbs' blended learning's structure, monitoring instruments and timeline



also a companionship that supported them during the whole curricula; 2) a second period (w5-w36) of intense online teaching and distant collaborative learning for a total of 27 ECTS; 3) a third period (w37-w49) for a total of 18 ECTS spent at one of the partner Universities where students could work on their thesis and take advantage of the scholarly competences available within the network. The final defence and award ceremony took place in Urbino (w52) and allowed students to physically meet each other again, receive their diploma and confront their research experiences and expectations.

In order to measure the learning effectiveness of the blended approach proposed and to monitor the overall learning process, a monitoring and assessment working group was established[6]. One of its responsibilities was to submit different questionnaires to the students. These included also specific learning curve questionnaires, whose results were used as an indicator of the effectiveness of the learning processes during different periods.

The integrated blended approach adopted during the E-Urbs MA, helped to overcome many of the challenges that a virtual-campus-like arrangement normally has to face. Thanks to the f2f period – targeted to both the acquisition of a common background on comparative urban studies and homogenising the very different competences and interests among students – it became clear that the different disciplinary backgrounds constituted an enrichment with new potential perspectives through which to address urban issues, rather than a problem. The same applied to the age differences among participants. It was clear since the very beginning that the experiences of the older students were perfectly complemented by the more up-to-date studies and the enthusiasm brought in by the younger students. For this reason the management committee opened the Master course to both target groups (students and professionals).

Differences in culture were emphasised and stressed during the f2f meetings. Living together in Urbino, a small university town in Italy, allowed students to experience and to relate with people with very different cultural backgrounds. The small number of participants coming from the same cultural environment, encouraged a group formation dynamic based on other parameters than culture, such as disciplinary interest, personality or simply personal preference. Knowing each other helped substantially to overcome most of the inter-cultural classical stereotypes. This process brought about the rise of several multicultural groups inside the class that also helped to improve language skills, English being the only common language in all groups. Within this context, English was not just strictly used for learning purposes, but also for participating in social gatherings and to be part of the community. This mere fact improved motivation and the language skills of the group. During the first f2f period students had also the possibility to meet their tutors, who accompanied their learning experience during the whole Master programme. The learning community and companionship built during the first f2f month of activity continued and consolidated during the online teaching period. Thanks to this organisation, typical problems afflicting online learning, such us the lack of motivation, weak sense of learning community, etc. (Irizarry, 2002) were substantially reduced. Distance induced difficulties both technical and motivational were weakened by the fact that people already knew each other and could easily compensate the informational gap, normally given in pure online activities, with the previously acquired knowledge of the group and its dynamics. Technical difficulties were also easier to solve thanks to the fact that students and tutors already had the possibility to experiment with the online platform when studying together in Urbino.

The third part of the E-Urbs MA included a f2f interaction. During this time students were asked to move to one of the partners' institutions for their comparative thesis work. There, students experienced different institutional settings, procedures and uses, but at the same time they met scholars, tutors and classmates with whom they interacted online for almost half a year. In synthesis, the blended approach improved the creation and the maintenance of a (quasi)spontaneous learning community that revealed itself as a fundamental element for scholarship throughout the duration of the Master.

The Creation of a Spontaneous Learning Community

The creation of a stable psychological sense of community, which can be defined as the perceived sense of belonging to a specific learning community, played a very important role in overcoming many of the challenges of a virtual-campus-like arrangement.

The blended approach we used in designing the learning path was one of the pedagogical solutions we used in order to promote the creation and maintenance of the community. What we needed in addition, was a series of precise instruments in order to measure and monitor the processes involved in the construction of a learning community. The monitoring and assessment working group (Pigliapoco & Bogliolo, 2007) used two main indicators: the Classroom Community Scale, proposed by Rovai (2002) and based on a 20-item questionnaire and SCITT, and a scale introduced by Pigliapoco & Bogliolo (2007), based on a 6-item questionnaire. Both indicators take into account 4 dimensions:

- Spirit (friendship, cohesion, bonding among students);
- Trust (credibility, benevolence, confidence among students);
- Interaction (honesty in feedback, trust and safety among students);
- Commonality of expectations (commonality of the same learning goal).

The perceived sense of learning community was monitored twice, at the end of first f2f based period (see Figure 1, w5, PSoC) and at the end of the second online based period (see Figure 1, w37, PSoC), to see if there were important changes in the quantity and quality of relations related to the different teaching methods. Even if the psychological sense of community slightly decreased during online courses due to distance, it remained exceptionally high, as did the uniformity of feelings of all participants. Indeed, the quantity of relations among the students during the online teaching period were lower than during the f2f periods, while their quality, i.e., the intensity of friendship and interaction among students, was higher.

Co-Opetition

As we have already maintained, distance poses serious problems to the students' motivation and the development of a psychological sense of learning community necessary to share knowledge in a virtual-campus-like arrangement. In order to foster interaction and engagement and making course-attendance more rewarding, we tried to overcome inhibitors like "forced participation" without interaction and decreased motivation by developing a co-opetitive examination setting. Co-opetition is a methodology that combines cooperation and solidarity with competition in the same given context. Co-opetition has been first developed in a business-market oriented context and it focuses on cooperation between actors in an imperfectly competitive situation (Brandenburger & Nalebuff, 1996).

In the E-Urbs MA we tried to apply the methodology to exams in order to enhance learning efficacy through the promotion of solidarity and competition between students, transforming in this way the differences present among them in a further opportunity to learn from each other. The system was structured in a way that points were given not just to the people who were answering their individual multiple-choice questions correctly, but also to students that were helping others in answering open-questions. Students were organised in small groups of about 3-4 students each and points were given for individual, cooperative and collective outputs with truly Web 2.0 techniques.

The Methodological–Technical Dimension

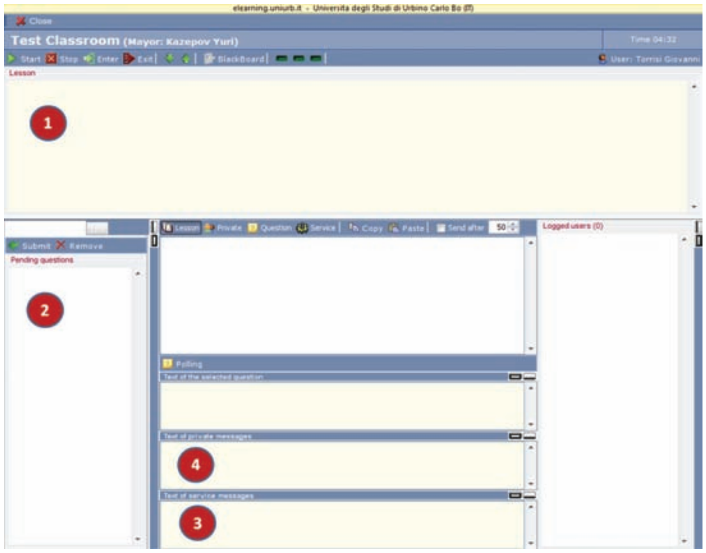
In order to be able to fully deploy the multifaceted pedagogical approach described earlier, we needed to have widely customisable technical instruments allowing us to organise the learning paths as they were developing. For this purpose we used *ad hoc* developed learning objects: a) OpenLol, an E-Learning platform; b) ePortfolio, a tool to promote self-reflection and strategic planning for students' personal and professional development; c) the QMT, the question management tool developed at the University of Urbino.

Open Land of Learning (OpenLol)

The partnership with an IT company (MEC Informatica) in the development of an open Web 2.0 learning platform was a crucial starting point. OpenLOL[7], a LMS based on Java and aimed at the delivery and use of broad and narrow-band Web 1.0 and 2.0 based services, was the result of several years of joint collaboration. The entire online environment has been localised in order to be used in several languages[8], giving everybody the possibility of integrating the familiar feeling of using one's own language in the learning process while keeping a common working language and helping to solve language difficulties.

OpenLoL is based on open standards like linux, mysql and java, and it is a modular environment in which modules and learning objects can be used and/or developed as needed. Among the modules that have been developed and used for the E-Urbs Master we find: a classroom, a shared repository, a forum, a visual lab, an ePortfolio and the Web 2.0 question management tools. The "Classroom" is based on an innovative concept of multi-layer interactive synchronous chat where courses or lessons can be held involving the interaction of professors, tutors, students, technicians

Figure 3. The OpenLol classroom



and learning advisors, yet keeping narrow band compatibility. This easily allows the overcoming of technical connection problems that may arise when people connect from narrow band connections around the world, as the classroom works mainly through layered text-based communication protocols.

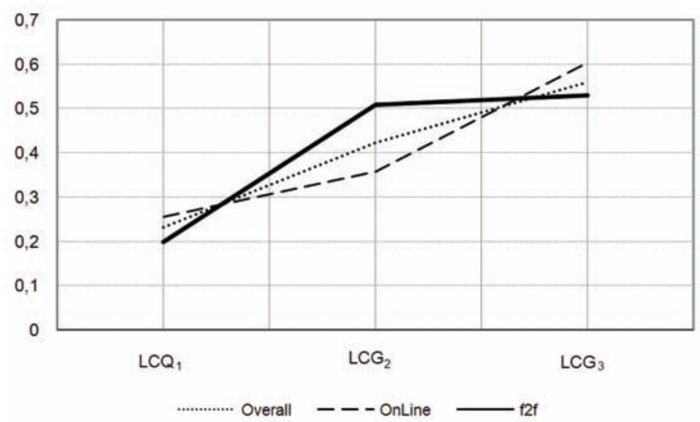
The Different windows of the classroom (see Figure 3) have different functions. In the lesson window (1) the content based interaction between scholars and students takes place. Tutors submit the questions asked by the students via the question window (2). The service window (3) is targeted to informal or service messages, while the private window (4) allows private p2p communication between participants. The coexistence of these parallel interaction layers reproduces a physical-like class arrangement and helps to booster the psychological sense of community and somehow to overcome the classical problems present in distance based courses such as fading motivation. At the end of a lesson, all text written in the lesson window is immediately available for downloading in RTF format, so that the participants may have the transcript of the relevant interaction. An interesting aspect of the classroom is the presence of a real-time blackboard on which users may draw graphs, write, paste pictures, slides, etc. The rest of the participants in the classroom are able to view and can interact on the blackboard in real time. This increases the efficiency of the interaction making explanations more visual.

The software includes other Web 2.0 features which have been used to foster collaborative learning strategies. Asynchronous interactions are taking place on thematic forums and a repository provides a virtual space where both the professors and the students can distribute teaching materials and share documents of any type. This repository becomes the central point of reference for sharing ideas, texts, data and information. Moreover, an online searchable database (called visual lab) helps with the organisation of multimedia content.

The ePorfolio

The ePortfolio is an educational tool used in many universities in Northern Europe and in the United States. It was developed as a methodology to address the growing demand for a personalisation of curricula, competence-focused learning and goal-oriented education. For this reason it is

Figure 4. The learning curve in the E-Urbs MA 2006-2007, adapted from (Pigliapoco et al., 2007)



mostly applied in advanced courses, especially Master programmes. With the Portfolio, a student can critically decide according to his/her goals to attend one particular course and, on the basis of this decision, build a structured archive of the most relevant materials collected during the course. Portfolio is, in synthesis, a methodological frame for self-reflection and strategic planning that some students found very useful in order to exploit the potential of a program for their personal and professional development. An ePortfolio has been included in the E-Urbs MA design aimed at providing students with some strategic career planning guidelines and a constant open channel with the organisation. Furthermore, we received very valuable information about the progress of the individual students, helping us, and the tutors in particular, to offer a more goal-oriented support.

Questionnaire Management Tools (QMT)

In order to manage properly a complex project like the E-Urbs MA virtual campus, there was the need to structure a regular feedback flow using various instruments – including non conventional ones[9] – at different points in time (see Figure 2). For this reason we used the QMT that colleagues from the Applied Informatics Department developed and integrated into the OpenLoL platform. The QMT has a double functionality: firstly, it allows instructors to create their own databases of questions and to build meta-questionnaires made up of questions randomly taken from the databases; secondly, it is a test management tool (TMT), allowing instructors to set-up exam calls, to administer both supervised written exams and unsupervised self-evaluation tests, feedback forms, and allowing students to fill in online questionnaires and to check their scores.

The QMT feedback management tool was used both for indirect and direct monitoring of the learning activity. Indirect monitoring was carried out by means of three types of questionnaires aimed at monitoring Customer satisfaction (CS), Learning styles (LS), and Psychological sense of community (PSoC). Direct monitoring was carried out by means of three types of questionnaires: Self-evaluation test (SE), exams, and Learning curve questionnaires (LCQ). Exams were setup by tutors and instructors as online tests made up of multiple-choice and open-text questions. The tests were submitted at the end of each teaching activity to evaluate students' preparation. Self-evaluation tests were made available to allow students to check their preparation and were associated with each lecture of each course. All self-evaluation tests were the same format as the exams.

As mentioned earlier, a learning curve questionnaire was prepared in order to monitor the effectiveness of the learning process considering the learning curve of the overall class at specific points in time. The questions were prepared with the contribution of all professors teaching in the Master and the learning curve results were calculated in terms of correct answers to a given number of randomly selected questions. Figure 4 shows the average scores obtained by students before the beginning of the Master (w0), after the first f2f period (w5) and at the end of third online period (w37) (see Figure 2 for timing and Figure 4 for results).

The scores confirm a general and decisive improvement of the learning outcomes. The line summarises the results of the overall questionnaire, while the dotted and dashed lines are calculated taking into account respectively the scores related to the f2f and online courses. As we expected, the dotted line grows much faster in the first part, while the dashed line grows faster in second one. There is also a significant cross-talk effect between f2f and online courses. This proves the coherent organisation of the overall programming of the Master with a strong correlation between the different topics of the course.

We used also customer satisfaction questionnaires, which were submitted three times to monitor the satisfaction of the stakeholders and the suitability of the proposed teaching methodology (see Figure 1).

Complementary to the other tools briefly described, a learning style classification questionnaire was used in order to infer how students learn and to make them aware of their learning attitudes and give them advice on how to take advantage of the teaching activities of the Master. Out of the different learning-style classification models which could be used for this purpose – e.g., Honey & Mumford, (1982; 2000), who use 40-80 questions to distinguish between activist and reflectors and between theorists and pragmatists; or by Kolb (1984), who uses 12 questions to classify students in accommodating, diverging, converging and assimilating – we adopted Felder's (2005) model, in which a 44-item questionnaire is used, and 4 dichotomised styles are identified: active/reflective, sensing/intuitive, visual/verbal, and sequential/global. For the results of these questionnaires see Pigliapoco et al., (2007).

The Way Content is Conveyed

In order to put forward an innovative and truly international curriculum, the contents of the course have been designed considering three interrelated aspects: a) the multidisciplinary competences existing within the network of universities; b) the complementarities of theoretical and empirical approaches based on a long-lasting joint-research activity carried on by the project partners; c) the comparative perspective informing all issues addressed from the content point of view. In designing the content of the learning path, we tried to balance these three aspects by organising a general introduction to the basic theoretical and empirical elements of urban studies during the first f2f period, followed by a second period online, characterised by more targeted and specialised topics (e.g., governance, quality of life, planning, etc.). Thanks to this organisation, the different disciplinary backgrounds of the students were considered to be consistent with the comparative spirit of the MA and they were an enrichment from the very beginning. Those students with less related disciplinary background (e.g., psychology) had to face more difficulties at the beginning but were able to catch up, also thanks to the crucial role of the local tutors, overcoming the difficulties and writing a good thesis. Also linguistic diversity among professors, tutors and students allowed the construction of a highly flexible yet powerful multi-linguistic educational model. English remained the working language for common interaction and study, while a variety of languages were used when needed. When comparative research is done, multi-linguistic skills are to be taken into account. Far from being just a problem, language diversity is also useful both for literature review of the social reality that is to be compared and for empirical research (interviews, questionnaires, etc). When language skills were not directly available local and thematic tutors were available.

The different teaching activities were organised according to a modular approach, whose main idea is based on breaking down educational content into small teaching units that can be reused in various learning environments, in the spirit of object-oriented programming. These learning objects are autonomous yet complementary units and can be used also in other learning contexts.

The Institutional Dimension

The institutional dimension was one of the most difficult challenges to deal with in the virtual campus organisation. This is particularly true when we consider the ambitious goal we had to institutionalise the international curricula. For this purpose we addressed three interrelated issues: a) the Bologna process frame; b) the adoption of multiple institutional solutions; c) partners' agreements.

The Bologna Process Frame

In order to design a truly European higher education learning path we wanted to ground it within the frame of the Bologna process so that we could contribute to the development of a European higher education area. From this point of view we needed to foster the approval of a joint E-Urbs MA degree. However, different national regulations and highly bureaucratised systems made this goal extremely difficult to attain. We organised several f2f project meetings, where we also invited representatives of the administration in order to better deal with the difficult legal and administrative issues we needed to address. This helped us partly to overcome the classical division that can be found in most universities between teaching and administrative staff, even though, despite all expectations, the existing legislation and universities internal procedures do not always comply with the Bologna process.

Multiple Institutional Solutions

In this context, all parties involved worked together with the common objective of finding the best possible solutions given the existing constraints related to legislation, bureaucracy and internal rules. Multiple solutions were found, ranging from a joint degree among two Universities, a double degree with other two universities and the mutual recognition of ECTS credits acquired during the MA course by all parties involved.

Partners' Agreements

From the administrative point of view, the multiple and flexible solutions mentioned under 'multiple institutional solutions' were not easy to achieve and required *ad hoc* agreements with the partners. The final diploma acknowledged the contribution of all institutions and the specific *ad hoc* solutions. Moreover, the E-Urbs consortium also issued the Diploma Supplement, in order to increase the transparency and recognition of qualifications across Europe in line with the Bologna Process. The latter offers important additional information on the learning path, describing the level and nature of the qualification attained by the students in order to ease the access to further studies and/or employment.

The Risks of Unbalanced Virtual Campuses

Keeping a dynamic balance can be very difficult and an accurate and strong coordination is needed in order to face and overcome the challenge of the existing differences, whether they are cultural, linguistic, and institutional. Considering one single dimension isolated from its synergic effect with the others would produce an unbalanced virtual campus and would affect the learning results. Focusing too much on didactics might produce excellent learning material, but if this is done without considering the synergic effect with the technical requirements of accessibility, students from African countries with no broadband connection, for instance, might not access videos or other multimedia content. Underestimating the institutional dimension, would not guarantee the achievement of an internationally recognised diploma degree, making the pedagogical results obtained less spendable (or not spendable at all) in the European context. If the technical dimension outweighs other aspects and contents tend to be left behind, students would have the means, but not enough value added in terms of knowledge. All this implies the following risks:

- Too much "pedagogy" → pedagogism: The pedagogical approach, and the consequent organisation of the learning activities, is a crucial dimension of the success of a virtual campus, but needs to be fine tuned in relation to the other dimensions to be effective in its objectives, otherwise it remains a mere expression of pedagogical theory.
- Too much "technology" → technicism: The belief that newer and more recent technologies are simply "better", might overestimate the role of the technical means in order to organise a virtual campus, fostering a blind acceptance of the online learning platform developments and completely forgetting the pedagogy that should direct its improvements.
- Too much "institution" → Bureaucratisa tion:Bureaucratisation puts an excessive attention on procedures and formalities, blocking potential institutional innovation. If every stakeholder within a large consortium is not willing to interpret its own rules and procedures with a problem-solving oriented attitude, the institutional building process entailed in the vision behind the Bologna process would not progress.
- Too much "content" → eclectic disarticulation: An excess of unstructured information would be highly ineffective. The added value of a well organised learning path relies on the fact that information is focused and needs to be integrated in its modules (pedagogy), transmitted (technology) and recognised as valid (institutional dimension).

In synthesis, the need to relate each dimension to one another becomes not only a crucial element of the blended approach we adopted in order to avoid the risks just described, but also to maximise the outcome in terms of quality and effective teaching.

ENRICHING THE BLENDED APPROACH WITH WEB 2.0 OPTIONS

As we maintained in the previous section, the flexible blended approach we adopted was crucial in overcoming the complex challenges faced by the E-Urbs Master course. Yet flexible integration of f2fs and online learning experiences are not always enough to address in an adequate way, some of the challenges involved in distance learning. From this point of view, the Web itself is going through a major change. The new emerging Web 2.0 characteristics are modifying the way in which information is treated and blended approaches improve substantially when integrated with a Web 2.0 perspective. The expression "Web 2.0" became famous after the first O'Reilly Media Web 2.0 conference in 2004. It does not refer to any technical update of the World Wide Web technologies, but it is a new way in which users can take advantages of the Web. The classical role division present in the mass media system between the information producer (i.e., writer, movie director, and singer), the information editor (i.e., editor, major, and publisher) and information user (i.e., reader, viewer, and end-user) suddenly tend to collapse. With youTube you can broadcast yourself without any filter; with Wikipedia you can write an encyclopaedia, not just read it; with del.icio.us you can directly share your bookmarks. No major, no editor, no publisher, no filters. All information is collectively created and immediately shared. Distribution of information loses its hierarchy and it becomes peer-to-peer based. Keywords of the Web 2.0 become wikis, p2p, blogs, folksonomies. As Högg et al., (2006) explained the Web 2.0 is a philosophy that mutually maximises collective intelligence, producing in this way added value for each participant by formalised and dynamic information sharing and creation.

Such a huge change in the way in which the Web is used and conceived could not have any impact on virtual-campus-like arrangements. The E-Urbs MA in Comparative Urban Studies, with students with very different backgrounds and age groups, eager to learn from each other and to share and compare their experiences, was the perfect opportunity to test a Web 2.0 approach in an E-Learning environment. E-Urbs changed traditional learning paradigms based on a clear hierarchy between the professor (who possesses the information) and student (who has to receive it), in order to take into account the new peer-to-peer learning paradigms in which the information increases its value and richness because each participant shares it, regardless of his/her statutory role. Albeit this major change in the way in which society treats information cannot be forgotten, a learning agency such as a university needs to retain some structure and integrity against the relativism of the informational value. Old learning paradigms need to be updated, but Universities remain the institutions where the means for processing information are given and where it is learned how to distinguish among the indefinite informational flow of information that the Internet presents to us.

In order to integrate Web 2.0 approaches and to address learning needs more adequately, as we have already seen in the previous paragraphs, we used several tools and methods: a) we incorporated a co-opetitive learning approach, as an innovative way of fostering p2p collaboration and motivation for achieving excellence; b) we developed a visual laboratory on “changing cities”; c) we used a repository for sharing documents, gray literature, provisional papers, etc.

A Co-Opetitive Approach

As described in the previous section, co-opetition is a methodology that combines cooperation with competition by contextualising and promoting both among students. In this way, existing differences are transformed in further learning opportunities, in the spirit of Web 2.0. Nonetheless, the method we adopted retains some more traditional elements since it is structured in a way in which students’ performances can be clearly assessed by the professors of the single courses. The co-opetitive exam we proposed, was made up of two parts: i) the first part with 5 multiple choice questions to be answered in 20 minutes on an individual basis; ii) the second part with one open question to be answered in a collaborative way within a given number of small groups. Important elements for the assessment of the outcome of the second part of the exam were: a) all members of the group have to agree on the answer; b) the division of the tasks should be discussed online within the group; c) the author of each part of the answer should be identifiable, i.e., we should be able to allocate the different parts to the different members; d) All interactions (i.e., including the discussion on the answering strategy) are tracked and considered in the assessment of the exam. Responses by the students to the learning methodologies we proposed were enthusiastic, not just for the excellent results they achieved and for the further consolidation of the psychological sense of the community they felt to belong to (as a side-product), but also because they were feeling that the over-all system was enriched by their own contribution and discussion.

The Visual Lab

The comparative Urban Visual Lab implemented in the E-Urbs project pushed the p2p 2.0 collaborative approaches even further. It allowed MA participants to develop collaborative and comparative projects using images of several cities (both European and non-European). The visual activities were focused on taking and collecting pictures that would illustrate the way in which different sociological concepts embody themselves differently across various cities. Participants were using visual tools to reach a deeper and richer understanding of urban societies and the respective metropolitan areas. Since the images produced were available in an open wiki-image-gallery, they were used not just by the Master participants, but also for teaching all over the world. Currently, in the E-Urbs MA the visual lab contains more than 9,000 pictures on more than 40 cities around the world uploaded by more than 100 MA and PhD students in Urban studies around the world and classified according to 9 main thematic issues relevant to urban studies[10].

The Repository

Inside OpenLoL, E-Urbs students had access to a digital repository where they could download most of their teaching materials. The peculiarity of this repository relied not so much on the instrument itself rather on the way in which we decided to use it, giving to each student the possibility to share his or her own resources, being those links, notes, pictures or draft papers. The Web 2.0 is not based on a major technological change, but much more on the way in which users make use of technology for collaboration. The repository is the digital place in which the concept "collaborative learning" became real, giving to the E-Urbs learning model a solid ground on which to base its collaborative memory.

FUTURE AND EMERGING TRENDS

As we have seen throughout the chapter, the results of integrating a blended learning approach with Web 2.0 features can be considered a viable model for experimenting and implementing new learning (and social) environments. Web 2.0 is changing the way in which information is constructed, (re) produced, distributed, consumed, allowing users to take advantage of the changing paradigms of the information society, in order to enhance the quality of the learning process. In synthesis, blended approaches improve substantially when integrated with a Web 2.0 perspective. This needs to be considered not just as an emerging trend, but a requirement for any future E-Learning development that has the ambition of being an innovative and effective means of education. The knowledge-based society has changed the way in which information is conveyed and learning agencies cannot simply avoid dealing with the issue.

Coherently, future trends in this domain will integrate electronic and paper based knowledge and between traditional and new paradigms of education. This will mean that most of the courses offered by Universities (or other learning agencies) are moving towards some form of blended 2.0 learning approach. Traditional face-to-face courses become each day more electronic dependent (assignments and research is done most of the time on the Web) as well as traditional online experiences need some physical attachment in order to be effective. So much differentiation is moving toward many variations of a single model in which physical presence remains fundamental but just if it is coupled with the e-means that allow students to navigate and process the fluidity of this new knowledge-based society.

As counter-intuitive as it may seem, the Web increased the number of flights instead of diminishing it, as virtual mobility has amplified physical mobility and computers (and printers) have boosted the amount of printed paper, the evolution of the Web 2.0 complexity and fluidity is calling for more structured and organised learning agencies that will have the responsibility not just of using, but also of educating people in dealing with the new information paradigms.

CONCLUSION

The main objective of the E-Urbs project was to contribute to the development of a higher education area in urban studies in Europe taking advantage of the use of virtual campus arrangements. The E-Urbs learning model, which characterises the virtual campus we described in this chapter balances different strategic dimensions like the institutional, technical, pedagogical and content related ones into a flexible blended approach with new co-opetitive Web 2.0 learning paradigms built in. The outcomes of the complex and interrelated arrangements resulting from the monitoring processes reported in the last two sections are rewarding and consistent with our initial hypothesis: a balanced approach taking into consideration the relevant dimensions not only adequately addresses the challenges of distant and transnational teaching arrangements, but also fosters a strong sense of psychological community that also improves learning effectiveness by building a collaborative learning culture.

REFERENCES

Alberto, A., & Magalhães, A. (2002). Epidemiology and the Bologna saga. *CHER Annual Conference 2002*, Vienna 5-7 September.

Bacsich, P. (2004). *The e-university compendium*. London, UK: Higher Education Academy.

Bakutes, A. P. (1998). An examination of faculty development centres. *Contemporary Education*, *69*(3), 168–171.

Barr, R. B., & Tagg, J. (1995). From teaching to learning: A new paradigm for undergraduate education. *Change*, *27*(6), 13–25.

Bauman, Z. (2000). *Liquid modernity*. Cambridge: Polity Press.

Bauman, Z. (2005). *Liquid life*. Cambridge: Polity Press.

Bauman, Z. (2008). *Liquid fear.* Oxford: Blackwell Publishers.

Berge, Z. L. (1998). *Changing roles of teachers and learners are transforming the online classroom*, Online-Ed. Retrieved on August 30, Doc. No. 74 from: http://www.edfac.unimelb.edu.au/online-ed

Berge, Z. L., & Collins, M. P. (1995). *Computer mediated communication and the online classroom*, Volumes I, II and III, Cresskill, NJ: Hampton Press.

Brandenburger, A., & Nalebuff, B. (1996). *Co-opetition: A revolution mindset that combines competition and cooperation.* New York: Currency/Doubleday.

Carr, S. (2000). As distance education comes of age, the challenge is keeping the students. *The Chronicle of Higher Education*, *46*, A39–A41.

Cereijo, M. V. P., Young, J., & Wilhelm, R. W. (2001). Factors facilitating student participation in asynchronous web-based courses. *The Journal of Computing in Teacher Education*, *18*(1), 32–39.

Coffield, F., Moseley, D., Hall, E., & Ecclestone, K. (2004). *Learning styles and pedagogy in post-16 learning: A systematic and critical review*. www.LSRC.ac.uk: Learning and Skills Research Centre. Retrieved January 15, 2008 from: http://www.lsda.org.uk/files/PDF/1543.pdf

Colis, B., & Moonen, J. (2001). *Flexible learning in a digital world: Experiences and expectations.* London: Kogan-Page.

Collins, M. (1999). I know my instructional technologies: It's these learners that perplex me! *American Journal of Distance Education*, *13*(1), 8–23.

DeVillar, R. A., & Fallis, C. J. (1991). *Computers and cultural diversity: Restructuring for school success.* Albany, NY, Sunny Press.

Draffan, E. A., & Rainger, P. (2006). A model for the identification of challenges to blended learning, in *ALT-J. Research in Learning Technologies*, *14*(1), 55-67

Dziuban, C., & Moskal, P. (2001). Evaluating distributed learning in metropolitan universities. *Metropolitan Universities*, *12*(1), 41–49.

Felder, R. M., & Silverman, L. K. (1988). Learning and teaching styles in engineering education. *Engineering Education, 78*(7), 674–681. Preceded by a preface in 2002. Retrieved on April 17, 2006, from: http://www.ncsu.edu/felderpublic/Papers/LS-1988.pdf

Felder, R. M., & Soloman, B. A. (1997). *Index of learning styles questionnaire*. Retrieved on April 22, 2006, from http://www.engr.ncsu.edu/learningstyles/ilsweb.html

Felder, R. M., & Spurlin, J. (2005). Applications, reliability and validity of the index of learning styles. *International Journal on Engineering Education*, *21*(1), 103–112.

Fletcher, G. H. (2004). [Technological Horizons In Education]. *T.H.E. Journal*, 32.

Forsyth, I. (1996). *Teaching and learning materials and the internet*. London: Kogan Page Ltd.

Gardiner, L. (1994). Redesigning higher education: Producing dramatic gains in student learning. [Washington, DC: George Washington University.]. *ASHE-ERIC Higher Education Report*, 7.

Hara, N., & Kling, R. (2001). Student distress in web-based distance education. *EDUCAUSE Quarterly*, *3*, 68–69.

Hiltz, S. R. (1994). *The Virtual classroom*. Norwood, NJ: Ablex.

Högg, R., Meckel, M., Stanoevska-Slabeva, K., & Martignoni, R. (2006). Overview of business models for web 2.0 communities. *Proceedings of GeNeMe*, (pp. 23-37).

Honey, P., & Mumford, A. (1982). *Manual of learning styles.* London: P. Honey.

Honey, P., & Mumford, A. (2000). *The learning styles helper's guide*. Maidenhead: Peter Honey Publications Ltd.

Irizarry, R. (2002). Self-efficacy and motivation effects on online psychology student retention. *USDLA Journal*, *16*(12), 55–64.

Kok, W. (2004). Facing the challenge. The Lisbon strategy for growth and employment. *Report from the High Level Group*, retrieved on March 3, 2007 from: http://ec.europa.eu/growthandjobs/pdf/kok_report_en.pdf

Kolb, D. A. (1984). *Experiential learning*. Englewood Cliffs, NJ: Prentice-Hall.

Lihui Sun., & Xiaowen, Xu. (2005). *Coopetitive game, equilibrium and their applications,* Berlin: Springer.

Marino, T. A. (2000). Learning online: A view from both sides. *The National Teaching & Learning Forum*, *9*(4), 4–6.

McCarthy, M. E., Pretty, G. M. H., & Catano, V. (1990). Psychological sense of community and student burnout. *Journal of College Student Development*, *31*, 211–216.

Melton, R. (1997). *Objectives, competencies and learning outcomes: Developing instructional materials in open and distance learning*. Stirling, VA, Kogan Page Ltd.

Morgan, C. K., & Tam, M. (1999). Unraveling the complexities of distance education student attrition. *Distance Education*, *20*(1), 96–108. doi:10.1080/0158791990200108

Munro, R., & Munro, E. (2004). Learning styles, teaching approaches and technology. *Journal for Quality and Participation*, *27*(1).

O'Reilly, T. (2006). *Web 2.0 compact definition: Trying again.* Retrieved on January 20, 2007 from: http://radar.oreilly.com/archives/2006/12/web-20-compact-definition-tryi.html

Pigliapoco, E., & Bogliolo, A. (2005). Global accessibility of higher education: Using ICT to build a worldwide campus. *In Proceedings of EISTA (EISTA-05).*

Pigliapoco, E., Lattanzi, E., & Bogliolo, A. (2007). Non-conventional online feedback tools for learning process monitoring. *Paper presented at the international conference "Challenging Boundaries. The Role of ICT in the creation of a European higher Education Area.* Urbino 15th November.

Privateer, P. M. (1999). Academic technology and the future of higher education: Strategic paths taken and not taken. *The Journal of Higher Education*, *70*(1), 60–79. doi:10.2307/2649118

Rovai, A. P. (2002). Development of an instrument to measure classroom community. *The Internet and Higher Education*, *5*(3), 197–211. doi:10.1016/S1096-7516(02)00102-1

Sanchez, I., & Gunawardena, C. N. (1998). Understanding and supporting the culturally diverse distance learner. In C.C. Gibson (Ed.), *Distance Learners in Higher Education*. Madison, WI: Atwood.

Schifter, C. C. (2000). Faculty participation in asynchronous learning networks: A case study of motivating and inhibiting factors. JALN, 4*(1)* - June 2000.

Sewart, D., Keegan, D., & Holmberg, B. (1983). *Distance education: international perspectives.* New York: Croom Helm.

Taylor, J. C., & White, J. V. (1991). Faculty attitudes towards teaching in the distance education mode: An exploratory investigation, *Research in Distance Education*, July, 7-11.

Voos, R. (2003). Blended learning: What is it and where might it take us? *Sloan-C View*, *2*(1), 2–5.

ENDNOTES

[1] The MA has been jointly organised with the following institutions: University of Urbino (Italy) (Coordination); Sako Musterd, Johan Post and Marco Bontje (University of Amsterdam, Netherlands); Marisol Garcia (University of Barcelona, Spain); Hartmut Haeussermann (Humboldt University at Berlin, Germany); Enzo Mingione and Giampaolo Nuvolati (University of Milan-Bicocca, Italy); Alan Murie and Rob Rowlands (University of Birmingham, United Kingdom); Hans Thor Andersen (University of Copenhagen, Denmark); Chris Kesteloot, (Catholic University of Leuven, Belgium); Grzegorz Weclawowicz (Polish Academy of Sciences, Poland). The arguments presented in this chapter are based on data gathered within the E-Urbs Master (www.e-urbs.net)

and all partners are to be gratefully thanked for their support throughout the project.

2 The project was funded for 24 months (January 2006-2008) under grant agreement: 2005-3870/001-001ELE ELEB12E.

3 For the basic documents of the Bologna process, see: http://www.sociologiadip.unimib.it/unimon/ Unimon is a European project aimed at monitoring the harmonisation processes of tertiary education in EU countries. For a critical perspective, see Amaral 2002.

4 Participants have been selected on the basis of an application procedure. Students had to include 2 recommendation letters, all relevant information on their previous studies, a statement on their knowledge of English and a motivation statement of why they wanted to embark on a Masters programme in comparative urban studies.

5 Mec Informatica is a software house based in Rome (www.mecinformatica.it), interested in the development of a LMS for online teaching and training activities. We had the opportunity to collaborate in the development of Land-of-Learning (LoL), building-in all features we considered relevant for a high level university online teaching. From 2008 onwards LoL will be made available as open source software.

6 The working group was headed by Alessandro Bogliolo from the Applied Informatics Department of the University of Urbino "Carlo Bo."

7 For further information see www.landoflearning.it

8 The software localisation was carried out in the languages spoken by the partner universities: Italian, Spanish, German, Dan-

This work was previously published in Institutional Transformation through Best Practices in Virtual Campus Development: Advancing E-Learning Policies, edited by M. Stansfield; T. Connolly, pp. 215-234 , copyright 2009 by Information Science Reference (an imprint of IGI Global).

Section VI
Managerial Impact

This section presents contemporary coverage of the managerial implications of Web technologies. Particular contributions address Web software engineering and Web-enabled employee life-cycle process management. The managerial research provided in this section allows executives, practitioners, and researchers to gain a better sense of how Web technologies can inform their practices and behavior.

Chapter 6.1
Enterprise 2.0:
Collaboration and Knowledge Emergence as a Business Web Strategy Enabler

Javier Soriano
Universidad Politécnica de Madrid, Spain

David Lizcano
Universidad Politécnica de Madrid, Spain

Marcos Reyes
Telefónica I+D, Spain

Fernando Alonso
Universidad Politécnica de Madrid, Spain

Genoveva López
Universidad Politécnica de Madrid, Spain

ABSTRACT

The Web is becoming in many respects a powerful tool for supporting business strategy as companies are quickly becoming more and more reliant on new Web-based technologies to capitalize on new business opportunities. However, this introduces additional managerial problems and risks that have to be taken into consideration, if they are not to be left behind. In this chapter we explore the Web's present and future potential in relation to information sharing, knowledge management, innovation management, and the automation of cross-organizational business transactions. The suggested approach will provide entrepreneurs, managers, and IT leaders with guidance on how to adopt the latest Web 2.0-based technologies in their everyday work with a view to setting up a business Web strategy. Specifically, Enterprise 2.0 is presented as a key enabler for businesses to expand their ecosystems and partnerships. Enterprise 2.0 also acts as a catalyst for improving innovation processes and knowledge work.

INTRODUCTION

There is no doubt that the Web is in many respects a powerful tool for supporting business strategy. Emerging Internet technologies continue to enable businesses to expand their ecosystems and partnerships. This, however, introduces additional managerial problems and risks that have to be taken into consideration to avoid being left behind.

This chapter explores the Internet's present and future potential in relation to information sharing, knowledge management, innovation management, and the automation of cross-organizational business transactions. It points out how a business Web strategy that takes into account this potential will help not only to improve the existing information sharing and knowledge management processes, but also to protect investments in technology that would otherwise have resulted in expensive failures and severe losses. The suggested approach is based on the emerging Web 2.0 vision and will help to minimize the risk of key information and knowledge being lost or simply not being available on time for the stakeholder, projects started and never finished, worse time-to-market, results not meeting expectations, failure of global, cross-organizational IT integration processes, or even incoherencies between technology and company strategy or structure and so on (Argyris, 1998, pp. 98-105). All managers, and particularly IT leaders, must be aware of this new potential and its implications in order to come up with innovative and effective answers to both known and new problems related to information sharing and knowledge management within their organizations (McAfee, 2006).

The chapter's contents are designed to guide entrepreneurs, managers, and IT leaders through the adoption of the latest Internet technologies, such as Web 2.0, Enterprise 2.0, and the global service oriented architecture (SOA), and their application to their everyday work with a view to setting up a business Web strategy. Musser and O'Reilly (2006) claim that by defining and following a set of architecture building blocks, architectural design decisions, and normative guidance, they can build flexible, extensible, and reusable solutions for exploiting the best features of the emerging Web 2.0 technology suite to achieve the best return on investment (ROI) by leveraging the upcoming Web of user-centered services.

BACKGROUND: THE ADVENT OF ENTERPRISE (WEB) 2.0

There are several different definitions of Web 2.0 (a.k.a. *social networking*) that mostly only describe certain aspects of the overall concept. Tim O'Reilly (2007), who originally coined the term, initially identified seven major characteristics inherent to the Web 2.0 concept. First, the Web is considered as a platform for building systems that do not necessarily have a specific owner and are "tied together by a set of protocols, open standards and agreements for cooperation." Harnessing Web users' collective intelligence represents the second major paradigm. This promotes architecture by participation and democracy and encourages users to add value to the application as they use it. The ownership of mission-critical data is regarded a further cornerstone of numerous Web 2.0 applications. Fourth, O'Reilly propagates the end of the software release cycle as another central paradigm. The use of lightweight programming models that allow for loosely coupled systems and applications, the provision of software above the level of a single device, and the realization of rich user experience represent the last major paradigms inherent to the Web 2.0 concept. Besides such analyses that properly describe parts of the super-ordinate concept, there are only very few comprehensive scientific definitions available. An in-depth investigation of numerous different, successful Web 2.0 applications conducted by Högg, Meckel, Stanoevska-Slabeva, and Martignoni (2006) condensed the respective characteristics into the following statement, which serves as underlying definition for this chapter: "Web 2.0 is defined as the philosophy of mutually maximizing collective intelligence and added value for each participant by formalized and dynamic information sharing and creation" (pp. 23-37).

The Enterprise (Web) 2.0 concept (henceforth referred to as *Enterprise 2.0*) is related to its big brother Web 2.0, because, to a certain extent, it can be viewed as many existing Web 2.0 consumer-oriented services maturing to include features that are important for enterprise users. Enterprise 2.0 represents on its own a new paradigm in which employees, regarded as *knowledge workers*, are coproducers of content, knowledge, applications, and services. Therefore, there is an imperious need to revisit and reconsider the very definition of knowledge worker during this chapter, because it is used extensively throughout the chapter from Davenport's (2005) viewpoint, that is, to refer to employees, partners, suppliers, customers and other possible stakeholders. Enterprise 2.0 provides enterprises with new models and tools for emergent collaboration and cocreation. Enterprise collaboration is thus being enhanced by virtual communities that leverage social linking and tagging tools (e.g., tools for social networking, social bookmarking, and social search), user-contributed content management platforms (e.g., enterprise wikis, blogs, and forums), tools that leverage user opinions (e.g., tools supporting comments and voting), subscription-based information distribution tools (e.g., Enterprise really simple syndication [RSS] feeds), user-centered services (e.g., mash-up- and pipe-based services), and so forth (Drive et al., 2004).

These digital platforms are already popular on the Internet, where they are collectively labeled Web 2.0 technologies (Laso, 2006). Now though, a number of Enterprise 2.0-based collaboration platforms are beginning to proliferate. These platforms are aimed at providing enterprises with specialized subsets of these out-of-the box capabilities (Coveyduc, Huang, Ostdiek, & Reif, 2006; Intel Corporation, 2006). These new collaboration platforms provide enterprises with an ecosystem of knowledge workers who collaborate to develop capabilities by collectively generating, sharing, and refining information, business knowledge, and services. Enterprise 2.0 collaboration enables firms to leverage desirable Web 2.0 attributes, including harnessing collective intelligence and architecture by participation.

The user's production is now primarily based on the customization, composition, remix, and reuse of existing material, such as sampling or mash-ups, by the employees themselves. Enterprise 2.0 technologies have the potential to usher in a new era by making both information sharing, service provision, and consumption, and knowledge work practices and their outputs more visible. At the same time, they provide support for the extraction and the emergence of both knowledge and organizational structure.

In representation of the Gartner Group analyst firm, Smith (2006) recently predicted that by 2008 the majority of Global 1,000 companies will adopt several technology-related aspects of Web 2.0 to advance their businesses. As companies quickly increase their reliance on new Web-based technologies to capitalize on new business opportunities, the industry is showing greater demand for technology experts who can build and manage Web 2.0 resources, including blogs, wikis, forums and user groups, and mash-up enabler tools to centralize the management of all of these resources, supporting technology and knowledge experts' work.

RESEARCH FRAMEWORK AND OBJECTIVES: FOSTERING INNOVATION THROUGH AN ENTERPRISE 2.0-BASED BUSINESS STRATEGY

As global market opportunities, competition, and availability of human resources increase, enterprises are assigning high strategic priority to speeding up innovation, even by modifying their own business and global presence strategies. Enterprises want to speed up innovation to improve their market and business opportunities in the fierce global competition by collaborating and cocreating with partners and users (Coveyduc et al., 2006). The desired outcome is improved time-to-market and increased value of more new products and services. The emergence of Enterprise 2.0 Web-based platforms provides enterprises with new models and tools for collaboration and cocreation.

Enterprise collaboration can be fostered by virtual communities that leverage user content sites (e.g., Wikipedia, Flickr), social linking, tagging, and searching tools (e.g., MySpace, del.icio.us), and sites that leverage opinions of all who participate (e.g., Amazon ratings). Enterprises already leveraging cocreation are eBay API and SalesForce.com AppExchange. These platforms, as Weill and Ross (2004) note, provide enterprises with an ecosystem of partners, suppliers, and customers collaborating to develop capabilities by integrating knowledge and services. Enterprises want a solution that delivers these capabilities out of the box.

The key idea behind Enterprise 2.0 vision, and the lesson many businesses must learn, is that next generation IT systems must be conceived to acquire the knowledge they operate on directly from who really has it, that is, the employees (seen as knowledge workers) and from the operation and communication processes employees enter into (Morris, Pohlmann, & Oliver, 2005). The knowledge of a business has less to do with the IT infrastructure than with the employees themselves. The IT infrastructure must be capable of extracting and managing that knowledge for it to evolve and adapt to the business processes. Any other means to model and exploit the business knowledge will never be flexible enough. If user knowledge changes (and it does change), the IT infrastructure must seamlessly adapt to such changes.

In any case, the design of both traditional and Enterprise 2.0-based solutions has focused primarily on creating a structure that supports common processes and stores information to assure that it is easy to find, reliably available, and backed up. They all have been conceived under the premise that teams need to focus on their core business rather than IT issues. The entire operating environment has therefore been traditionally subordinated to IT departments. Nevertheless, this approach has proved to have a number of collaboration-related drawbacks that slow down the pace of innovation. Knowledge workers are thoroughly acquainted with routine procedures and are capable of extracting automatic behavior, suggesting improvements on the IT systems they use through their operating environment and, more importantly, innovating new operating procedures. Operational innovation is an essential requirement in today's competitiveness-driven business markets, enhancing as far as possible collective intelligence-based knowledge work.

With this in mind, this chapter's main objective is to elaborate on the synergies the Web 2.0 concept and several IT technologies have with regard to the enterprise innovation. Web 2.0's focus on the inclusion of human beings and the exploitation of users' collective intelligence is considered a key enrichment of the knowledge emergency within enterprises. This research objective will be contrasted with relevant literature reviews. The remainder of the chapter is structured as follows. First of all we revisit the notion of knowledge worker and its duty, analyzing its features, needs, and problems they find in their daily innovation effort. In this section we elaborate on factors that can facilitate or instead inhibit a knowledge work process. We then present solutions for fostering enterprise innovation based on Enterprise Web 2.0 ideas and technologies, a pragmatic guideline recommending the alignment of the business strategy exploiting Enterprise 2.0 advantages and the application of this idea to other key Web strategy areas. Finally, the final section concludes this chapter and presents a brief outlook on future trends.

ISSUES, CONTROVERSIES AND PROBLEMS IN ENTERPRISE COLLABORATION AND KNOWLEDGE EMERGENCY

Knowledge Work Revisited: Novel Ways to Foster Innovation through Social Capital and Collective Intelligence

Collective intelligence has existed for at least as long as humans have. Ancient social groups, nations, and modern corporations all act collectively with varying degrees of intelligence. But this ancient phenomenon called innovation emergence is now occurring in dramatically new forms. With new communication technologies, and especially the Internet, huge numbers of people all over the planet can work together in previously unsuspected ways. For this reason, it is more important now than ever before to have an in-depth understanding of collective intelligence to be able to create and take advantage of the new possibilities. Our current definition of collective intelligence is "*groups of individuals doing things collectively that seem intelligent*" (Davenport, 2005).

In general, "collective intelligence" is a *perspective* that can be applied to many different kinds of phenomena. For instance, this perspective suggests another way of thinking about things like "firm productivity," "organizational effectiveness," "teamwork," "firm profitability," and "leadership." When people hear the term "collective intelligence," they tend to assume that it implies individuals giving up their individuality to be somehow subsumed in a group or team. This is *not* what we mean. Collective intelligence, as we understand and explore it, is not about false consensus, hive minds, cults, or groupthink. Collective intelligence relies upon and emerges from a synergy between the individual knowledge, creativity, and identity of its constituent parts (Brown & Duguid, 2000). In its highest forms, participating in collective intelligence processes can actually help people self-actualize while solving collective problems. This collective intelligence is developed within enterprises by the innovative engines called knowledge workers.

As we mentioned above, nowadays enterprises need to accelerate innovation to improve their market and business opportunities in global competition, and therefore it becomes essential to revisit features, profiles, and characteristics of key actors in every innovation process, hence, the knowledge workers. This will help in understanding of the relevance of Enterprise 2.0 technologies and models for both improving their work and expanding their productivity, as we will tackle in subsequent sections.

Knowledge Workers. Mission, Relevance and Novel Ways to Improve their Work

There is a range of ideas about what *knowledge workers* are and what characterizes them. Some examples are:

- "The term *knowledge worker* was coined by Peter Drucker some thirty years ago to describe someone who adds value by processing existing information to create new information which could be used to define and solve problems. Examples of knowledge workers include lawyers, doctors, diplomats, law makers, marketers, software developers, managers and bankers." (Fallows, 2005)

- "Knowledge workers use their intellect to convert their ideas into products, services, or processes" (Davenport, 2005).
- "Their main value to an organization is their ability to gather and analyze information and make decisions that will benefit the company. They are able to work collaboratively with and learn from each other; they are willing to take risks, expecting to learn from their mistakes rather than be criticized for them." (Davenport & Prusak, 1997)
- "Knowledge workers are continually learning, aware that knowledge has a limited shelf life" (Davenport & Prusak, 2000).

What then is a knowledge worker?

- A problem solver vs. a production worker
- A person who uses intellectual rather than manual skills to earn a living
- An individual who requires a high level of autonomy
- A manipulator of symbols; someone paid for quality of judgment rather than speed of work
- A worker who uses unique processes
- Someone who possesses uncodified knowledge which is difficult to duplicate
- A worker who sources between the ears
- Someone who uses knowledge and information to add to deeper knowledge and information

Fewer and fewer people are subordinates, even in fairly low-level jobs; increasingly they are knowledge workers. Knowledge workers cannot be managed as subordinates; they are associates. The very definition of a knowledge worker is one who knows more about his or her job than anyone else in the organization (Davenport & Harris, 2007).

According to Nonaka and Takeuch (1995), what motivates workers—especially knowledge workers—is what motivates volunteers. Volunteers have to get more satisfaction from their work than paid employees precisely because they do not get a paycheck. They need, above all, challenge. They need to know the organization's mission and to believe in it. They need continuous training. They need to see results. Implicit in this is that employees have to be managed as associates and/or partners, and not in name only. The definition of a partnership is that all partners are equal.

Davenport (2005) elaborates on this idea when stated that the productivity of the knowledge worker is still abysmally low. It has probably not improved in the past 100 or even 200 years for the simple reason that nobody has worked at improving the productivity. All our work on productivity has been on the productivity of the manual worker. The way one maximizes their performance is by capitalizing on their strengths and their knowledge rather than trying to force them into molds.

Types of Knowledge Workers

From the practical perspective outlined in this chapter, it can be very useful to consider three separate types of knowledge worker: "core knowledge workers," "*high-end* knowledge workers," and "everyone else."

- Core knowledge workers are those in specific "knowledge management" roles and enterprise duties. Examples of these kinds of roles include chief information/knowledge officers, librarians, knowledge managers, content managers, knowledge analysts, information officers, and so forth.
- "High-end" knowledge workers, or those with the highest degree of education and expertise, would seem to be particularly important to enterprise innovation. They are the scientists who develop the new products, the professionals who plan and sell the big consulting or legal projects, or the hardware or software architects who

envision and deliver the new product line. In the knowledge economy, these should be the horses that pull the plow (the people to whom we should look for the new ideas, products, and services that fuel revenue growth and ensure organizational longevity) (Davenport & Harris, 2007).

- Everyone else is all the other knowledge workers (e.g., dentists, doctors, nurses, managers, pharmacists, technicians, administrators, etc.). In short, everyone in the organization engaged in some form of "knowledge work."

Of course, there is not always a clear dividing line between these classes, but the distinction can be a helpful one at the start. It can be particularly useful for helping people to understand that everyone in a company is a knowledge worker to some degree, and knowledge work is everyone's responsibility, not just that of a few people with "knowledge" or "information" in their job title.

Features of HEKWs (High-End Knowledge Workers)

1. They control their own work structure; high-end knowledge work remains relatively unstructured and autonomous. No one generally tells these workers where to work, when to work, or what specific tasks to perform during work.
2. They are highly collaborative.
3. They work in multiple settings.
4. They do individual and group work.
5. They have high levels of passion, power and occupational mobility.

Knowledge Types Managed by Knowledge Worker

Here is one classification for different types of knowledge.

- **Logical:** There is knowledge that is the result of the understanding of how ideas relate to one another in a domain.
- **Semantic:** There is knowledge that is the result of learning the meaning of words or concepts. Knowledge of words is knowledge of definitions. Such definitions are set out in dictionaries. You can look this knowledge up.
- **Systemic:** There is knowledge of mathematics and geometry, for example, which is the result of learning a system of words, or symbols, and how they relate to one another and the rules for operating in that system. Any claims made that are consistent with those definitions and rules are called knowledge.
- **Empirical:** There is knowledge that comes through our senses. This is empirical knowledge. Science is the best example of a method for ascertaining the accuracy of such knowledge. Scientific knowledge is a result of practicing the scientific method, that is, observation, abduction of a hypothesis, careful observation, refinement of the hypothesis, deduction of a test for the hypothesis, testing and experimentation, and confirmation or falsification of the hypothesis.

In addition, knowledge can be viewed from another point of view as implicit and explicit knowledge.

Informative Channels used by Knowledge Workers

Traditional knowledge management programs attempt to manage the process of creation or identification, accumulation, and application of knowledge or **intellectual capital** across an organization.

By exploiting several informative channels, the guidelines of core knowledge workers can be made accessible for everyone else in the company. This approach looks to emerge and apply collaborative and social knowledge to create a **social capital** across the organization (Lin, Burt, & Cook, 2001).

The following is a list of informative channels used by core knowledge workers in Enterprise 2.0:

- e-mails
- chats
- blogs
- RSS feeds
- portal or Web content
- links and reverse references (links in blogs)
- wiki tools
- folksonomies
- bookmarks (tagged or not)
- documents of every kind, including files in ftp, printed papers, and so forth
- physical communication in person
- physical communication in distance (e.g., phone)
- common applications as enterprise mash-ups

Factors in Knowledge Work

Facilitating Factors

Bloom (2000, pp. 42-44) identifies the following five elements as causing a group to be intelligent (a "collective learning machine"):

1. **Conformity enforcers:** Mechanisms that cause consensus and similarities among most members of the group.
2. **Variety generators:** Mechanisms that cause some differences and discussion among members of the group.
3. **Inner judges:** Mechanisms that cause individual members of a group to reward themselves for successes and to punish themselves for failures, and cause everyone to evaluate a concept or idea, and validate it after their own experience-based verification.
4. **Resource shifters:** Mechanisms that shift resources (e.g., admiration, information, data, concepts, knowledge, money, or influence) to members of the group.
5. **Intergroup tournaments:** Competitions between subgroups or departments (such as games, corporate competitions, rivalry discussions, etc.)

Other authors, like Surowiecki (2005), say that there are three conditions for a group to be intelligent (for a "crowd to be wise"):

1. **Diversity:** The group includes members with a wide diversity of knowledge or abilities (and the ability to recognize successful and unsuccessful outcomes).
2. **Independence:** Group members use their own knowledge and abilities without being overly influenced by others. (When group members have too much influence over each other, various kinds of bad outcomes can result. See inhibitory factors section below.)
3. **A particular kind of decentralization:** Group members' actions are *aggregated* in a way that finds the right balance between: (a) "making individual knowledge globally and collectively useful," and (b) "still allowing it to remain resolutely specific and local."

Inhibitory Factors

Finally, there are several general factors that can inhibit collective intelligence, such as groupthink and informational cascades, social dilemmas, coordination failures, or failures in thinking itself (Malone, Jay, Legay, & Kosorukoff, 2006).

The first barrier to collective intelligence is called groupthink and social conformity, which was described by Janis (it is perhaps the best explored factor) and developed further in numerous experimental studies. The key point of this research is that people's tendencies to conform, imitate, and avoid conflict can reduce the effective diversity of opinions, and lead to judgments and decisions that are inaccurate, premature, systematically biased, and so forth. The analogue to this phenomenon in pragmatic distributed collective intelligence, as James Surowiecki points out in *The Wisdom of Crowds*, is the informational cascade, where imitation produces fads and conformity instead of individual decision making. In other words, the knowledge of the whole turns out to be less than the sum of the parts, because ***only some parts are actually contributing while everyone else conforms or imitates.*** For this reason, mechanisms that foster diversity and independence might improve collective intelligence. At the interpersonal level, this means practices and norms surrounding respect for individual ideas and contributions (as in the early stages of a brainstorming session). At a distributed level, this results in structural barriers in the physical, legal, or IT "code" (e.g., the walls around a voting booth).

An excellent example based on heuristic experiments of how early decisions by some group members can unduly bias the decisions of later group members is the study of online music ratings by Salganik, Dodds, and Watts. Here is a summary of that study from its online abstract:

Hit songs, books, and movies are many times more successful than average, suggesting that 'the best' alternatives are qualitatively different from 'the rest'; yet experts routinely fail to predict which products will succeed. We investigated this paradox experimentally, by creating an artificial 'music market' in which 14,341 participants downloaded previously unknown songs either with or without knowledge of previous participants' choices. Increasing the strength of social influence increased both inequality and unpredictability of success. Success was also only partly determined by quality: the best songs rarely did poorly, and the worst rarely did well, but any other result was possible. (Salganik, Dodds, & Duncan, 2006)

A second category of barriers includes prisoners' dilemmas, social loafing, and tragedies of the commons (Davenport & Prusak, 2000). These dilemmas or barriers, which involve disincentives for collective performance, are less explored, and described in less depth in economics. They apply at both the interpersonal level (e.g., social loafing in teams) and at the distributed level (low voter participation in democracies). The knowledge and intelligence of the whole turns out to be less than the sum of the parts because ***some parts contribute but others slack off.*** This tells us that it would be important to consider carefully structured incentives to reward individual participation as well as collective intelligence. This naturally occurs in futures markets and betting, but further application and innovation on such ideas is possible.

While the first two barriers involve mechanisms that suppress or delete individual contribution, a third category involves failures to integrate contributions when they are made adequately. Surowiecki offers the traffic jam as a simple example. Information overload on the Internet is another. The knowledge and intelligence of the whole turns out to be less than the sum of the parts because ***the parts' contributions interfere with or cancel each other***. Solving this problem chiefly involves evolving structures and practices that coordinate individual and group contribution. At a distributed level, structures are highly visible, albeit incompletely studied. For example, there are congestion pricing for traffic systems, eBay's auction and interpersonal rating system, Amazon.com's collaborative filtering, Google's algorithm for search ranking, Wikipedia's review practices (as studied by Giles [2005, pp. 900-901]), and so forth. There is similar wealth on the interpersonal side, although many of the practices remain proprietary or tacitly in the hands of professional facilitators.

A final category of barriers to collective intelligence and innovation emerges from William Isaacs' work on dialogue. His theories build upon work by physicist David Bohm on "thought as a system," a new perspective in which all thinking and intelligence is understood to be collective. Within the system of collective thinking, Isaacs identifies four key pathologies that decrease collective intelligence. For each of pathology, he describes a principle that should be kept in mind and a dialogue practice for individuals to refine their own awareness and intelligence quality and contribute to fostering collective intelligence (Malone et al., 2006):

- **Abstraction/Fragmentation:** The tendency to hold oneself distant or separate from the world, for example, by abstracting or compartmentalizing it.
  - Siloing is a clear symptom of this kind of phenomenon: "That's an economics problem, not a psychology problem"; "That's a marketing problem, not a manufacturing problem"; "Not invented here."
  - Staying high on the so-called "ladder of inference" (Argyris, 1998), that is, arguing at the level of fragmented and reified ideas instead of about the flow of experience and data.
  - An example of this issue is: "This is a unique case" instead of "this is a symptom of how the whole thing is working."
  - Antidote: Listening (to data, to people, to the innovation emergence); that is, the principle of *holographic participation*, which is all things are whole, connected.
- **Idolatry of Memory:** The repetition of automatic answers, routines, stereotypes, and behavior patterns from memory.
  - "We solved that problem years ago"; "That's just the way we've always done things around here"; "We have a human resources department, therefore we're taking care of our people."
  - Antidote: Voicing what is actually new and emergent in one's understanding and experience; that is, the principle of *unfolding potential*, which is, the universe is always unfolding and producing the new.
- **Certainty:** The "knowledge" that one's view (often a manager or chief officer) is correct.
  - "That's impossible"; "There's no way that could be true."
  - Antidote: *Suspending* one's assumptions and prejudices for personal and collective reflection, that is, the principle of *proprioceptive awareness*, which is learning to see and feel how your assumptions are affecting your thinking and actions.
- **Violence:** The repression, disrespect, and destruction of alternative points of view in order to force acceptance or consensus of one's own understanding.
  - "No educated person could take that view"; "You're an idiot for believing that"; "That's all well and good, but…"
  - Antidote: Respecting diversity of opinion, style, and knowledge; that is, the principle of dif*ferentiation*, which is, diversity is natural and valuable, and collective intelligence means fostering differentiation and integration.

Enterprise 2.0 ground rules are related to these strong and weak factors of collaborative intelligence (Davenport & Prusak, 1997) and tackle each one of them to adequately emerge social capital.

SOLUTIONS AND RECOMMENDATIONS TO FOSTER COLLABORATION AND KNOWLEDGE EMERGENCY THROUGH ENTERPRISE 2.0 IDEAS

Enterprise 2.0 Key Technologies and Models for Improving Knowledge Work

As mentioned above, Web 2.0 and its application in enterprises can be seen as the computer industry's business revolution caused by the move to the Internet as a platform, and an attempt to understand the rules for success on that new platform. The key rule is to build applications that harness network effects to improve as they are used by more and more people.

The concept of Web-as-participation-platform captures many of these characteristics from the viewpoint of the new software as a service (SaaS) paradigm. Bart Decrem, founder and former CEO of Flock, calls Web 2.0 the "participatory Web" and regards Web-as-information-source as Web 1.0 (O'Reilly, 2004).

The following sections deal with the different existing Enterprise 2.0 technologies from two different viewpoints: first, a service-oriented perspective that is paving the way for a user-centered Web of services, recently termed as global SOA (Schroth & Christ, 2007); and second, a user-centered content driven perspective, comprising enterprise blogs, wikis, RSS, and other business knowledge channels.

Service-Oriented Enterprise 2.0 Technologies

The number of enterprises that are bringing their business systems to the Web to automate cross-organizational business transactions is constantly growing. Porter (2001) says that benefits of performing such transactions electronically include extending market reach, saving time, cutting costs, and responding to customer queries more agilely. Renowned scientists such as Malone (2001) cite the relentless march of improvements in the cost-performance ratio of information technology as the main driver of this development. SOAs have attracted, as McAfee (2005) notes, a lot of interest during the last few years as they are expected to play a key role as enablers of seamless application-to-application integration, both within company boundaries and on a global, cross-organizational scale, required to build this scenario.

From a technological viewpoint, Web services (Alonso, Casati, Kuno & Machiraju, 2004) have been massively adopted as the technical foundation for the realization of SOAs. Even so, Web Ssrvices-based SOAs mostly only exist within company boundaries at present (Roman et al., 2005, pp. 77-106). The global provision and consumption of services over the Internet is still at an early stage and has not yet taken on a significant role in realizing cross-organizational collaboration in an *Internet of services.*

Several reasons, such as high technical complexity, implementation and maintenance costs, inflexibility, and the lack of widely accepted standards for defining service choreographies as well as message semantics, have been repeatedly identified as key factors that have prevented the emergence of a global mesh of interoperable Web services, as Hinchcliffe says (2007). Further hurdles on the path to a "global SOA" include the lack of global-scale service discovery, as well as platforms allowing for intuitive human-guided service interaction and composition. Recently, the emergence of the Web 2.0 phenomenon is expected to act as a facilitator of such a global SOA (McAfee, 2007). Novel Web 2.0 technologies and design principles are now about to experience increasing acceptance as they allow for reusing, customizing, interconnecting, composing, and finally exposing Web-based content or functionality again as new resources. They are, therefore, considered not as a substitute for, but as an enrichment of, SOA concepts and technologies (Schroth & Christ, 2007; Schroth & Janner, 2007).

User-Centered Global Service Oriented Architectures (Global SOA)

A number of Enterprise 2.0 collaboration platforms are beginning to proliferate. By leveraging desirable Web 2.0 attributes, these platforms provide enterprises with an ecosystem of employees, partners, suppliers, and customers who collaborate to develop capabilities by collectively generating, sharing, and refining business knowledge. Nevertheless, enterprise collaboration should evolve towards a new paradigm in which knowledge workers are considered as coproducers not only of information, but also of software services and applications that promote specific competitive advantages and/or meet their immediate needs, without involving IT departments. The Web 2.0-based approach to a global SOA empowers users to coproduce and share instant applications and thus represents a major step forward to evolving the above ecosystem into one in which all the stakeholders will also be able to collaboratively develop capabilities and innovate operating procedures by remixing and integrating already available services through the emerging ideas of Enterprise 2.0 mash-ups ("Mashing the Web," 2005).

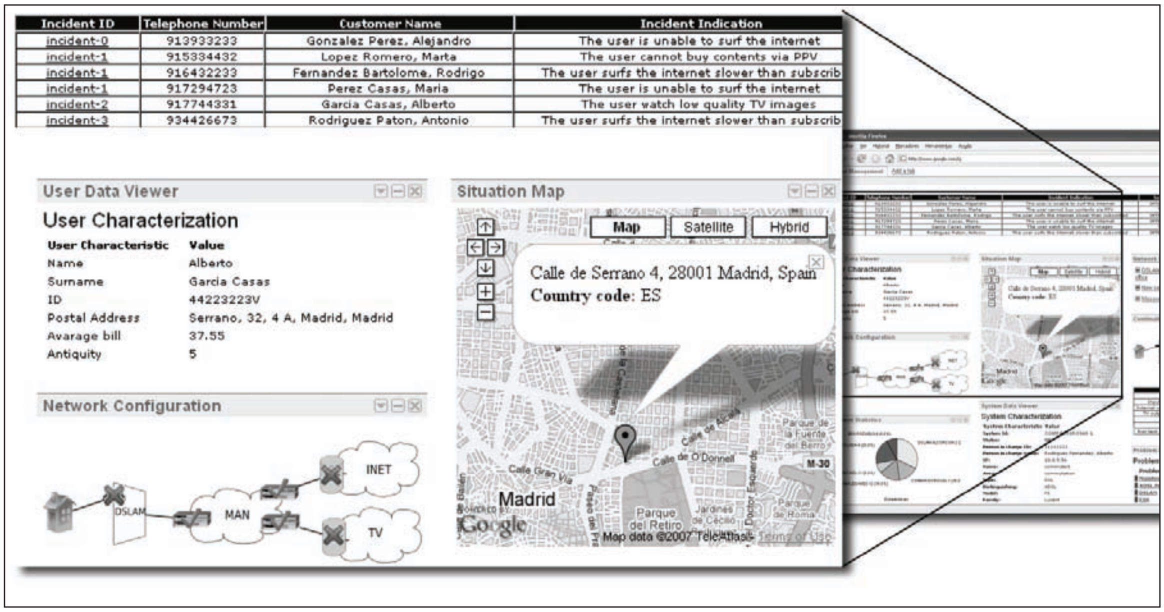
Enterprise 2.0 Mash-Ups

Content-driven mash-up-oriented programming (a.k.a. situational programming or instant programming) (Smith, 2006) is a new agile application development paradigm[1] in which knowledge workers, who do not have previous coding skills but do have extensive domain expertise, visually assemble and combine off-the-shelf gadgets (a.k.a. widgets), that is, discrete self-contained domain data-oriented components, with both development (service and data binding and interconnection) and runtime rendering capabilities. These gadgets represent the basic building blocks for knowledge workers to assemble new services (e.g., SOAP or REST-based lightweight Web services), data sources (e.g., Atom/RSS feeds) and other gadgets, and to render them as necessary to develop the application they need in a very short time. The kind of hybrid application that results from applying this new paradigm is often called enterprise mash-up (a.k.a. situational application or instant application) (Hof, 2005).

A simple example would be a mash-up connecting three gadgets: a list of tasks involving customers, an agenda of customers, and a Google map. By attaching the three to each other, the agenda gadget will display the customer's details and the Google map gadget will display the customer's address on a map as you scroll the task list. This would be useful, for example, for an employee responsible for the task of geographically locating the customer. It is the knowledge worker who can develop this "service," and do it on the fly with the help of mash-up enablers. This way a business person could build a "dashboard" to see how weather is affecting sales at retail outlets. By aggregating information from public Web sites, such as mapping and weather services, the business person could assemble a very useful, albeit simple, content-driven application. Companies are trying to capitalize on these technologies (Smith, 2006) with software and services for relatively short-lived, quick-to-build applications.

Figure 1 depicts a real scenario extracted from a Telefónica-based mash-up which connects four gadgets: a list of tasks involving customer requests, a customer agenda, a Google map, and a network status map. Figure 1 shows how Telefónica's operational support systems' (OSS) knowledge workers create a fully functional environment on their own by visually attaching these gadgets to each other and to the enterprise backend; the agenda gadget will display customer details and have a customer/task selection option, the network map will represent the selected customer's network status, and the Google map gadget will display the selected customer's address on a map when a given task is selected from the list. This

Figure 1. Creation of an EzWeb platform-based enterprise mash-up



enterprise mash-up environment is useful for a user responsible for the task of testing the status of all systems used by a customer. In the event of a problem in the customer's local telecommunications infrastructure, customer geographical location is a big help for the technician to prepare the visit to the customer's home.

As mentioned above, traditional Web services are provided as functionality described by arbitrary (mostly WSDL compliant) interfaces that define input and output messages, as well as the supported service functions. These interfaces are not human-readable and do not facilitate the interaction of users with the underlying services. In the mash-up context resources no longer target technical experts in the corporate context but now go for the huge number of individuals (Anderson, 2006), of Internet users who require intuitive visual means for retrieving resources on the Web and for capturing their respective functionality.

This way, the Web 2.0-based approach to a global SOA delivers a mash-up-enabled infrastructure to help businesses share and collaborate with the business ecosystem and partners instantly. In doing so, enterprise collaboration architectures introduce the mash-up-oriented lightweight programming model as a means for knowledge workers to collaborate in solving an immediate, specific business problem by blending externalities with private business content and services.

The way services are discovered, used, and managed by knowledge workers is fundamental in terms of both the ICT technology and the cultural aspects involved in implementing this enterprise collaboration paradigm shift (Salganik et al. 2006, pp. 854-856). In this respect, user-service interaction must embrace a number of principles to ensure the widest acceptance by knowledge workers. The most important that have been identified are:

1. *Knowledge workers must feel fully empowered* and able to serve themselves from available resources that provide them with access to the content and services they can use to set up their own personalized operating environment in a highly flexible and dynamic way.

2. *Active user participation must be enabled*. Knowledge workers must be able to contribute new and improved versions of resources, as well as share further knowledge about these resources, their use, and their interrelationships.
3. *Community-based collaborations need to be fostered*. The introduction of a *share, reuse, and assembly culture of collaboration* will boost and speed up this process thanks to the network effect.

To exploit this approach to the maximum, IT departments will need to embrace the SaaS model as an effective software-delivery mechanism. This will change the department's focus from deploying and supporting applications to managing the services that those applications provide. Knowledge workers will now extend and improve these services in a collaborative fashion to exploit their extensive domain expertise and their thorough business knowledge.

Enterprise Mash-ups as a Means to Drastically Improve Time-to-Market

Internet technologies continue enabling businesses to expand their ecosystems and partnerships. This expansion process means, on the one hand, that the information technology effort focuses on work items related to integration, usually requiring a minimum of 6 months per request. On the other hand, partnerships change, and some business collaborations last less than 12 months. Consequently, there is a whole bunch of applications not being written today because they are not affordable due to time-to-market constraints and/or because there is no justification for IT investment. Knowledge workers' needs are typically of short duration (ranging from one week to several months), thereby limiting justification for IT investment too. It would be sufficient if they were provided with informal, just-in-time access to domain content, and were able to create their own short-lived ad hoc application for each individual need, without the need for IT investment.

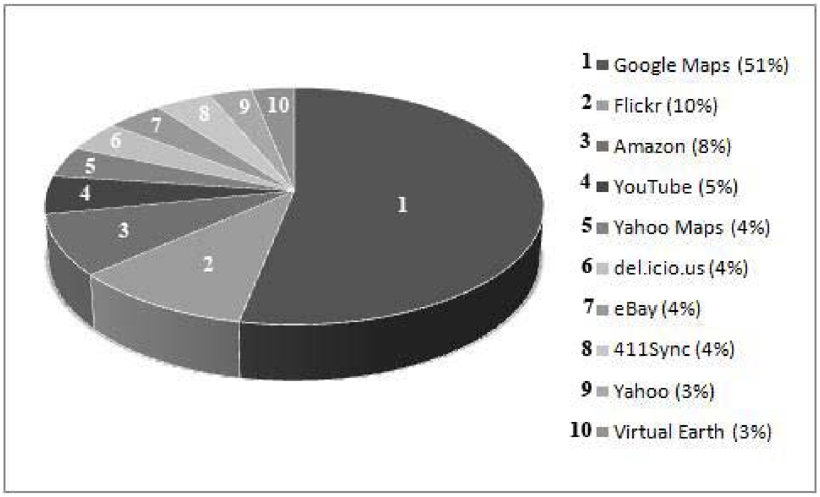
Enterprise 2.0 mash-ups, and the associated application development paradigm, clearly fulfill these needs. Mash-ups will help businesses share and collaborate with the business ecosystem and partners instantly. This in turn will help evolving enterprise collaboration towards a new paradigm in which knowledge workers (without previous programming skills, but with thorough business knowledge) are considered as coproducers not only of information, but also of software services and applications that promote specific competitive advantages and/or meet their immediate needs, without involving IT departments, and share the solution with the remainder of the organization. This will lead at last to an ecosystem of knowledge workers collaborating to develop capabilities and innovate operating procedures by remixing and integrating available services, exploiting the collaborative intelligence. Community-based collaborations could speed up this process thanks to the network effect caused by broader community support and participation (architecture by participation). Additionally, the introduction of a reuse and assembly culture will boost this process, allowing strong business value synergy and linkages.

Current Impact of Enterprise 2.0 Mash-Up Technology

The ProgrammableWeb.com Web site acts as a major aggregator of numerous (currently over 1,600) references of assorted mash-ups existing all over the world. It provides statistics about used resources, a classification of mash-ups by categories, as well as user statistics to evaluate mash-up popularity. Therefore, it serves as a central source of empirical data to work with. Several facts can be inferred from these data:

- First, large companies such as Google, Amazon, and Microsoft started to successfully provide Web-based resources that were

Figure 2. Chart of most popular mash-up resource providers (according to ProgrammableWeb.com)



leveraged by numerous users to create mash-ups. At least 836 Web-based applications have already integrated the "Google Maps," a resource that offers geographical data.

- Second, there is a shift away from the professional corporate context towards a private, end-user driven field of applications. As opposed to the focus of traditional Web services, the resources used for building mash-ups target the long tail of Internet users and deal with media management, shopping functionality, entertainment, or desktop applications. Heavy-weight enterprise applications for automating business transactions can only rarely be found in this context.
- Third, besides the SOAP protocol, which is also used in the context of traditional Web services, "light-weight" protocols, like REST (Fielding, 2000) and RSS, are widely leveraged to allow for fast and seamless mashing of different resources.

Summing up, the Web can increasingly be considered as a comprehensive and global development platform containing numerous easily usable and mashable resources that are provided by large firms, as well as by SME and even individual end-users. As argued above, the provision of resources that use lightweight protocols based on uniform interfaces, such as REST or RSS, and also the focus on end-user requirements rather than business-to-business relationships represent core success factors for this new global SOA consisting of numerous mash-ups. Figure 2 shows the current most popular providers of resources leveraged for the creation of mash-ups.

User-Contributed Content-Driven Enterprise 2.0 Technologies

Of the existing categories of Enterprise 2.0 technologies now available, technologies for content management (enterprise *wiki*/*blogs*), contract management, project management, enterprise mash-up platforms, messaging and e-mail, signaling (feeds), listing services, social network analysis and analytics, social search, media, collaborative categorization (a.k.a. *folksonomy*), online Web storage, supporting infrastructure, organizing, social networking, consumer or workgroup *wikis*, massive collaboration, and business process management are worth mentioning (Reding, 2006). There will be significant differences in companies' abilities to exploit these technologies due to the challenges they bring with them.

Because of the opportunities these technologies offer, these differences will matter a great deal. It is important to get an understanding of their real potential and drawbacks, as well as how to take advantage of them holistically.

Andrew McAfee (2006, pp. 21-28) first introduced the acronym SLATES to indicate the six components of Enterprise 2.0 technologies: **s**earch, **l**inks, **a**uthoring, **t**ags, **e**xtensions, and **s**ignals (SLATES). As technologists build Enterprise 2.0 technologies that incorporate the SLATES components, they seem to be following two intelligent ground rules following McAfee's vision of SLATES. First, they are making sure their offerings are easy to use. Second, Enterprise 2.0 technologists are trying hard not to impose any preconceived notions on users about how work should proceed or how output should be categorized or structured. Instead, they are building tools that enable these aspects of knowledge work to emerge. In the following sections we elaborate on the key Enterprise 2.0 technologies that will help a business to successfully exploit every SLATES component.

Enterprise Blogging

Blogging allows users to keep track of ideas and their authors, including concept redefinitions or business process information on a time line in a Web page, as if it were traditional Web content information. This information evolves like an approximating definition cycle based on dialogue and creative discussion, looking around for near ideas and issues to express enterprise knowledge.

There are several social networking functions in enterprise blogging that show the difference between simple blogging systems and systems used for triggering network effects across an organization. Some interesting insights about internal blogging in the *Enterprise Blogging in Practice* case study follow (Rand, 2004).

Michael Cot notes, *"At the department level, I wouldn't say that blogs have been a wide-reaching, raging success, primarily because people don't post to them as much as you'd hope. However, for the people who do post to and read the blogs, they've been very successful"* (Rand, 2004).

Among many other things, people usually post information at their company about their own stories and experiences, the status of tests, brainstorming-based ideas or issues, customer visit/phone call notes, comments that are only useful to bring attention to or track an employer with a project, requests for ideas or help, and even off-topic posts that can be useful for enforcing social networking relationships.

But the biggest problem is limited search capabilities. Without a Google-like quality search (i.e., near real time and full indexing, page-rank, quick search results, etc.) of the enterprise Intranet, it is very hard to find anything, let alone blog posts on relevant topics. People have been shown to get smarter a lot faster using social networking software (a lot of organizations of all sizes are using this Enterprise 2.0 application to trigger network effects and for innovation fostering). In addition to the enterprise blogging platform, it is interesting to stimulate several social networking system functions in Enterprise blogs (Wacka, 2005):

- Contact list (create informal groups or social networks) to maintain easy contact with community members
- Private messages
- Attach files (.doc, .pdf, .zip) and add tags and notes to them
- Advanced taxonomy with both structured and unstructured (free tagging) support
- Powerful ajax editor
- Tasks (to-do lists) for easy collaboration and project management
- Threaded comments for robust discussions
- Revision control
- Basic polls (advanced polling, surveying, and quizzes are available)
- Profiles
- Advanced search
- News aggregator (read RSS feeds and XML files)

- Syndication (generate RSS feeds and XML docs for content, profiles, tags, categories, etc.)

This way, it would be easier to make content and profiles precise. With structured and unstructured categories, it is easier to find people or content on the "long tail" of a curve (Kline & Burstein, 2005). In addition, it is possible to use tracking of individual profiles and posts to monitor user activity (i.e., to find out what others in the community are reading and writing).

Enterprise blogging should offer a useful way for connecting, creating, and collaborating on project management, help desks, finding and identifying experts (people aggregator), recruiting experts, talent, or ideas or innovation management, open innovation and its visibility, knowledge management, product development, and other off-topic tasks.

This platform would provide a comprehensive online ecosystem, a tightly integrated set of publishing, communication, and networking features that support and enable an online experience like never before. Users could engage, create, and share their content online (publicly or privately) in a multitude of ways to achieve greater performance. This technology will help to discover the purpose (i.e., connecting, creating, and collaborating) of an enterprise community in order to improve the company's business strategy. Knowing the purpose determines how the system is configured, how it flows, and how well it succeeds, and therefore the best strategy to carry out.

The following are some tips from Dion Hinchcliffe's (2007) "Nine Ideas for IT Managers Considering Enterprise 2.0," which are easily adaptable to enterprise blogging:

1. It is about ease-of-use, first and foremost.
2. Change requires motivation. Provide it.
3. Emergent does not mean a blank slate.
4. Discoverability is not an afterthought, it is the core.
5. It is okay to fear loss of control and misuse.
6. Dynamic, effective advocates are a key enabler.
7. The problems will be with the business culture, not the technology.
8. Triggering an Enterprise 2.0 ecosystem quickly is likely to be an early activity driver.
9. Allow the tools to access enterprise services

Enterprise Wiki: The 2.0 Approach of Content Management Systems (CMS)

A wiki can be defined as a piece of server software that allows users to freely create and edit Web page content using any Web browser. Wiki supports hyperlinks and has a simple text syntax for creating new pages and crosslinks between internal pages. Because it allows "everyday users to create and edit any website page, it is exciting in that it encourages democratic use of the Web and promotes content composition by non-technical users."

There are some fundamental Wiki design principles, but the most important principle that makes wiki different from any other Web site source is that it allows any enterprise readers to edit the page content as they see fit, if they feel the content is insufficient or poorly organized. If you come across any mistakes in the document as you read an information item, or you have more information that you would like to add to the item, just click on "Edit Text" and you, too, can change the content.

Because wiki is mainly designed to promote content composition by nontechnical users, the formatting rules for editing a wiki are fairly simple, and there are no complicated markup languages. Wiki content generally contains a lot of accurate information, because inaccurate information will be very quickly corrected by other readers.

The main idea of not having anyone to control the content in a centralized way, and/or of allowing *anyone* to edit and publish a document real-time is inconceivable for most people.

Therefore, a wiki can be the underlying technological support for creating a common sharing, emergence, and conceptualization of mash-up data, playing a similar role to ontologies in the Semantic Web vision. In fact, wiki software has demonstrated that ***it works well in a small community of like-minded people, like an enterprise community.*** Documents build up very fast as many people contribute small, manageable pieces. Some contribute contents and information, some contribute links, some correct grammar, while others fix the structure and formatting. Therefore, enterprise wiki allows the input of common descriptions and definitions of key business concepts, that is, everyone in the company can find a resource, bring experience to bear to evaluate and improve the resource. Wiki contents and data follow an iterative lifecycle, and their description undergoes constant improvements, refinements, and evolution.

There are several applications defined as enterprise wiki enablers, like Confluence, Social Text, or Twiki (*InfoWorld* proclaimed 2004 to be the "Year of the Enterprise Wiki"). Heightened interest comes in response to the increasing number of organizations like Google, Nokia, and Yahoo! who are turning to wikis as a way to improve internal efficiency.

Wikis and Content Management

Wikis fall conceptually under the broad concept of content management systems, and users could certainly use the existing CMS to create a *wiki-like* site. However, wikis have unique characteristics that differentiate them from traditional CMS (Choate, 2006).

Wikis emphasize ease of content creation. This simplicity has many sources: a wiki markup language that provides a short-hand way of linking documents and formatting text; the ability of users to edit and create pages independently and directly; a bottom-up approach to site navigation and structure; a very simple templating framework; and, finally, a conscious decision to eschew workflow or even simple approval steps.

Content Creation and Editing

Wiki software focuses on the empowerment of users to create and edit their own pages, but content management systems provide tools for creating and editing content, too. The difference is in their approach. When wikis first came out in 1995, there were not a lot of options for WYSIWYG editing from within a browser, so the wiki markup language (sometimes called "wikitext") provided a particularly valuable short-hand for formatting text that was much easier to learn than pure HTML (Heigl, Glaser, & Anja, 2006).

A good CMS will offer a WYSIWYG interface that makes writing content for the Web, like using a word processor. More wikis nowadays have WYSIWYG editing features, so the wiki markup language is a less interesting feature in terms of formatting, although it does provide the benefit of being supported by all browsers on all platforms, something that is not usually the case with rich-text editors. Many wikis support both wikitext and rich-text editors.

However, there is one area where wikitext still retains its power: linking resources and knowledge. Wiki software still provides a much easier way to link pages within the wiki to each other. Links are made based on the title of a page, so the author does not need to use, remember, or type long URLs in order to link one page to another.

Site Structure and Navigation

Contributors can create new knowledge, pages, contents, and can easily link one page and data to another (Venners, 2003); wikis offer a new unique approach to navigation and site structure.

Traditional information systems usually take a more formal approach to site structure and navigation through enterprise knowledge, with the site organized into a hierarchy by an information architect. User-created pages in a wiki mean that the hierarchy and structure of the site is created in an ad hoc way. Navigation tends to be simple, and the hierarchies are flat. For example, the Wikipedia online encyclopedia has hundreds of thousands of articles on a broad range of topics, but these topics are not arranged in any conceptual hierarchy. The entry for dogs serves as a good illustration. The URL for the article about dogs is: http://en.wikipedia.org/wiki/Dog

A pug is a kind of dog, and the URL for the pug entry is: http://en.wikipedia.org/wiki/Pug

Since a pug is a kind of dog, you might expect to see the following URL for pugs: http://en.wikipedia.org/wiki/Dog/Pug

But it is not there. Several wiki software solutions support more complex content categorization, but many are totally flat, like Wikipedia. Even if the software supports subconcepts, contributors are still allowed to create subpages in an ad hoc fashion and there is no systematic approach to the architecture of enterprise knowledge.

Content Repository and APIs

An experienced architect or administrator will ask of any content technology what the repository looks like. This is a good approach, because they are concerned about back-up, compatibility, performance, and a raft of similar issues.

Wikis have traditionally taken a very simple approach to information storage. Original wikis stored content in plain text files written with a wiki markup language. When a reader requested a page, the page was rendered. This was not speedy, but it worked. These days, wiki packages employ one of several different back-ends, with many housing their content in databases.

A transcendental consideration is whether the software supports automatic back-ups (commercial wiki applications often do). Another thing to think about is what this means in terms of integrating wiki content with content managed by other systems. For example, should an enterprise search system be able to index wiki content, and should the indexed content be raw wikitext, or rendered HTML pages?

This question leads on to the issue of wiki APIs, which, in fact, very few wikis have.

Templates

When a wikitext page is required, it is rendered as HTML in a two-part process. First, the wiki markup is converted to HTML, and links are created between pages. Then, this content is wrapped by a template that provides a consistent look to all the pages in the wiki.

Comparing wikis to a CMS, most wikis have template systems that are very simple, often only enabling one general template for the entire site. Wiki templates (and page rendering in general) are often not cached, so the page is rendered with each request. From an enterprise perspective, a lack of caching can obviously limit system scalability. On the other hand, there is no finicky caching mechanism to deal with.

Workflow

Wiki software completely changes the main idea of a workflow. Wikis are decentralized and typically lack the controlling mechanism of a workflow system with a formal approval process. Wikis workflow systems often lack sophisticated and disaggregated and approval processes is commonly considered a feature and not a fault of wikis. This is contrary to the basic idea of many CMS, which prioritize control over empowerment. Despite this decentralized approach, there is one important thing to remember: the main idea that anyone can edit content is only a general policy and not an inherent characteristic of wiki software.

Control vs. Flexibility

There is a traditional trade-off between control and flexibility in information systems software. Decision-making is centralized by some sort of editor that verifies and approves content prior to publishing in a traditional CMS. With a wiki, the writer writes then publishes without editorial oversight or approval. This direct channel to publication is what makes wikis so wonderful in scenarios that emphasize speed and flexibility.

An important issue to deal with is what possibilities there are if enterprises want to exercise at least some control. In the absence of workflow controls, content creation in a wiki is managed through change monitoring, automated spam prevention, and user access control. Let us look at each one in turn.

Change Monitoring

One simple defense mechanism is to monitor changes in the wiki and enable the rolling back to a previous version through versions control.

Recent changes can be monitored as follows:

- Most wikis have a "Recent Changes" page that lists all the pages that have been changed.
- E-mail notification of changes or support for RSS syndication.
- If more than one person has been tasked with monitoring changes, some wikis offer the capability to track whether a recently changed page has been checked yet, reducing the chances of the work being done twice.
- More sophisticated systems identify and differentiate "trivial" changes from more substantive ones.

In addition to the above ideas, it is important to consider that people make mistakes and sometimes deliberately do things badly. Therefore, the ability to roll back changes is a necessity. Features to look for include capabilities similar to what you would find in a CMS, like the ability to roll back changes to the previous version, to compare different versions side-by-side or the use of diffs between versions so that specific differences between them can be easily identified.

Spam Prevention

Another approach is to monitor the content of changes programmatically, that is, to manage spam prevention. This differs from user access control in the sense that it monitors wiki edits based on the content itself, or patterns of user behavior. Some systems can block access to IP addresses and URLs, or they can block the posting of individual changes based on restricting the use of certain words or phrases, using word lists or regular expressions, and blocking access based on excessive activity.

User Access Control

Enterprise wiki usually means that it has user access control. An increasing number of wiki projects offer sophisticated **more granular level** in user access control issues. Users and groups can be assigned rights to tasks such as reading, editing, writing to, and rolling back a resource to a previous version. There is a lot of variance among wiki packages in terms of how those rights are applied to the site. A less common but useful feature is the ability to restrict access to parts of resources. The most sophisticated enterprise wikis work with single sign-on security systems like Siteminder, or offer network and directory integration (e.g., LDAP and Active Directory) for user authentication and authorization.

Contrary to their reputation, wikis are CMS that can be managed efficiently. They simply take a different approach to content management by choosing to emphasize speed and flexibility rather than strict controls. In order to successfully implement a wiki software package you will need to look at workflow from a different perspective and be sure to select wiki software that provides the right level of content monitoring and access control for your organization.

Enterprise RSS

RSS is a family of Web feed formats used to publish frequently updated digital content, such as blogs, news feeds, or podcasts.

End-users that receive data from this technological channel use programs called feed "readers" or "aggregators", as follows. The user "subscribes" to a feed by supplying to their reader a link to the feed; the reader can then check the user's subscribed feeds to see if any of those feeds have new content since the last time it checked, and if so, retrieve that content and present it to the user.

The initials "RSS" are variously used to refer to the following standards:

- Really Simple Syndication (RSS 2.0)
- Rich Site Summary (RSS 0.91, RSS 1.0)
- RDF Site Summary (RSS 0.9 and 1.0)

RSS formats are specified in XML (a generic specification for data formats). RSS delivers its information as an XML file called an "RSS feed," "Webfeed," "RSS stream," or "RSS channel."

Essentially, Web 2.0 is fully centralized in its conception. Why are skype, del.icio.us, or Flickr Web sites instead of protocols (as foaf is)? The reuse of Web 2.0 data is limited only to the hostside and only with the help of feeds are data able to break out from centralized sites (Hammond, Hannay, & Lund, 2004, pp. 1082-9873). Therefore, feeds and RSS are the key to a new data-distributed model in the Web 2.0, where data are disaggregated on the Internet, and RSS allows data, information ,and remote events to be distributed to end-users through the Internet.

Content Tagging

Content tagging is a growing Internet trend that empowers users to add their own contextual tags to Web content, information, or resources. Typically, as Gruber (2005) affirms, this results in excellent content categorization in a way that is relevant to the needs of users.

Tags are, therefore, Web page and/or database descriptors (e.g., title, author, language, date, subject) that are assigned to knowledge (e.g., information, Web content, distributed resource, etc.). One of their main purposes is to help people find information. Tags can be assigned to document descriptions (e.g., card catalog cards in a library) or they can be assigned to the documents themselves or both.

Tags can be assigned by document authors, information professionals, editorial assistants, or even by computer programs. Artificial intelligence programs are a fast and easy (but not always the most accurate) method of tagging. Information professionals can produce highly accurate and effective tags that take into account all the nuances of language and subject matter, but there is a limited supply of people with these skills.

These tags are useful for creating an emerging user-centric categorization of content in a folksonomy (a user-generated taxonomy used to categorize and retrieve Web content). Folksonomic tagging is intended to make a body of information that is increasingly easy to search, discover, and navigate over time. A well-developed folksonomy is ideally accessible as a shared vocabulary that is both originated by, and familiar to, its primary users. Two widely cited examples of Web sites using folksonomic tagging are Flickr and del.icio.us.

Folksonomies are developed in Internet-mediated social environments. Therefore, knowledge workers can discover who has created a given tag for a concept, and see the other tags that this person created. In this way, folksonomy users often discover the tag sets of another user who tends to interpret and tag content in a way that makes sense to them. The result is often an immediate and rewarding gain in the user's capacity to find related content. Part of the appeal of folksonomy is its inherent subversiveness: when faced with the choice of the search tools that Web sites provide, folksonomies can be seen as a rejection of the search engine status quo in favor of tools that are created by the community.

Folksonomy creation and searching tools are not part of the underlying World Wide Web protocols. Basically, these folksonomies arise in Web-based communities where special provisions are made at site level for creating and using tags, as in del.icio.us. These communities are established to enable Web users to label and share user-generated content or to collaboratively label existing content. Since folksonomies are user-generated and therefore inexpensive to implement, advocates of folksonomy believe that it provides a useful low-cost alternative to more traditional, institutionally supported taxonomies or controlled vocabularies like enterprise IT solutions. An employee-generated folksonomy could therefore be seen as an "emergent enterprise taxonomy." Some folksonomy advocates believe that it is useful for facilitating workplace democracy and the distribution of management tasks among people actually doing the work.

As many authors note in blogs and articles, "workplace democracy is also seen as a utopian concept at odds with governing enterprise reality, the majority of which exist and thrive as hierarchically-structured corporations not especially aligned to democratically informed governance and decision-making." Also, the folksonomy may facilitate workflow, but it does not guarantee that the information worker will tag and, then, tag consistently, in an unbiased way, and without intentional malice directed at the enterprise.

Strategic Sensemaking

The increased importance of sensemaking will prove to be one of the central drivers for Enterprise 2.0 technologies adoption. The organizational theorist Karl Weick says that sensemaking is a central task in new organizations. Dan Russell at Creating Passionate Users provides a definition of sensemaking that will serve as a useful starting point: "Sensemaking is in many ways a search for the right organization or the right way to represent what you know about a topic. It's data collection, analysis, organization and performing the task" (Dervin, 1983).

Sensemaking can be a solution for constructing sensible accounts out of ambiguous, ambivalent, equivocal, and conflicting data in organizational settings for managers and leaders in the knowledge organization and management. In a world characterized by significant technology and strategic change, the problem of sensemaking becomes more acute.

One of the attractions of Enterprise 2.0 technologies is that they make business strategies more feasible and scalable. Most of the technologies depicted in this section take participation as far as what face-to-face methods can support. They make it possible to generate and organize more extensive raw materials and inputs to planning/sensemaking processes. Wikis with good version tracking and refactoring capabilities make it both safer and easier to generate and work through alternative representations/sensemakings. Realizing this sensemaking potential will require brokering some introductions and partnerships. Those adept in the techniques are likely not to be versed in the ways that the technologies reduce or eliminate some of the key barriers to successfully using the techniques. Those who understand the technologies may not be aware that the techniques exist, much less that they could benefit from technological improvement. One starting point is to investigate the sensemaking planning techniques and practices and map points where the technologies enable, simplify, or improve the techniques for those promoting Enterprise 2.0 technologies.

Social Networking

Nohria and Eccles (1992) give a common definition of social network as "a social structure made of nodes (which are often organizations or individuals) tied by one or more specific types of relationships, such as values, visions, idea, financial exchange, friends, kinship, dislike, trade, web links, etc."

Social network analysis approach consists of relations in terms of *nodes* and *ties.* Nodes are actors within the networks, and ties are the relations between the actors. There can be many kinds of ties between the nodes. Research in a number of academic fields has shown that social networks operate on many levels, from families up to the level of nations, and play a critical role in determining the way problems are solved, organizations are run, and the degree to which individuals succeed in achieving their goals.

A social network is a map of all of the relevant ties between the nodes being studied. The network can also be used to determine the social capital of individual actors. These concepts are often illustrated by means of a social network diagram, where nodes are the points and ties are the lines. In traditional social network communities, an initial set of founders sends out messages inviting members of their own personal networks to join the site. New members repeat the process, adding to the total number of members and links in the network. Sites then offer features, such as automatic address book updates, viewable profiles, the ability to form new links through "introduction services," and other forms of online social connections. Social networks can also be organized around business connections, as in the case of LinkedIn.

The combination of networking is a new point of view to social networking that combines both off-line elements (face-to-face events) and online elements. The newest social networks on the Internet are becoming more focused on niches such as travel, art, and so forth. Other social networking sites focus on local communities, sharing local business, and entertainment reviews, news, event calendars, and happenings.

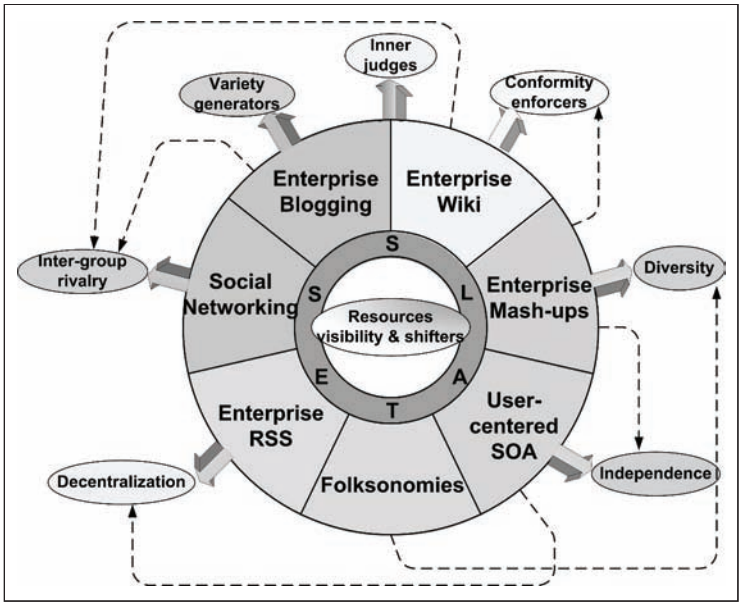
Traditional social networks on the Internet were public, and any user could participate. However, large enterprises and organizations also have access to private social networking applications often called enterprise social networking software. For example, Microsoft released an enterprise social networking application in the form of a free add-on for Microsoft Office SharePoint Server called Knowledge Network (currently in beta) in February 2007. Organizations install these applications on their own servers and enable employees to share their networks of contacts and relationships to outside people and companies.

Aligning the Business Strategy with Enterprise 2.0 Ground Rules, Diving Forces and Best Practices

Figure 3 shows a subset of the main Enterprise Web 2.0 technologies described throughout this chapter and what innovation promoting factors are fostered by each technology or ideology. Generally, each Enterprise 2.0 approach follows the SLATES principle described by McAfee. In addition, one of the most important innovation factors described in high-level knowledge working, that is, resources and content visibility through the Web and its simple modification and reedition in a shifting of resources is fostered by all technologies in the framework.

- Enterprise Wiki is used to input common descriptions and definitions of main business concepts. Everyone in the company can bring his or her own experience and knowledge to bear to find, evaluate, and improve a content resource. This iterative process enforces common data and knowledge conformity across the enterprise. Each knowledge worker must review the wiki concepts and will use an inner judge of the content to do his or her best to improve the concepts. In addition, this phenomenon shows group and department knowledge outputs, allowing the management of constructive intergroup rivalry in order to improve the descriptions of concepts or knowledge about business processes.
- Enterprise blogging is useful for keeping a track on ideas, concept redefinitions, or business process information on a time line. This visibility stimulates each knowledge worker to discuss this information, generating a variety of ideas in an approximate definition cycle. Obviously, this technology, as the wiki

Figure 3. Enterprise 2.0 technologies: Innovation fostering framework



approach, fosters the ideas of judgment and inter-group rivalry through the contribution of their own experiences. Therefore, wiki and blogging together can tackle the emerging knowledge, social capital, and collective enterprise intelligence created by the groups of knowledge workers.

- The social networking idea promotes the rivalry and communication between enterprise employers in a social net, using several communication channels to increase the dialogue in the enterprise and its departments.
- Enterprise mash-ups deal with the diversity of solutions to a definite problem within the enterprise. Software solutions based on a heterogeneous merge of separate components communicated and parameterized by knowledge workers fosters diversity and originality in the enterprise, eliminating barriers to the innovation bloom. This approach also helps to enforce conformity in the enterprise about systematic knowledge, that is, what mash-up solution is the best to afford a particular solution for a business process. Users follow a "do-it-yourself" ideology that encourages the independence of end-users from the service providers and legacy.
- Global (user-centered) SOA mainly fosters the idea of the independence of end users and end-user innovation from the content and services providers, breaking down traditional innovation barriers in SOA approaches. In addition, it motivates a decentralization of resources across the Web, fostering client-computation and disaggregated data models and composition against the service traditional front-end.
- Enterprise RSS can manage the decentralization of data in Web 2.0 and Enterprise 2.0. Feeds and RSS are the keys of a new data-distributed model where data are disaggregated in Internet, and RSS is useful for distributing data, information, and remote events to end users through the Internet.
- Finally, folksonomies are related to the visibility of resources, and resources discovery and recommendation issues. The use of a

relaxed taxonomy based on tagging by end users improves the diversity of knowledge related to these resources, and the collaborative intelligence present in enterprises can be better emerged.

Table 1 shows how the Enterprise 2.0 technologies are related to the factors that inhibit or slow down innovation.

These relationships are of three types: a technology can be appropriate for deleting a negative factor (shown as ✓), could cause this factor (shown as ✗), or must be applied very carefully because it can cause or delete a factor depending on its use.

Next these specific relationships between technologies and inhibit factors are depicted in more detail:

- An enterprise Wiki is useful for collaboratively editing contents through a Web platform. In this technology, a new content or concept would commonly be inserted to the wiki, imitating the description, structure, and form used in other previous concepts, that is, it leads to a negative conformity and imitation process without concern about enterprise integration. On the other hand, it is very simple to edit a concept or contribute with one's own experience and knowledge to a wiki. Therefore, this idea reduces social loafing and the slack-off caused by traditional complex content management systems. The wiki uses an iterative description of concepts, that is, a new edition overwrites the previous one and could cause coordination failures. Finally, a wiki content is usually anonymous and iteratively improved. For this reason, it is easy to eradicate the idolatry of memory in enterprise knowledge, constantly improving the contents and anyone in the enterprise to properly discuss the manager's ideas.
- Enterprise RSS is a communication channel to manage the decentralization of data in Enterprise 2.0. Therefore, some factors like conformity or slack-off depend on the content transmitted through the feeds. The RSS architecture, based on client aggregators and feed channels, decreases the coordination failures during contents accessing in the enterprise. In addition, this data communication channel, split into several different disperse channels merged in the client, is a correct way to deal with data and their abstraction and fragmentation. The negative part of RSS is that the data origin is known, and this could foster the feeling that data created by heads or managers and distributed by RSS are certain.

Table 1.

	Conformity & imitation	Social loafing & slack off	Coordination failures	Abstraction/ Fragmentat.	Idolatry of memory	Certainty
Enterprise Wiki	✗	✓	✗		✓	✓
Enterprise RSS	?	?	✓	✓		✗
Folksonomy	✓	✓			✓	✓
Enterprise Blogging	✗	✓	✓		✗	✗
Enterprise Mash-ups	✗	✓	✗	✗	✓	✗
Global SOA	✓	✗	✗	✓	✓	
Social Networking	?	?	✓			✗

- Folksonomies can create informal taxonomies based on tags (anonymous or not) in a very simple way using a Web platform. A new concept is very quick to tag. Therefore, this technique decreases the conformity in the conceptualization and the social loafing surrounding tagging contents and applying knowledge. Each knowledge worker will use his or her own experience, refining the tags used, even if these tags have been imposed by managers. For this reason, the knowledge is extrapolated and transformed, and therefore social capital emerges as obsolete enterprise conceptualizations are forgotten.
- Blogging keeps a track of ideas (and their authors), concept redefinitions, or business process information on a time line. This technology often causes a conformity feeling among the personnel, imitating structures, ideas, natural language descriptions, and schemas during a new track of knowledge in other enterprise blogs (in fact, wiki could cause a parallel effect). As with wiki editing, it is so simple, friendly, and quick to edit a new comment, or to refine an idea that social loafing is evidently decreased. However, each contribution is logged and stored in a blog, which has several consequences: coordination failures and overwrite issues are decreased but knowledge workers could be afraid of expressing their ideas in public or of arguing a traditional notion (provoking an idolatry of memory) or a leader's opinion.
- Enterprise mash-ups motivate software solutions based on a heterogeneous merge of separate components. A working mash-up often causes a conformity feeling in knowledge workers. Therefore, it is recommendable to force them to create their own solutions fitted to their own problems. In this same sense, nonprogrammer users should be offered a simple way to create mash-ups and reduce slack-offs. This approach may often cause coordination failures across departments, creating solutions to similar business problems. Likewise, this technique can foster a wrong abstraction level at solving problems, creating partial software solutions without considering the whole problem dimension. This approach can improve previous mash-up-oriented solutions in a very fast and simple way, reducing the traditional idolatry of previous software or enterprise solutions so harmful in business strategies. Finally, it is very dangerous to publish strict mash-up compositions as unique software solutions created by managers or specialized departments, because this could provoke an innovation barrier to new ideas or improvements to these solutions.
- User-centered SOA mainly fosters the idea of independence of end users and end-user innovation from the content and services providers, focusing on reducing conformity and imitation issues caused by the traditional rigid SOA approach. This idea must be applied carefully because it could foster a slack-off in nonprogrammer users that have a poor perspective of Web services, or problems coordinating efforts in pragmatic developments. This issue can be easily improved using this approach and a mash-up orientation together. A strong point of this technique is the correct abstraction/fragmentation view of enterprise complex problems, coordinating or orchestrating user-centered services to tackle with whole problems through interface charts and storyboards linked to concrete workflows. One of the most important ideas is the high parameterization in enterprise services. This makes it easy to forget traditional memorized solutions, improving them in an adaptable way.
- The social networking is a general philosophy focused on improving effort coordination across a social group, fostering collective intelligence emergence and exploitation,

reducing the possible coordination failures, and increasing outsourcing visibility. Like RSS, social networking can be considered a family of communication channels. Therefore, aspects like imitation, conformity, or social loafing depend directly on the content and the management of social information and knowledge. In this philosophy, it is harmful to introduce managers and manager-generated knowledge that can cause the social group concern and delimit its innovation process and evolution.

Application to Other Key Business Web Strategy Areas

This section looks at each of the implications of the explained Enterprise 2.0 vision for communication and information sharing, knowledge management business intelligence and business process management, and its application to key business web strategy evolution.

High-Performance Collaboration and Community-Building

More and more often organizations tend to behave like dynamically reconfigurable networked structures that carry out their tasks through collaboration and teamwork. Effective teamwork is an essential part of any nontrivial engineering process, and collaborative capabilities are an essential support for these teams.

Traditionally, collaboration has been a means for organizations to do their work. As illustrated throughout this chapter, however, the context in which they do this work is changing, especially in regards where the work is done, how the work is organized, and who does the work, and with this the characteristics of collaboration. Work teams face sizeable collaborative challenges, for which they have need of tools that they can use to communicate and coordinate their work efficiently. These challenges have been tackled traditionally by profuse research in *computer supported collaborative work* (CSCW). CSCW has a great deal of drawbacks can be dealt with under the Web 2.0 vision. Web 2.0 has taken a step forward in this respect with the emergence of *social networking* and *communities*, where the emphasis is on *open source communities.*

Open source communities are one of the most successful—and least well understood—examples of high-performance collaboration and community-building on the Internet today. *Open source communities* began as loosely organized, ad hoc communities of contributors from all over the world who shared an interest in meeting a common need. However, the organization of these communities has proven to be very flexible and capable of carrying out all kinds of developments, ranging from minor projects to huge developments. Businesses following the *Enterprise 2.0* vision can benefit enormously by learning what open source communities are and how they work. It is important to remember how the use of the *Enterprise 2.0*-based IT infrastructure will transform today's Intranets into virtual spaces where all project stakeholders, possibly distributed in time and space, can negotiate, brainstorm, discuss, share knowledge and resources, and, generally, work together to carry out some task. The vision presented in previous sections will help to definitively change Tom Allen's well-known "30-meter" rule, stating that two scientists or engineers whose desks are more than 30 meters apart have a communication frequency of almost zero.

Collaborative Knowledge Emergence and Management

The concept of knowledge management introduced previously has been an elusive chimera to corporations since the mid-1990s. Ever since employees came to be seen as knowledge workers, companies have been searching for ways to capture and disseminate the stuff inside their heads. Knowledge management systems have traditionally tried to do this by both relying on distributed production and providing high commonality. This way, they have sought to elicit tacit knowledge, best practices, and relevant experience from people throughout a company and put this information in a widely available database.

Nevertheless, a corporation's knowledge is scattered across a multitude of communication media channels, including e-mail threads of conversation, instant messaging talks, and communication media platforms or Intranets, corporate Web sites, and information portals. Production in the first group is distributed (that is, knowledge workers can create and distribute digital information free of charge), and many of them leave communication traces (e.g., instant messaging talks or e-mail threads of conversation). However, the degree of commonality of this information is low (e.g., only the participants in an e-mail exchange have access to the knowledge held in the thread of conversation). In the second group, commonality is high, but production is centralized and visits to platforms leave no traces. Both the "low commonality" factor in current channels and the "centralized production" and "lack of traces" factors in current platforms imply that most knowledge work practices and output are invisible to most people in most companies. For this reason, it is very important to understand the presented vision and especially how *Enterprise 2.0*-based IT introduces new channels and platforms that enable distributed production, communication tracing, and high commonality of information and services simultaneously to improve user productivity in the way explained in the framework of Enterprises 2.0 technologies and their application to innovation, knowledge emergence, and content visibility.

On the other hand, current knowledge-work-specific technologies, like highly structured knowledge management systems using complex taxonomies and/or ontologies are not doing a good job at capturing, sharing, and applying their knowledge, which is typically highly unstructured and textual. In this respect, a recent study (Morris, 2005) has shown that only 44% of respondents agreed that it was easy to find what they were looking for on their Intranet. The channels and platforms in traditional use are not much good at providing answers to questions like who is working on a similar problem right now. Or what is the right way to approach this analysis? The presented practical Enterprise 2.0-based collaborative (and social) approach can catalog and search knowledge so that employees can easily leverage it throughout the firm. Briefly, the application of the broad spectrum of Enterprise 2.0 of technologies to business Web strategy should be considered.

Finally, most current platforms, such as knowledge management systems, information portals, Intranets, business process management (BPM), business activity monitoring (BAM), and workflow applications are highly structured from the start, and users have little opportunity to influence their structure or to customize their functionality and their interfaces. Emerging platforms, like wiki, blogging, or folksonomies (explained throughout this chapter), for generating, sharing, and refining information under the Enterprise 2.0 vision umbrella focus not on capturing actual knowledge, but rather on knowledge workers' practices and output.

In conclusion, the Enterprise 2.0 vision is significant in this respect because it can potentially knit together an enterprise and facilitate knowledge work in ways that were out of the question before. Putting it simply, Enterprise 2.0 technologies have the potential to make the knowledge management infrastructure of a corporation what the Internet already is, that is, an online platform with a constantly changing, searchable structure built by distributed, autonomous, and largely self-interested peers. Technologies like blogs, wikis, and labeling systems capable of emerging folksonomies make a decisive contribution to the elicitation of knowledge, best practices, and relevant experience that is scattered across the corporation and make this information trustworthy, searchable, and accessible to people throughout a company, at the same time as creating a cooperative and helpful culture capable of boosting knowledge production and guaranteeing convergence and quality through highly egalitarian collaboration.

Social Network Analysis and Business Intelligence

Howard Dresner, a Research Fellow at Gartner Group, popularized the term business intelligence as an umbrella term to describe a set of concepts and methods to improve business decision making using fact-based support systems. This discipline aims to describe how end users could access and analyze information stored on their company systems in order to provide a better understanding of the business and its customer. To do this, they used a broad category of applications and technologies for gathering, providing access to, and analyzing data from the earlier business activities for the purpose of helping enterprise users make better business decisions.

Much of the early research in this field took place before the widespread use of the Internet and even corporate e-mail. We now need to understand the implications of Enterprise 2.0 technologies in the process of sharing business intelligence (Barquin, 2006). Web 2.0 can be seen (as it has been explained previously) as the ability to communicate and share knowledge with other participants that have similar interests, resulting in a key means of producing, communicating, and sharing business intelligence. In this line, this chapter focused on describing how Enterprise 2.0 can be used to spread, publish, and manage data from previous business activities through a new breadth of collaborative social knowledge networking tools like blogs, wikis, messaging, e-mail channels, and so forth, improving the added value outlined in enterprise business intelligence.

In previous sections, we showed how the Enterprise 2.0 vision introduces new channels and platforms that enable distributed production, communication tracing, and high commonality of information and services simultaneously. Apart from making information and knowledge work practices and outputs far more visible to the entire company, another key advantage pointed out during the chapter is that it can be used to carry out *social network analysis* of *business intelligence*. This is primarily concerned with the paths that information exchanges take between individuals and the fact that some individuals act as key nodes in the network and become critical factors in the successful communication and sharing of business intelligence.

Social network analysis allows us to identify the pathways that *business intelligence* must travel if it is to be used, integrated, enriched, and applied by individuals within enterprises. It will help to identify which the critical nodes are. Star nodes usually represent key points in the routing at which important intelligence can be either effectively tunneled to other network members that share the same interests or are working on the same type of problems or tragically choked, as we have seen up to now. Star nodes have been identified in almost any work environment grouping knowledge workers. Social network analysis is a powerful tool for improving a company's intelligence capabilities. Therefore, business intelligence analysis should be carried out not only on the content of communications, but preferentially on the structure of the social network, its topology, communication patterns, and links to identify the key nodes.

Business Process Management Revisited

Finally, BPM represents another key area in which Enterprise 2.0 vision will help to make great improvements. It is generally accepted that the current orchestration, BPM and workflow systems and technologies have failed to achieve a seamless automation and integration of business processes. Although there are currently a lot of approaches and standards in this respect, none of them seems to offer a sound solution to real enterprise needs. Most of the problem lies in how they handle the knowledge acquisition process for automating business operations. As we have seen, *Enterprise 2.0* provides a new approach to this problem that integrates flexibility, human interaction, and modeling facilities (different from the flowchart-based or the algorithmic approaches, and now based on networks of resources and interrelationships). Following the exposition carried out it is simple to find out the need to apply the Enterprise 2.0 approach to improve and evolve this kind of systems.

CONCLUSION AND FUTURE TRENDS

In this chapter we have explored emerging Internet technologies, highlighting their potential for supporting business Web strategy as companies' reliance on new Web-based technologies to capitalize on new business opportunities increase quickly. Specifically, Enterprise 2.0 has been presented as a key enabler for businesses to expand their ecosystems and partnerships, as well as acting as a catalyst for improving innovation processes and knowledge work in general. On the one hand, we have elaborated on the concept of Enterprise 2.0 mash-up as the main technological enabler of a global, user-centered SOA that spans company boundaries. On the other hand, we have discussed the most relevant Enterprise 2.0 models and tools suitable for fostering emergent collaboration and cocreation, thus enabling firms to leverage desirable attributes, including harnessing collective intelligence and organization by participation.

The key idea behind the Enterprise 2.0-based approach to a business Web strategy, and the lesson many businesses must learn, is that next generation IT systems must be conceived to acquire the knowledge they operate on directly from who really has it, that is, the employees (seen as knowledge workers) and from the operation and communication processes employees enter into. The knowledge of a business has less to do with the IT infrastructure than with the employees themselves. The IT infrastructure must be capable of extracting and managing that knowledge in order to evolve and adapt it to the business processes, and finally to the business strategy. Any other means to model and exploit the business knowledge will never be flexible enough. If user knowledge changes (and it does change), both the IT infrastructure and the business strategy must seamlessly adapt to such changes.

Future work would concentrate on evolving practical Enterprise 2.0 frameworks, as open source packages of technologies, tools, and platforms that build on all the key technical and theoretical enablers described above and on the proposed model of collaboration and enterprise knowledge emergency. We expect this framework and its pragmatic application could have great research opportunities within the domain of the topic.

REFERENCES

Alonso, G., Casati, F., Kuno, H., & Machiraju, V. (2004). *Web services concepts, architectures and applications.* Springer.

Anderson, C. (2006). *The long tail: Why the future of business is selling less of more*. Hyperion Books.

Argyris, C. (1998, May-June). Empowerment: The emperor's new clothes. *Harvard Business Review, 76*, 98-105.

Barquin, R. C. (2006, July 18). *Social network analysis and business intelligence*. Retrieved January 8, 2006, from http://www.b-eye-network.com/view/3088

Bloom, H. (2000). *Global brain: The evolution of mass mind from the big bang to the 21st century* (pp. 42-44). Wiley.

Brown, J. S., & Duguid, P. (2000). *The social life of information*. Boston: Harvard Business School Press.

Choate, M. (2006, April 28). What makes an enterprise wiki? *cmswatch*. Retrieved at April 15, 2007, from http://www.cmswatch.com/Feature/145-Wikis-in-the-enterprise

Coveyduc, J., Huang, C., Ostdiek, L., & Reif, J. (2006, October). IBM innovation factory. An integrated solution for accelerating innovation. *IBM*. Retrieved July 3, 2008, from http://ibm.com/websphere/developer/zones/hipods

Davenport, T. H. (2005). *Thinking for a living: How to get better performance and results from knowledge workers.* Boston: Harvard Business School Press.

Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning.* Boston: Harvard Business School Press

Davenport, T. H., & Prusak, L. (1997). *Information ecology*. Boston: Harvard Business School Press.

Davenport, T. H., & Prusak, L. (2000). *Working knowledge: How organizations manage what they know.* Boston: Harvard Business School Press.

Dervin, B. (1983, May). *An overview of sense-making research: Concepts, methods, and results to date*. Paper presented at the annual meeting of the International Communication Association, Dallas.

Drive, E. et al. (2004, August 2). *Road map to an enterprise collaboration strategy.* Forrester Research.

Fallows, D. (2005, January). *Pew Internet & American Life Project. Search engine users.* Retrieved January 1, 2006 from http://www.pewinternet.org/PPF/r/146/report_ display.asp

Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures*. Unpublished doctoral thesis, University of California, Irvine.

Giles, J. (2005, December). Internet encyclopaedias go head to head. *Nature, 438*(7070), 900-901.

Gruber, T. (2005). *Ontology of folksonomy.* Paper presented at the First on-Line Conference on Metadata and Semantics Research (MTSR'05).

Hammond, T., Hannay, T., & Lund, B. (2004). The role of RSS in science publishing. *D-Lib Magazine*, *10*, 1082-9873.

Heigl, R., Glaser, M., & Anja, E. (2006). *Wiki: Web collaboration*. Springer. ISBN 3540229396.

Hinchcliffe, D. (2007). i-Technology viewpoint: Is Web 2.0 the global SOA? *SOA Web Services Journal*. Retrieved February 23, 2007 from http://webservices.sys-con.com/read/164532.htm

Hof, R. D. (2005). Mix, match, and mutate. *Business Week Magazine*. Retrieved February 23, 2007 from http://www.businessweek.com/@@76IH*ocQ34AvyQMA/magazine/content/05_30/b3944108_mz063.htm

Högg, R., Meckel, M., Stanoevska-Slabeva, K., & Martignoni R. (2006). Overview of business models for Web 2.0 communities. In *Proceedings of GeNeMe 2006* (pp. 23-37). Dresden.

Intel Suite Two. *The Enterprise 2.0 Suite, powered by Intel*. Retrieved July 3, 2008, from http://www.suitetwo.com.

Kline, D., & Burstein, D. (2005). *Blog!: How the newest media revolution is changing politics, business, and culture*. Squibnocket Partners, L.L.C. ISBN 1-59315-141-1.

Laso, I. (2006, October 18). *A Web 2.0 research perspective. DG information society and media*. European Commission Representative at the 6th Internet Next Generation Workshop 2006, Madrid, Spain.

Lin, N., Burt, R. S., & Cook, K. (Eds.). (2001). *Social capital: Theory and research*. New York: Aldine de Gruyter.

Malone, T. (2001). The future of e-business. *Sloan Management Review, 43*(1), 104.

Malone, T. W., Jay, J., Legay, L., & Kosorukoff, A. (2006). *Handbook of collective intelligence*. Retrieved at January 15, 2007, from http://www.eu.socialtext.net/mit-cci-hci/index.cgi?handbook_of_collective_intelligence

Mashing the Web. (2005). *The Economist*, *376*(8444), 4.

McAfee, A. P. (2005). Will Web services really transform collaboration. *MIT Sloan Management Review, 46*(2).

McAfee, A. P. (2006). Enterprise 2.0: The dawn of emergent collaboration. *MIT Sloan Management Review, 47*(3), 21-28.

McAfee, A. P. (2006). *The impact of information technology (IT) on businesses and their leaders.* Retrieved January 1, 2006, from http://blog.hbs.edu/faculty/amcafee

McAfee, A. P. (2007). Enterprise 2.0 vs. SOA. *Harvard Business School faculty blog.* Retrieved June 8, 2007, from http://blog.hbs.edu/faculty/amcafee/index.php/faculty_amcafee_v3/enterprise_20_vs_soa/

Morris, M., Pohlmann, T., & Oliver, Y. G. (2005, April 8). *How do users feel about technology? Business technographics United States.* Forrester Research.

Musser, J., & O'Reilly, T. (2006, November). *Web 2.0 principles and best practices.* O'Reilly Radar.

Nohria, N., & Eccles, R. (1992). *Networks in organizations* (2nd ed.). Boston: Harvard Business Press.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press.

O'Reilly, D. (2004, 7 May). Web-User satisfaction on the upswing. *PC World*. Retrieved January 1, 2006, from http://www.pcworld.com/news/article/0,aid,116060,00.asp

O'Reilly, T. (2007) *What is Web 2.0*. Retrieved June 9, 2007, from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Porter, M. (2001). Strategy and the Internet. *Harvard Business Review.*

Rand, M. (2004, December 13). Best of the Web: Extreme blogging. *Forbes*. Retrieved January 1, 2006, from http://www.forbes.com/best/2004/1213/bow001.html

Reding, V. (2006, December 3). *The disruptive force of Web 2.0: How the new generation will define the future* (European Commission press release). Hong Kong: Youth Forum, ITU Telecom World.

Roman, D. et al. (2005). Web service modeling ontology. *Applied Ontology*, *1*(1), 77-106.

Salganik, M. J., Dodds, P. S., & Duncan, J. (2006, February). *Experimental study of inequality and unpredictability in an artificial cultural market*. Watts, Science 10, 311(5762), 854-856.

Schroth, C., & Christ, O. (2007). Brave new Web: Emerging design principles and technologies as enablers of a global SOA. In *Proceedings of the 2007 IEEE International Conference on Services Computing (SCC 2007),* Salt Lake City.

Schroth, C., & Janner, T. (2007). Web 2.0 and SOA: Converging concepts enabling the Internet of services. *IT Professional, 9*(3), 36-41.

Smith, D. M. (2006, May 16). *Web 2.0 offers many opportunities for growth, but few enterprises will immediately adopt all aspects necessary for significant business impact*. Paper presented at the Gartner Symposium/ITxpo, San Francisco.

Smith, R. (2006, June 14-16). *Enterprise mashups: An industry case study*. Paper presented at the New York PHP Conference & Expo 2006, The New Yorker Hotel, Manhattan.

Surowiecki, J. (2005). *The wisdom of crowds* (pp. xviii – xix, Chapters 2-4). Anchor Books.

Venners, B. (2003, October 20). Exploring with wiki: A conversation with Ward Cunningham, part I. *Artima developer*. Retrieved January 1, 2006, from http://www.artima.com/intv/wiki.html

Wacka, F. (2005). Enterprise blogging in practice. *webpronews*. Retrieved at January 1, 2007, from http://www.webpronews.com/topnews/2005/01/24/enterprise-blogging-in-practice

Weill, P., & Ross, J. W. (2004, May). *IT governance: How top performers manage IT decision rights for superior results*. Boston: Harvard Business School Press.

ENDNOTE

[1] Note that we are not talking here about things like agile development or eXtreme programming, because the target audience is the knowledge workers not a development team.

This work was previously published in Business Web Strategy: Design, Alignment, and Application, edited by L. Al-Hakim; M. Memmola, pp. 61-93, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 6.2
Web Engineering in Small Jordanian Web Development Firms:
An XP Based Process Model

Haroon Altarawneh
Albalqa' Applied University, Jordan

Asim El-Shiekh
The Arab Academy for Banking and Financial Sciences, Jordan

ABSTRACT

Small firms do not have the managerial experience, the financial resources and the methodological know-how to manage web-based applications projects the way large firms do. Many small firms are unaware of existing software process assessment models and standards. There's often the assumption that assessments conformant to these models and standards can be expensive and time consuming, and therefore difficult to perform in small companies. This chapter proposes a theoretical model for small Web project development and its special features in the context small Web firms, which are capable of being "tailor able" to the particular stage of organizational development of small Web firms . The process model derived form Web engineering best practices, real case studies from Jordanian Web firms and agile development methodologies (extreme programming) . This chapter also contains results from tow surveys: a questionnaire to Web developers and interview with Web mangers in Jordan.The results reflect the Web industry situation in small Jordanian firms, and the major problems they face. Most of small Web projects in Jordan run over time and budget, due to the ad hoc development and the weakness of Web project management. The results showed that there is a weakness in applying Web engineering practices in small Jordanian Web development firms.

INTRODUCTION

"Web Engineering is the application of systematic, disciplined and quantifiable approaches to development, operation, and maintenance of Web-based applications"(Deshpande Y and et al 2002). It is a response to the early, chaotic development of Web sites and applications as well as recognition of a divide between Web developers and conventional software developers (Murugesan, S et al 1999, Pressman 1998). Viewed broadly, Web Engineering is both a conscious and pro-active approach and a growing collection of theoretical and empirical research. Web engineering is the process used to create high-quality Web-based systems and applications that deliver a complex array of content and functionality to a broad population of end-users (bouchaib bahli and dany di tullio 2003). Web Engineering is concerned with the establishment and use of sound scientific, engineering and management principles and disciplined and systematic approaches to the successful development, deployment and maintenance of high quality Web-based systems and applications (Web Engineering Home Page 2003).

Web-based applications are becoming so popular in our daily life in the sense that it would not go a single day without we use them. These applications range from simple to sophisticated ones, where millions of dollars in revenue are generated. Developing, testing and quality assuring these applications become a challenging task (Abdesselam Redouane 2002). Although the development of Web-based applications made many improvements, there is still a lack of an established software engineering methodology for constructing Web-based applications. Consequently, much of the development is carried out without a true understanding of analysis and design issue.

The development of Web applications (E-commerce systems, Web portals, etc.) is subject to different conditions than that of conventional software systems (Said Hadjerrouit 2001). Such idiosyncrasies include: usability, rapid development lifecycle and short time to market. Web based systems and applications deliver a complex array of content and functionality to a broad population of end users. They require new approaches to design and development but present the same issues and challenges as traditional information systems. Therefore, the same software engineering techniques are still necessary but the process should take these differences into account.

Web-based applications differ from other applications from both the product and process point of view. As products, they differ from traditional systems in the following ways:

1. Web based applications are distributed and component based.
2. High reliability
3. High Usability
4. Security

Web applications also differ from traditional applications from the process point of view: there are more Technologies (HTML, XML, network protocols, multimedia, and Java and script languages) and thus, many Roles (authors, developers, graphic designers, legal issues etc.) that have to be managed. In addition, the shorter time to market, shorter product life cycles and continuous maintenance are much more pronounced in the case of Web applications as compared to traditional ones(D. Rodriguez et al 2002).

WEB PROJECTS DEVELOPMENT

The history of Web development is relatively short. Initially, many Web applications were small and simple with little thought given to planning or design before constructing the site, and few have been tested properly. Today, many applications are large-scale and involve sophisticated interaction with visitors and databases; such sites are often regarded as mission critical. In parallel with this evolution, a need for Web engineering has become apparent. Yet, within education, the plethora of Web courses primarily address the implementation of Web sites with very little about the analysis and design of Web applications. We believe that an early consideration of a Web engineering process suited for inexperienced users is important.

(Pressman 2003) says Web Engineering is not a perfect "Clone" of software engineering but it borrows many of software engineering's fundamental concepts and principles, emphasizing the same technical and management activities. The brief history of systems development methodologies identifies and explores eras of development and speculates on their future. Today's "post methodology" era involves methodologies that can be viewed by developers as outdated and inappropriate for rapid development, Web applications, and other current requirements. Perhaps we are in danger of returning to the bad old days of the pre methodology era and its lack of control, standards, and training (David E. Avison and Guy Fitzgerald 2002).

(Yogesh Deshpande and Martin Gaedke 2005) mentioned that "There are very few standard methods for the Web developers to use. Hence, there is a strong need to understand and undertake Web Engineering". Ad-hoc development of WBA has brought disasters to many organizations. A survey on Web based project development by the (Cutter Consortium 2000) highlighted problems for Web-based projects:

- Delivered systems didn't meet business needs 84 percent of the time.
- Schedule delays plagued the projects 79 percent of the time.
- Projects exceeded the budget 63 percent of the time.
- Delivered systems didn't have the required functionality 53 percent of the time.
- Deliverables were of poor quality 52 percent of time.

SOFTWARE PROCESS IMPROVEMENT

Software process improvement (SPI) assessments are considered by many small software development firms to be too expensive (Aileen 2004). Software process improvement (SPI) is defined as having the potential to improve competitiveness by increasing productivity; reducing costs, defect and rework; and improving time to market and customer satisfaction (M. E. Fayad et al 2000). Small software development firms recognize that software process assessments play a valuable role in improving a firm's processes and products, but most feel that SPI costs too much and takes up resources needed to deliver products. SPI is very important now days, where in the last 15 years, interest in SPI has increased as evidenced by the growing number of journal articles which include the phrase 'process improvement' in their title or abstract (L. Pringle 2001) .

Most of the empirical studies on SPI relate to large well-resourced organizations. It has been noted that very little is known about the experience of small software development firms in regard to SPI. The growth of the software industry has produced many small companies that do not do contract software, but rather compete in other areas. This gives rise to at least four significant development issues that have not been adequately addressed in software engineering literature: company size, development mode, development size, and development speed. . Definitions of "small" businesses vary by industry and by government agency from 100 to 500 employees or more. These bounds are somewhat broad for our purposes. Based on census data, we define companies of 50 or fewer employees as small (Mohamed E. Fayad et al 2000).

First step toward process improvement is identifying the strengths and weaknesses of an organization's software processes to determine effective improvement actions. An assessment can help an organization examine its processes against a reference model to determine the processes' capability or the organization's maturity, to meet quality, cost, and schedule goals (Christiane Gresse et al 2006).

A study by (Sebastian Stein 2006) showed that software process improvements are required to increase the productivity of software companies. Generally, it is the aim to increase the quality of the produced software and to keep budget and time. Quality models for software process improvements were developed in context of large organizations and multi-national companies.

AGILE DEVELOPMENT METHODS AND WEB ENGINEERING PRACTICES

The field of software development is not shy of introducing new methodologies. Indeed, in the last 25 years, a large number of different approaches to software development have been introduced, of which only few have survived to be used today. The term agile has recently been used to categories a number of lightweight approaches to building software. These include: Extreme Programming (XP), Adaptive Software Development and Dynamic Systems Development Methodology (DSDM). Seventeen advocates and methodologists of the aforementioned and other agile processes convened in February 2001. The result of this meeting was the formation of the Agile Alliance (Beck K. et al 2001)and the production of The Manifesto for Agile Software Development (Fowler M. & Highsmith J 2001).

The following quote from The Manifesto for Agile Software Development1 gives a summary of it's purpose:

"We are uncovering better ways of developing software by doing it and helping others do it. Through this work we have come to value:

i. Individuals and interactions over processes and tools.
ii. Working software over comprehensive documentation.
iii. Customer collaboration over contract negotiation.
iv. Responding to change over following a plan.

That is, while we value the items on the right, we value the items on the left more."(Constantine L 2001).

Ultimately we believe that the developers and organizations involved in Web engineering projects are the primary factor in the success or failure of Web application development. Given the diversity of disciplines required to develop Web-based applications, we are of the opinion that the AWE Process, or any other process or methodology, can only hope to have a second order effect on project success. Thus, we hold the belief that the agile route with its focus on people.

Our belief that people are the most important factor in project success is the fundamental reason why we have not tried to develop a monumental process to tackle the problems associated with Web application development. Many monumental processes attempt to codify good practice and experience in too much detail and for developers who do not understand the importance of what they are doing! This often results in development projects using monumental processes as cookbook recipes, where developers are lulled into a false sense of security by following the recipe in detail rather than using the ingredients selectively to help them build software deliverables that solve their problem space.

AWE is an iterative and incremental process; we believe this will allow for: early and continuous delivery of valuable software; the ability to harness changing requirements, even late in development; and the delivery of working software frequently. The AWE Process supports multidisciplinary development treating business experts, domain experts, and creative designers as developers along side software engineers (Miller, G 2001)

The following are the characteristics of agile software processes from the fast delivery point of view, which allow shortening the life-cycle of projects:

1. Modularity on development process level
2. Iterative with short cycles enabling fast verifications and corrections
3. Time-bound with iteration cycles from one to six weeks
4. Parsimony in development process removes all unnecessary activities
5. Adaptive with possible emergent new risks
6. Incremental process approach that allows functioning application building in small steps
7. Convergent (and incremental) approach minimizes the risks
8. People-oriented, i.e. agile processes favor people over processes and technology
9. Collaborative and communicative working style (Miller, G 2001).

Surveys of Web engineering practice by (Andrew McDonald and Ray Welland 2002) have identified seven characteristics of Web engineering that must be addressed by a Web engineering processes. These are support for:

1. Short development life-cycle times
2. Different business models (Business Process Re-engineering)
3. Multidisciplinary development teams
4. Small development teams working in parallel on similar tasks
5. Business Analysis and Evaluation with End-Users
6. Explicit Requirements and rigorous Testing against requirements
7. Maintenance

Since agility is desired, the principles of Agile Modeling should be followed.

Most of today's Web application development processes are extensions of standard software engineering processes. The usual iterated waterfall model is too rigid an approach to developing Web Applications. The waterfall model process was perfect for developing a file maintenance program for mainframes, but far too restrictive a process for building a Web application. Web application development needs to be an iterative process and most agree that a spiral approach is best. But, the exact steps at each cycle of the spiral are debated, as is the metric to be used to determine the completion of a cycle.

A few of today's Web application development processes have been derived from a business- oriented approach to applications development (Standing Craig 2002). Most of these processes develop a business plan for the e-business associated with the Web application, sometimes re-engineering the business along the way, and use things like Return on Investment (ROI) as a metric for the Web application development process

Extreme Programming

Extreme Programming (XP) has evolved from the problems caused by the long development cycles of traditional development models.

It first started as "simply an opportunity to get the job done" (Beck, K 1999) with practices that had been found effective in software development processes during the preceding decades (Beck, K 1999). After a number of successful trials in practice, the XP methodology was "theorized" on the key principles and practices used. Even though the individual practices of XP are not new as such, in XP they have been collected and lined up to function with each other in a novel way thus forming a new methodology for software development. The term 'extreme' comes from taking these commonsense principles and practices to extreme levels.

XP consists of 12 related practices and works best for small teams of 5 to 15 developers. Rather than focus on paper-based requirements and design documentation, XP concentrates on producing executable code and automated test drivers. This focus on source code makes XP controversial, leading some to compare it to hacking. We believe this comparison is unjustified because XP highly values simple design, and counters hacking claims by emphasizing refactoring, strong regression testing, and continuous code inspections through pair programming (J. Zettel et al 2001).

Small Firms and Standards

Many small firms are unaware of existing software process assessment models and standards. There's often the assumption that assessments conformant to these models and standards can be expensive and time consuming, and therefore difficult to perform in small companies. Small organizations also perceive assessment models and standards—including documentation and process-formalization practices—as targeting large organizations (M.C. Paulk 1998).

Small firms do not have the managerial experience, the financial resources and the methodological know-how to manage Web-based applications projects the way large firms do. Yet despite this, some small firms are satisfying their software development needs offshore (Brian Nicholson and Erran Carme 2002) There are two battles over process that every small software company must win to be successful. The first is the battle to convince the company to adopt reasonable development processes. Discussion of what makes up a good process may be an interesting meditation, but is entirely moot until the company commits to a policy of process improvement. The second battle is never over. It is to change existing processes to match changing circumstance (Robert P et al 2001).

A first step toward process improvement is identifying the strengths and weaknesses of an organization's software processes to determine effective improvement actions. An assessment can help an organization examine its processes against a reference model to determine the processes' capability or the organization's maturity, to meet quality, cost, and schedule goals. Several software process assessment models have been developed, such as CMM/CMMI, ISO 9001 Quality Management (including 9000-3), and ISO/IEC 15504—sometimes called SPICE. However, small companies (1–49 employees) find it difficult to run assessments (T. Mäkinen 2000).

SURVEY METHODOLOGY

The units of analysis for the survey were Jordanian small Web firms undertaking Web development. The target population included all small firms in Jordan which develop Web applications for sale as well as in-house software development groups within firms. The SEI [10] questionnaire was used as the survey instrument. Prior to the data collection, the survey instrument was pre-tested to enable clarification of constructs; to provide the means of operationalising selected constructs; and because pre-tests can be useful in qualitatively establishing the reliability, construct validity, and content validity of measure. In order to locate and correct weaknesses in the questionnaire, the questionnaire was pre-tested using face-to-face interviews with 10 random developers. The selection of interviewees for these pre-tests was designed to obtain maximum feedback from Web developers in various roles.

Prior to the pre-test, the following checklist was used to review the questionnaire instrument:

- Will the words be uniformly understood?
- Do the questions contain abbreviations or unconventional phrases?
- Are the questions too vague?
- Is the question too precise, biased or objectionable?
- Is it a double-barrel question?
- Does it have a double negative?

- Are the answer choices mutually exclusive?
- Has too much knowledge been assumed?
- Is the question technically accurate?
- Are the questions too cryptic?

During the pre-test, concerns were raised about the section headings and question groupings of the SEI questionnaire (such as metrics, standards and procedures, control of development Process). During the pre-test, the respondents completed the questionnaire in the presence of the researcher, and identified any difficulties with interpretation of words or questions. As well as testing the reliability and construct validity, the pre-tests served as 'dry runs' for the final administration of the instrument. The feedback was not adequate. So, some parts of the questionnaire were translated into Arabic Language (the mother tongue of respondents), and the pre-test was carried a second time. The sample used consisted of twenty small Web development firms in Jordan. Firms were eligible for inclusion in the survey when they (1) have had developed Web applications during the last 12 months and (2) did not have more than 50 employees.

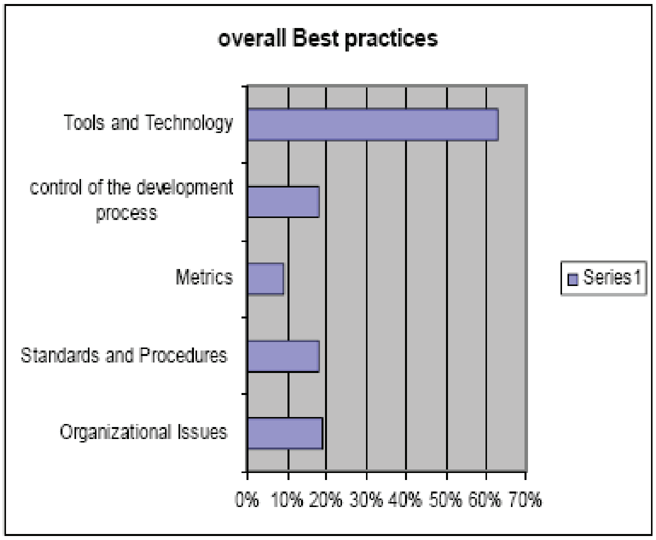
The questionnaire is organized into two sections. The first section collects information on the respondent's background.

This includes current position, past experience, roles played among others. The second section, entitled Web engineering best practices, concentrates on the adoption of best practices by the respondent's organization. These include organizational Issues, standards and procedures, Web metrics, control of the development process, and tools and technology.

RESULTS

The statistical analysis showed that the majority of respondents have 5 years or less of experience in their present organizations, and the same number of years of overall software experience. The majority of them are in software engineering process group positions. The highest percentage of the study participants is involved in design activities. The majority of the participants were never involved in software process improvement activities. The analysis showed a significant weakness in the levels of adoption of Web engineering practices by small Jordanian Web development firms. The ratios of adoption levels were as follow: organizational issues (19%), standards and procedures (18%), Web metrics (9%), control of the development process (18%), and tools and technology (63%). The statistical analysis shows that Web metrics got the lowest ratio; which implies that the majority of respondents are not familiar with this practice. Tools and technology got the highest ratio; which implies that this practice is the most applied practice in these firms, and this reflects the widespread use of these tools in the development process worldwide. Organizational issues, standards and procedures and control of the development process got similar ratios. But, for the organizational issues, and standards and procedures, the majority of respondents answered with "No". This implies that the majority of respondents are not familiar with these two practices. For the control of development process practice the majority of respondents answered with "Does not apply". This implies that the respondents are familiar with this practice, but they don't apply it, and this leads one to say that the development process models used by these firms are ad hoc. The analysis also showed that there are significant differences in the levels of adoption of Web engineering best practices between European firms and small Jordanian Web development firms. The overall average of adoption levels of best practices in European countries is 51%, while the overall average of adoption levels in small Web development firms in Jordan is 25% see (Figure 1). The big difference in average between them implies that there is a big gap between the adoption levels of best practices in European countries and

Figure 1. Overall best practices adoption in small Jordanian firms



small Web development firms in Jordan. The final conclusion is that small Web development firms in Jordan have a lot to do in order to succeed in a very competitive market.

The Model

According to (Abdesselam Redouane 2002) Web-based firms have very stiff and stringent conditions. They have limited resources. This will hinder the quality of the product and ultimately the success of these companies. It is usually the case that the few people, who carried out the development, will also perform the testing of the end product. This is a poor practice, as it does not allow the test to be carried out rigorously and it will be certainly biased.

According to the results we got from the surveys that we have done, we found that small Jordanian firms have many problems:

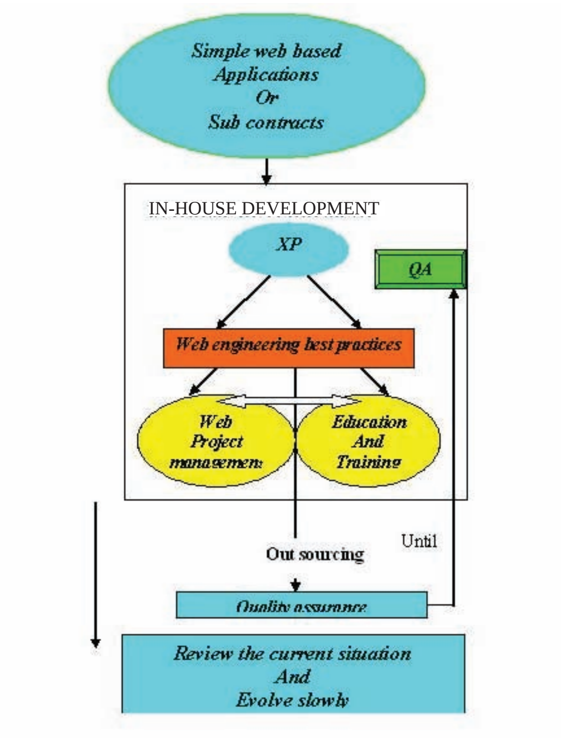
Web projects are always, in most cases (firms) run over budget and time , have limited resources, number of developers is very small, quality assurance activities are done by the people who carried out the development, add hoc development approaches used , poor Web project management, and Web engineering practices is not implemented .

Based on results of survey and literature analysis we proposed the following model (Figure 2):

1. Start with simple Web projects or implement sub contracts projects: The surveys and literature stated that most of Web development projects in small firms are small or medium size projects (1-2 months or 3-6 months). In small Web projects it is highly desirable to plan all phases at the start of the project. This means that the phase sections of the Web project management plan, Web configuration management plan and Web verification and validation plans are combined. The plans may be generated when writing the proposal. Typically, qualified developers are involved in large or medium-size projects whereas small Web projects are carried out by under-qualified or inexperienced Web developers. The reactionary development scenario and the lack of clear guidelines to face the process, push developers to follow an ad-hoc development process.

   The heavyweight software methodologies

Figure 2. The model



are limited to support such scenario. This is because they involve several stages and roles that require an important amount of communication and coordination in order to get a final product. On the other hand, there are the lightweight or agile methodologies that could have an interesting applicability to the described scenario.

2. Implement an agile development model, and we recommended using extreme programming that identify specific best practices within the Web development domain: Different software development methodologies and quality assurance methods are used in order to attain high quality, reliable, and bug free software. Extreme Programming (XP) is a software development methodology that integrates many of the known ideas (that we all were familiar with) in order to achieve such software systems. Specifically, XP emphasizes code-unit testing (preferably before its writing), and thorough testing of software functionality. The contribution of XP to software development is expressed, among other ways, in the quality improvement of both the entire process of software development and of the software quality itself. Currently, XP is used mainly in small-medium size software projects [19].
3. Developers must apply the lower limit of Web engineering best practices: Many practitioners in the field of Web engineering and software engineering have commented on the lack of suitable software engineering processes that can be used to build Web applications. We investigated the way industrial Web engineering is being carried out by making a survey consists of questions relating to the development process being used to develop Web applications and Web engineering practices. If a Web engineering process is to be successful then it must address the following:
    1. Short development life-cycle times.
    2. Delivery of bespoke solutions.
    3. Multidisciplinary development teams.
    4. Small development teams working in parallel on similar tasks.
    5. Analysis and Evaluation.
    6. Requirements and Testing.
    7. Maintenance.
4. Testing and quality assurance (QA) activities must be carried out by another qualified firm, until an internal qualified quality assurance department is established
5. Web project management issues specific to Web development; must be carried out.

   In order to organize and manage a Web development project successfully, one must combine specific knowledge, skills, efforts, experience, capabilities, and even intuition.
6. Training and education: Good Web engineering practice requires expertise in a complex

set of activities that involve the intellectual skills of planning, designing, evaluating, and revising. A Web engineering process must take into account the different types of developer required to build a successful solution.

In order to this framework to be successful, all of the people involved in Web development (developers, mangers,) must have a good knowledge in Web engineering development and they must be trained.

The best way to learn modified XP is in an experiential-learning training course. Your entire development team (including the testers, the customer, and the manager)

Should attend a one-week immersion course on XP. While many programming teams are learning XP based solely on books and information on the Web, it is important to actually do the practices with guidance.

7. Review the current situation; evolve slowly, until development method is clear.

Figure 2 shows the model and explains the model elements.

CONCLUSION

The results showed that there is a weakness in applying Web engineering practices in small Jordanian Web development firms. The results of the study lead to the following recommendations to improve Web development practice in small Web development firms in Jordan:

1. Development teams should be multidisciplinary: A Web engineering process must take into account the different types of developer required to build a successful solution. The process

   Should ensure that all involved understand their roles and responsibilities, and where overlap occurs understand how to resolve conflict in the best interests of the project in question.
2. Firms should apply project management best practices to improve organizational issues.
3. Firms should pay attention to the quality management and standards.
4. Firms should apply software engineering best practices to improve the execution of their Web engineering projects.
5. Software process improvement initiatives should be considered.
6. Education and training: A proper and regular training of employees especially on newer aspects in Web engineering will lead to an increase of acceptance and usage.

Small firms need to adopt a suitable Web development model and must apply Web engineering best practices in order to survive. Web project management and education is also important for the development process. Testing is the main key for quality, so small firms should outsource their testing activities at the initial phases when apply the model. The model may be helpful for small Web firms in improving their way in development, where small Web firms have very stiff and stringent conditions. They have limited resources.

The model may be considered useful for Web project development in small firms, where it based on agile development methods, a set of Web engineering practices, and Web project management.

REFERENCES

Avison, D. E., & Fitzgerald, G. (2003, January). Where Now for Development methodologies. *Communications of the ACM, 46*(1), 79-82.

Bahli, B., & Di Tullio, D. (2003). Web Engineering: An Assessment of Empirical Research. *Communications of the Associations for Information Systems, 12*, 203-222.

Beck, K. (1999). Embracing Change with Extreme Programming. *IEEE Computer, 32*(10): 70–77.

Beck, K. et al., (2001, February). Manifesto for Agile Software Development. *The Agile Alliance*, http://www.agilealliance.org/

Cater-Steel, A. P. (2004). Low-rigour, Rapid Software Process Assessments for Small Software Development Firms. *Proceedings of the 2004 Australian Software Engineering Conference (ASWEC'04).*

Constantine, L. (2001, April). Lightweights, Heavyweights and Usable Processes for Usable Software. *Keynote at Software Development 2001.*

Cuitter Consortium Research Briefs (2000). http://www.cutter.com

Deshpande, Y., & Gaedke, M. (2005). Web Engineering: Developing Successful Web Applications In A Systematic Way. *14th International World Wide Web Conference*, 10-14 May, 2005, Chiba, Jap.

Deshpande, Y., Murugesan, S., Ginige, A., Hansen, S., Schwbe, D., Gaedke, M., & White, B. (2002). Web Engineering. *Journal of Web Engineering, 1*(1), 3-17. Santa Barbara, CA.

Fayad, M. E., Laitinen, M., & Ward, R. P. (2000). Software Engineering in the Small. *Communications of the ACM, 43*, 115-118.

Fayad, M. E., Laitinen, M., & Ward, R. P. (2000, March). Software Engineering in the Small. Communications of the ACM, *43*(3).

Fowler, M., & Highsmith J. (2001, August). The Agile Manifesto. *Software Development Magazine*. http://www.sdmagazine.com/documents/s=844/sdm0108a/0108a.htm

Gresse von Wangenheim, C., Anacleto, A., & Salviano, C. F. (2006). *Helping Small Companies Assess Software Processes, IEEE Software*, (pp. 91-98).

Mäkinen, T., Varkoi, T., & Lepasaar, M. (2000). A Detailed Process Assessment Method for Software SMEs. *Proc. 7th European Software Process Improvement Conf. (EuroSPI).* www.iscn.at/select_newspaper/assessments/tampere.html.

McDonald, A., & Welland, R. (2002).. Available at http://www.dcs.gla.ac.uk/ accessed *Evaluation of Commercial Web Engineering Processes* 20/5/2006.

Miller, G. G. (2001). *The Characteristics of Agile Software Processes.*

Murugesan, S., Deshpande, Y., Hansen, S., & Ginige, A. (1999, May). Web Engineering: A New Discipline for Development of Web-based Systems. *Proceedings of the First ICSE Workshop on Web Engineering, International Conference on Software Engineering*, Los Angeles, http://aeims.uws.edu.au/WebEhome/ICSE99-WebE-Proc/San.do

Nicholson, B., & Carme, E. (2002). Offshore Software Sourcing by Small Firms. *An Analysis of Risk, Trust and Control.*

Paulk, M. C., (1998). Using the Software CMM in Small Organizations. *Joint 1998 Proc. Pacific-Northwest Software Quality Conf. and 8th Int'l Conf. Software Quality, PNSQC/Pacific Agenda*, (pp. 350–361). www.pnsqc.org/proceedings/pnsqc98.pdf

Pressman, R. S. (1998, September/October). Can Internet-Based Applications Be Engineered? *IEEE Software.*

Pressman, R. S. (2003). Software Engineering: a Practitioner's Approach (4th ed.). McGraw-Hill Publishing Company.

Pringle, L. (2001). Size does matter: improvement for SMEs. *Software (SEA National)*, (pp. 4-7).

Ray, K C. (2002, August). Adopting xp The path to—and pitfalls of—implementing Extreme Programming. *STQE* (pp. 34-40). www.stqemagazine.com

Redouane, A. (2002). Guidelines for Improving the Development of Web-Based Applications. *Proceedings of the Fourth International Workshop on Web Site Evolution (WSE'02)* 0-7695-1804-4/02 2002 IEEE.

Rodriguez, D., Harrison, R., & Satpathy, M. (2002). A Generic Model and Tool Support for Assessing and Improving Web Processes. *Proceedings of the Eighth IEEE Symposium on Software Metrics* (METRICS.02) 0-7695-1339-5/02 © 2002 IEEE.

Standing, C. (2002, March). Methodologies for developing Web applications. *Information and Software Technology*, *44*(3), 151-159.

Stein, S. (2006). *Software Process Improvements in a Small Organization*. Master Thesis ,Software Engineering Thesis no: MSE-2006:01 January 2006. Available at: http://quality.hpfsc.de/master-sstein.pdf accessed 4/6/2006.

Stein, S. (2006). Software Process Improvements in a Small Organization. Master Thesis ,Software Engineering Thesis no: MSE-2006:01 January 2006. Available at: http://quality.hpfsc.de/master-sstein.pdf accessed 4/6/2006.

Stein, S. (2006). Software Process Improvements in a Small Organization. Master Thesis. Software Engineering Thesis no: MSE-2006:01 January 2006. Available at: http://quality.hpfsc.de/master-sstein.pdf accessed 4/6/2006

Ward, R. P., Fayad, M. E., & Laitinen, M. (2001, April). Software Process Improvement in the Small. *Communications of the ACM*, *44*(4).

Web Engineering Home Page [http://fistserv.macarthur.uws.edu.au/san/WebEhome/]

Zettel, J., et al. (2001). LIPE: A Lightweight Process for E-Business Startup Companies Based on Extreme Programming. *Proc. Third Int'l Conf. Product-Focused Software Process Improvement (PROFES 2001)*. Berlin: Springer Verlag.

KEY TERMS AND DEFINITIONS

Agile Development: Refers to a group of software development methodologies that are based on similar principles.

CMMI: A process improvement approach that provides organizations with the essential elements of effective processes.

Extreme Programming: Is a software engineering methodology, the most prominent of several agile software development methodologies.

Small Firms: Small firms are companies which employ a relatively low number of workers.

Software Process: The total set of software engineering activities necessary to develop and maintain software products.

Web-Based Applications: Is an application that is accessed via Web browser over a network such as the Internet or an intranet.

Web Project Management: Project management is a methodical approach to planning and guiding project processes from start to finish.

This work was previously published in Utilizing Information Technology Systems Across Disciplines: Advancements in the Application of Computer Science, edited by E. Abu-Taieh; A. El-Sheikh; J. Abu-Tayeh, pp. 130-141, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 6.3
Employee Life-Cycle Process Management Improvement with Web-Enabled Workflow Systems

Leon Welicki
Microsoft, Canada

Javier Piqueres Juan
Systar, Spain

Fernando Llorente Martin
ONO, Spain

Victor de Vega Hernandez
ONO, Spain

ABSTRACT

Employee life-cycle processes management (hiring new employees, changing their conditions, and dismissing them) is a critical task that has a big impact in HR Information Systems. If these processes are not handled correctly the consistency of HR databases is compromised. In many cases (especially in small and mid-size business) these processes are implemented using semi-manual procedures based on unstructured information. In this chapter the authors will present the results of our real-world experience building a Web-enabled workflow system for managing employee life-cycle process instances in the context big Spanish telecommunications company.

INTRODUCTION

Employee life-cycle management is a critical task that affects all companies without regard of their size and business. These processes include hiring new employees, changing working conditions (promotion, demotions, change of cost centre, changes in the compensation package, change of function, change of organizational unit, etc.) and dismissals (end of relationship). In this paper we will present our real-world experience building a web-enabled

DOI: 10.4018/978-1-60566-304-3.ch020

workflow system for managing employee life-cycle process instances in a big Spanish telecommunications company. In the first section we will present ONO, our company, in order to set the organizational context. In the second section we will present the problem that we faced and set the requirements for building a tool to solve it. In the third section the web-enabled workflow system is presented, making special focus on the agile approach used to build it and how the previously stated requirements are met. Finally we will offer some conclusions and future lines of work

About ONO

ONO is the leading alternative provider of telecommunications, broadband Internet and pay television services in Spain and the only cable operator with national coverage. ONO offers its services to more than 1.8 million residential cable access and 69,000 business customers as of 31 March 2007, through its own state of the art networks which give direct access to nearly six million homes in franchises that cover the majority of Spain, including the nine largest cities. ONO is the principal competitor to the incumbent telecommunications and pay television operators in Spain. For the first Quarter 2007, ONO generated revenues of €1,608 million and EBITDA of €592 million, on an annualized basis. ONO has several offices all around Spain.

Ono is a young company in constant growth in search of excellence. Throughout its history has demonstrated great management skills and solid growth prospects, backed by a strong global investment in an infrastructure that reaches 6.8 million homes. Table 1 shows chronologically the main highlights that significantly transformed our company.

BACKGROUND

In this section, we will present our notion of employee life-cycle process and how it is related

Figure 1. ONO Spanish coverage and operating highlights. ONO offers their services to clients in almost all the national territory, covering more than 17.500 homes in Spain (according the Spanish INE)

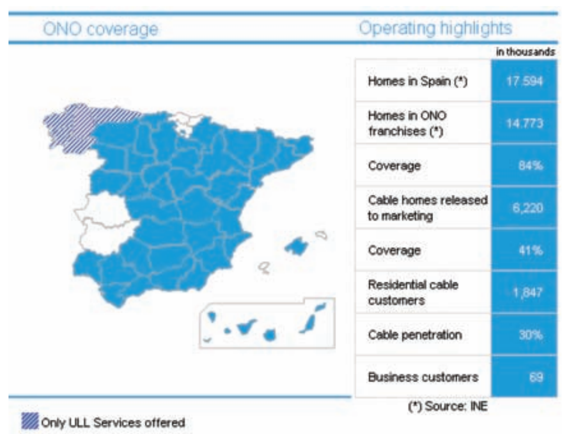


Table 1. Company background (chronologically)

Year	Event
1998	The ONO brand was launched in 1998, and was awarded the licenses to provide cable television and telecommunications services in the regions of Valencia, Castellon, Alicante, Murcia, Cadiz, Huelva, Cantabria, Mallorca, and Albacete.
2002	By the end of 2002, ONO was already present in approximately one in every three homes prepared to receive its services.
2003	In 2003, it was awarded the licence to operate in Castilla – La Mancha.
2004	In 2004, it acquired the telecommunications operator Retecal, covering Castilla-Leon.
2005	In November 2005, ONO closed the acquisition of 100% of the telecoms company Auna Tlc, extending its services to the communities of Madrid, Catalonia, Aragon, Andalusia (excluding Cadiz and Huelva, which already belonged to ONO), the Canary Islands, Catalonia, La Rioja, Madrid and Navarra.

to e-HRM. Following, we will detail how this processes where run at our company in our initial scenario to show their fragility and resource consumption intrinsic problems.

Employee Life-Cycle Process

We call "*employee life-cycle process*" to any process that modifies the conditions of an employee within our company. These processes include hiring new employees, modifying any of her contractual conditions (changes in the compensation, position, department, location, bonus, etc.), and dismissal. These processes are critical for the reasons detailed in the next paragraphs.

Employee Life-Cycle processes provides the input for HR databases. Therefore, if they are not managed correctly the consistency, organization, and trustworthiness of the information in HR databases are compromised.

They are run collaboratively by groups of persons that may not know each other and may be geographically dispersed (in different cities, states, countries, or continents, depending on the size of the company). Usually each process instance must go through a hierarchical approval workflow (e.g., hiring of a new employee requires an approval of the manager of the requester). When the processes are not formalized enough these workflows mutate on a case-by-case basis, hindering control over the ongoing processes.

Since they deal with information of people and this information is handled by people, there are non-formal issues that may arise (anxiety, envy, information protection, corporate politics, etc.) and may require special attention of the HR team.

The last (but not least) important issue regarding these processes is budget control: each one of these processes affects directly the HR budget for the company. Without budget control departments can hire above the HR budget forecast of the company.

Weak vs. Strong Employee Life-Cycle Process Management

An inefficient (non automated, non formal, non controlled) employee life-cycle process management results in lots of manual, repetitive, boring, low value, and error prone tasks that must be done by the HR team. All companies with a basic implementation of the HR module of SAP (or similar) may be doing a very basic management of employee life-cycle processes (all changes are recorded in a centralized database). However, there is no formal support for the approval workflow, nor management of the people issues at the stages before the data is introduced in the database. In these implementations, the processes may be performed manually (or semi-manually or using unstructured information sources like email). In this case the HR team needs to do an enormous amount of back-office low added-value work: paper chasing, consolidating requests, check for collisions when dealing with a particular request, manually creating reports, deal with each process as a "state-of-the-art" unit, control the HR budget, ensure privacy of information, normalize inconsistent input, etc. (the list could go for several lines more). We call this "weak employee life-cycle process management".

In contrast, when these processes are managed correctly they can change the way HR works within a company since all the manual, boring, error prone, and low added-value work is removed allowing HR people to innovate, drive, envision, and improve the organization. As a plus, upper management has up-to-date just-in-time information on the state of on-going employee life-cycle processes and real time information about the current HR budget. We call this "strong employee life-cycle process management".

Employee Life-Cycle Process and e-HRM

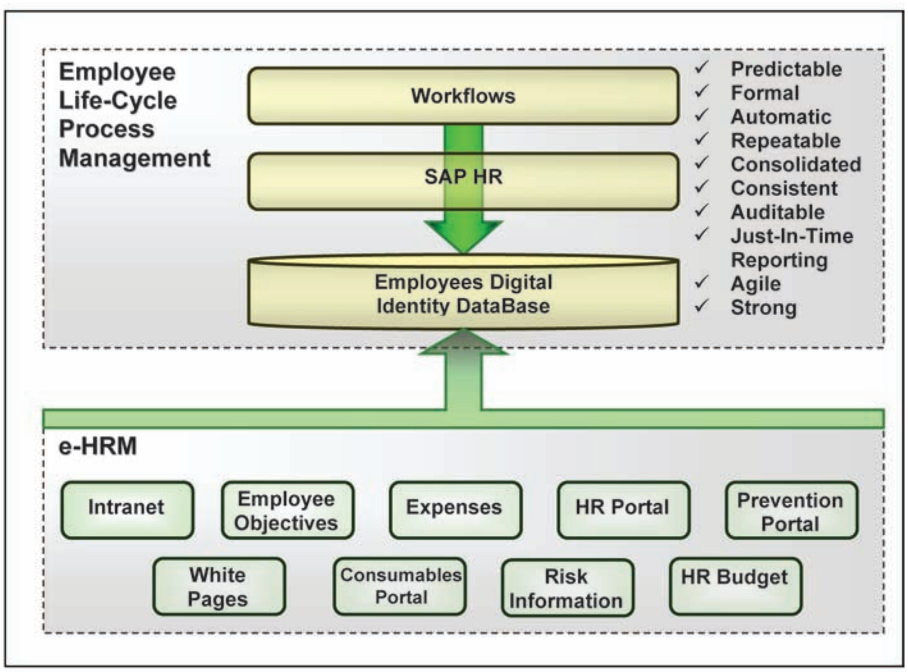
Ruël, Bondarouk and Looise (2004) state that e-HRM is a way of implementing HR strategies, policies, and practices in organizations through a conscious and directed support of and/or with the full use of web-technology. The use of web-technology may render in a set of Intranet based applications distributed within the company. Each one of these applications may focus on a concrete aspect of interest for the employee (Employee's Portal, Vacations Management, Travel Expenses, etc.) or the company (Objectives Management, Work Reports, Planning, etc.).

Each application may have been built to manage its "small universe" of information. However, there is a common denominator for all those applications: all of them deal with information about employees.

When strong employee life-cycle management is implemented within an organization we can expect a trustable source for accessing consolidated up-to-date employee information. In contrast, when weak employee life-cycle management is implemented we cannot guarantee the trustworthiness of our employee information repository (e.g. some processes may have been processed but not introduced in the system; a human error in consolidating thousand-of-excels may result in errors within the hierarchical structure of a department).

eHRM is feasible without a proper employee life-cycle process management, although not advisable (specially in large companies). Weak employee-life cycle process management has a big impact in all HR applications within a company regarding quality and reliability of its information and workflows. The lack of reliability, auditability,

Figure 2. Manual employee life-cycle processes management is error prone and time consuming, resulting in a weak employee's digital corporate identity on the employees. This creates the need for lots of manual or semi formal integration processes as an immediate negative consequence (represented by the arrows from each application to SAP HR)



and normalization in employee life-cycle process management leads to a weak employee's digital identity (who is a person within an organization, including position, department, location, duties, professional category, etc.). As a consequence, lots of manual or semi-formal integrations must be done, taking lots of hours of the HR department impeding them to do "interesting things". We could summarize this in the following phrase: "*HR is working for the System, but the System is not working for HR*".

Employee Life-Cycle Processes at ONO

In the last years, ONO experienced a very fast growth in a very short period of time (Table 1). The biggest leap was the merge with Auna (a company that had the same size than ONO at the time of the merge). One of ONO's main characteristics was being very agile, with very simple and human-centred processes. Some of them didn't scale to the new context of the company, since they were designed and implemented having another model of enterprise in mind. The employee life-cycle management support processes (hiring, changing working conditions, and dismissal) fell in this group.

Each of these processes were based on paper-printed documents (that in the best case where based in corporate Excel templates), with all the problems that this implies including inconsistency in the input (files were completed differently by the requestors), traceability (all the Excels were sent internal mail), problems for enabling teamwork (there wasn't any mechanism for dispatching the requests and therefore a single request could be handled incorrectly by several HR (Human Resources) employees at the same time), lack of a unified way for notifying the participants in the flow, lack of tools for reusing the requests, lack of reporting (there wasn't any automatic way to get a report of what was going on in the company), lot of human effort of very low added-value (HR managers consolidating hundreds of Excel files and e-mails), among other problems. To make things worst, ONO has a region-based organizational model (based on the division of the company in the Spanish geography), increasing the impact of the above mentioned problems.

Description of the Manual Employee Life-Cycle Processes at ONO

ONO is a young company, and therefore some of their processes matured in the last years and some of them are still maturing. The employee life-cycle management processes were manual, based in interactions among people that knew each other (at least by telephone). They were manual-run workflows where all the participants worked collaboratively on the same physical paper document.

Following, we will describe briefly (and in a broad way) how the processes were run:

1. A user creates a request document (for staring a new hire, a change in the conditions of one of his employees, or a dismissal). In the best cases this document was based on a corporate template. While this was the best case, it was far from being ideal: since the requestors weren't HR specialist (a petitioner could be any manager that needs to hire a new employee or to promote one of his employees) usually the document wasn't correctly fulfilled.
2. After the document was created and printed, it was signed by the petitioner and by her manager.
3. The petitioner notifies the HR department (by email or by phone) and then he sends the physical document by internal mail.
4. HR receives the document and validates it. If they need further clarification, they would contact the petitioner or her manager to discuss about the request.

5. In the case that the conditions of the hiring / promotion / change / etc. varied significantly after the discussion, a new physical document needed to be issued and signed again.
6. HR registers the transaction, contacts the employee that would be hired / promoted / fired / etc., does their usual tasks according the type of process, and registers all the information in their HR systems (payroll, SAP HR module, etc.)
7. Periodically (twice a month), the HR staff manually created reports of the ongoing processes to inform the upper management. They also used this information to verify that the HR budget wasn't overrun (this verification was also manual on a process-by-process basis).

The above description is a simplified version of the manual processes, with the goal to illustrate its fragility. It was very error prone, produced lots of unnecessary work, unreliable (in some cases, papers where lost and the process needed to be started over), and unique for each petitioner (according to her personality). It consumed lots of time of the HR team performing tasks of near-to-null value, such as verifying the input data and trying to interpret it, creating reports manually, tracking papers, notifying personally each of the actors in the process, and controlling the budget.

IMPROVING EMPLOYEE LIFE-CYCLE PROCESSES WITH A WEB-ENABLED WORKFLOW SYSTEM: AN AGILE APPROACH

In this section we present the main problem and our solution proposal. We start showing how our weak employee life-cycle processes (detailed briefly in the section "Employee Life-Cycle Processes at ONO") didn't scale when our company increased its size and rapidly became an important issue to be solved. Next is detailed how this problem derives in the need for a tool and establish a set of requirements for that tool. Following we elaborate on our response to the problem: creating a web-enabled workflow system using an agile approach. In order to make our ideas true we benefited from agile concepts for modeling the business processes and building the solution. The use of agile methods allowed us to create a high-functionality and quality application on a low budget with early and frequent deliverables.

ONO-AUNA Merge Related Problems

The process mentioned above worked fairly well in mid-to-low-size environment (less than 2.500 employees), with a relatively small HR department, and a very people oriented culture. Since the process was mainly based in human interactions, it worked better when the participants knew each other. Additionally, each instance of the process was highly dependant in the actors: some requesters (hiring managers in the hundreds of departments of the organization) where good "process players" and sent the information in good shape while others where very chaotic.

When ONO bought Auna and the merge started the company doubled its size. A manual process mainly based in interactions among people that knew each other didn't scale well in the new scenario, in one part because the size of the company increased but mainly because the people running the process didn't knew all their counterparts anymore (and given the new size of the company, it was very improbable that they would do ever).

To worsen things, the amount of HR transactions regarding employee life-cycle increased exponentially: is well known by everyone that in a merge process lots of new people come to the company, lots of people leave the company, and even more people changes position (in the average, there are two employees for each position, since both companies where in the same business).

A final added problem was traceability and auditability: since the process was based in physical paper sheets, it was mandatory to keep the original papers for a time period according with the Spanish law. During the merge, lots of work centre moves were done. Each move affected the paper files and therefore increased the difficulty to locate the papers (they could be lost or at any branch of the company). In these cases, lots of time of the HR was lost just doing "paper chasing".

The Need for a Tool

It was very clear that the company needed a tool to assist the employee life-cycle management, with the goal of making it easier, more reliable, predictable and auditable, to provide all the participants in the process the information they need just in time without any further hassle, and to provide the upper management with reporting tools to know the global picture regarding the overall HR budget of the company.

After jointly studying the problem ONO's HR and IT departments decided to build a set of tools to support the employee life-cycle processes. The main goal of these tools was to enable the collaborative work between all the actors implied in each of the employee life-cycle management processes. Each of these actors should be able to interact with the new tool in a very simple and efficient manner. Additionally, the tool should be proactive, providing a "push model": each participant in the process should be notified whenever a process instance requires his participation. The tool should support the approval workflow (Workflow Management Coalition, 1995) for each type of process, dynamic headcount validation and automatic reporting.

Requirements for the Tool

The following requirements where established by ONO's HR department:

- **HR budget control:** Provide an automatic control of the HR budget at all the appropriate levels. For example, when a new request is issued it should be checked against the requesting department HR budget to verify if the request is valid. At a higher level the upper management and HR directors need to have reporting facilities to have a general view of evolution the overall HR budget of the company.
- **Support for the approval workflows:** Model the manual processes using a workflow systems. This implies modelling semi-formal processes using formal specifications.
- **Enabling collaborative work:** The tool should allow each employee to participate in the approval workflow at the right moment. It should inform the participant in the flow of the new events that require their attention. It should also give a unique single point for checking the status of the on going processes (eliminating e-mail and phone).
- **Automatic reporting:** Generate all the reports automatically, without manual intervention.
- **Traceability and auditability:** All the process instances should be traceable and auditable, giving a fine grained control over the past events.
- **Reducing manual work:** Reduce all the low value added manual work.
- **Input consistency:** Simplify data input in order to give reduce ambiguity and improve consistency in data that enters to the approval flow.
- **Reduce bottlenecks in the process:** Provide tools to avoid a single person to stop all the ongoing flows when she is not available.
- **Ease of use and far reach:** Create a very easy to use tool in order to make the learning curve as low as possible. It should also

accessible by any employee of the organization without the need of any complex setup in his computer.

A Web-Enabled Workflow System as a Response to the Problem: Benefiting from Agile

After the problem was clearly established ONO's IT and HR departments started working on how to deliver a cost and time effective solution following ONO's cultural principles (Davenport & Prusak, 2000). The working group determined that the best approach for solving the problem was building a web-enabled workflow-based system, since it would empower the collaborative work on the employee life-cycle requests without any special requirement in the client computers.

The group also decided to build it with a low budget without sacrificing the quality and functionality, using the technologies and tools available at the company (mainly based in the development framework created by ONO's IT department). An agile approach was the methodological choice since it allowed to build small deliverables in short iterations, delivering value in small (but continuous) slots of time. In the economic side, this approach would allow to build with the resources at disposal: if the team was short on economic resources, it could do less ambitious iterations or even do no iteration without affecting the overall project (whenever any iteration was finished a new fully working module was delivered and deployed in production).

An Agile Approach for Modeling the Business Processes

The processes were modeled after a careful analysis of their manual counterparts. A joint work between One's HR and IT department was performed in order to come out with the most accurate representation of the business processes involved in the employee life-cycle.

The business process modeling was a very important part of the process: lot of tacit organizational knowledge needed to be transformed in explicit knowledge (Nonaka & Takeuchi, 1995) that can be used as a formal specification. It was a big challenge to come out with a model that clearly represents the interactions for each employee life-cycle process: since the process was semi-formal, there where significant ambiguities that needed to be resolved.

The main premise in this specification process was *simplicity*: our goal was to come out with the simplest process definition as possible. All the resulting flows are very similar: this is the result of a commonality analysis (Gamma, Helm, Johnson, & Vlissides, 1995) between all the processes in order to find all the similarities between processes and model them as uniform as possible to make them simpler and easier (Figure 3).

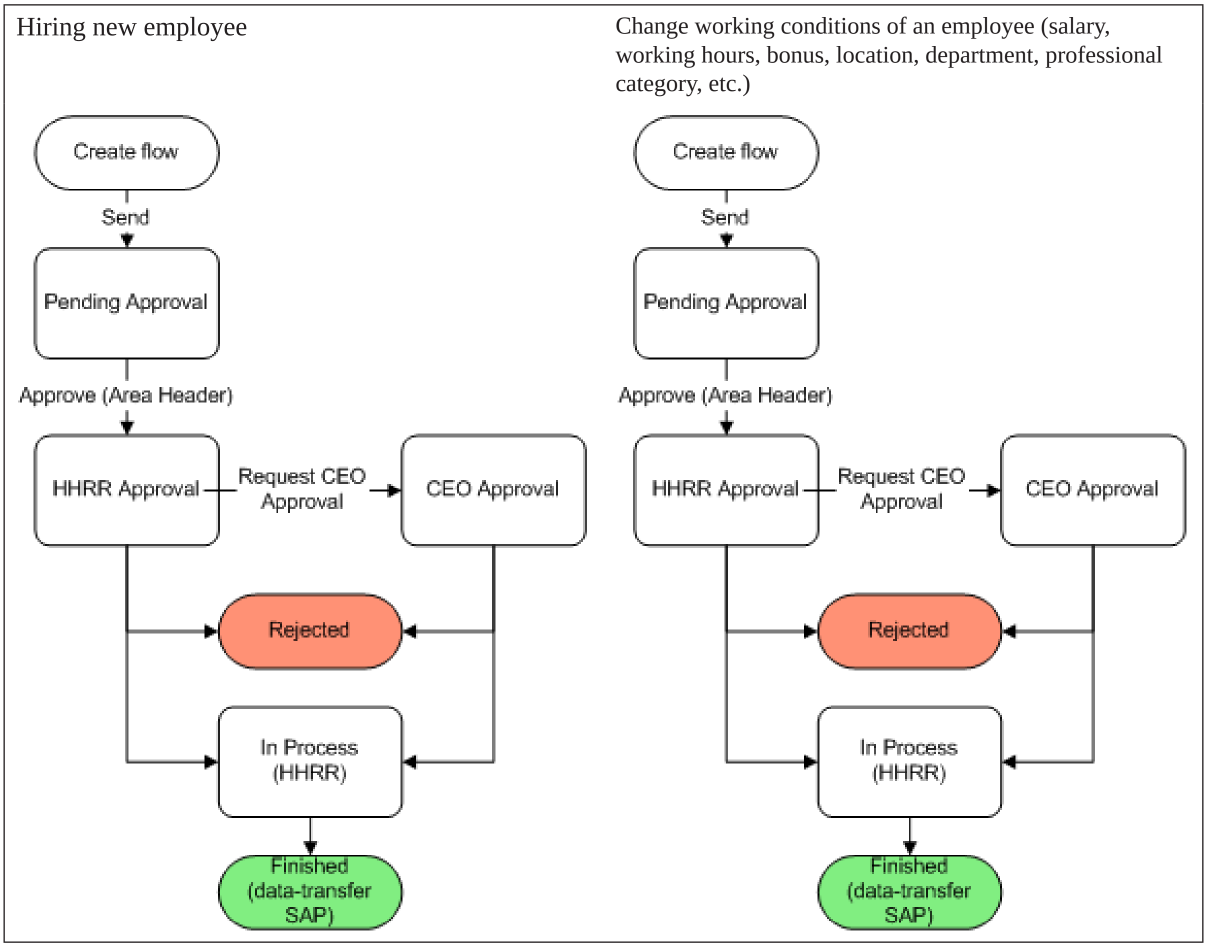
Observing, Learning Building, and Delivering: An Agile Approach for Modelling the Business Processes

We took an agile (Agile Alliance, 1999; Beck, 1999; Martin, 2002; Schwaber & Beedle, 2001) approach built our first model iteratively and incrementally: instead of trying to come out with the final version of the three flows up-front we went through several short iterations on the first one (the hiring process) and improved it using the feedback and experience gained from its real users.

We created a simple implementation of the hiring process according to the more immediate needs of HR. When it was in production we actively observed the problems and situations that arose to our users when using the tool and the error logs of the system. With all this information we built packages of fixes (in short sprints) and delivered new versions taking the same approach (observing, learning, building, and delivering).

This approach allowed us to constantly deliver value and to supply a working solution to our users without spending unnecessary time in the

Figure 3. After several iterations and commonality analysis, our "Hiring" and "Change of Conditions" flows are almost identical. In the manual process the flows where significantly different. Currently both flows are formal, determining automatically and in a predictable way their participants. The only difference among both flows is in the control of the HR budget (this is strongly checked on the hiring flow), but does not affect the flow neither the participants (the instance of the flow is marked with special information to indicate if it is or it is not within the forecasted headcount for the ongoing quarter)



analysis phase (lot of the changes that we included in the packages where product of our continuous improvement process).

An Agile Approach for Implementing the HR Workflow System

The Software Development area in ONO's IT department works under the principle of dividing big projects in small chunks in order to deliver functionality earlier in shorter intervals of time and with more client checkpoints. This allows building applications with smaller budget, deliver early working products to the clients, and giving them the chance to change or add new requirements after each iteration is finished. The main goal of our approach is to deliver value to clients in the shortest interval as possible and to evolve the functionalities based on real world information mined from the usage of the application (observe, learn, build, and deliver).

ONO's Software Development area uses a variant of SCRUM (Schwaber & Beedle, 2001). Each sprint (Schwaber & Beedle, 2001) starts with the creation of a scope document that is validated by IT and its clients (in this case, HR). After the document is approved, a high level architecture and high level designs are built and a product backlog is created. This backlog may also contain small user stories on each of its items. The backlog is revised on a daily basis doing stand-up meetings. In that stand-up meetings, some architectural or design issues may arise. In that case they are discussed later in other special purpose meetings, in order to make the daily stand ups as short as possible and not keeping all the members of the team busy with things that are not interesting for them. In our SCRUM variant, the sprints are not of a fixed size and we have an architect role to ensure conceptual integrity (Brooks, 1995) and alignment and synergies with Ono.CDI (Ono Content Driven Infrastructure), our corporate software development framework.

We calculate the size of the sprint according to the commitments established with the clients. However, we always work actively with the clients to keep the sprints shorter than six weeks. Ideally, and on the average, the sprints have three weeks duration.

After sprint is finished, a User Acceptance Test (UAT) with the client area is performed. In this test some minor issues may arise, so we always plan some days in our schedule to fix these minor issues. After the UAT is passed and the fixes are approved we deliver the product of the iteration.

Deliverables Planning: Technology for the Masses

HR and IT established to policy of "*Deliver value to 4980 employees first and later to the remaining 20*". This was our guiding metaphor (Nonaka & Takeuchi, 1995; Beck, 1999) and means that we work to deliver functionality to all the company first (the front-end of the application) and leave the detailed back-end operations for later iterations.

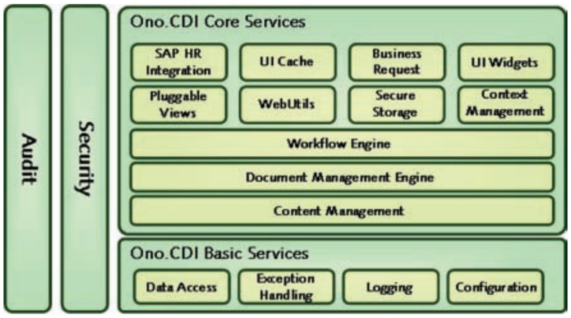
The following example illustrates the idea: we focused on creating client applications with very clear and crisp user interfaces to deliver good tools to all the employees in the company for going through the employee life-cycle processes. This eliminated the physical sheet of paper, the phone calls, and the ambiguities in the process, improving the operations of all the hiring managers and organizational unit managers in the company (and potentially of any employee in the company). However, in the first versions the back-end tools for the HR team where less sophisticated since we used all our resources for providing value to the biggest number of users.

Technology Planning: Leveraging In-House IT Assets

In order to create a reliable and fully functional tool with a low budget and fast delivery we leveraged Ono.CDI (Ono Content Driven Infrastructure Framework), our corporate software development framework for intranet based applications. Ono.CDI contains a set of tools and engines that provide core and basic services (Figure 2) that includes a Workflow Engine, Document Management Engine, Business Request Framework, and Caching Engine (Welicki 2006; Welicki & Sanjuan, 2007). It has crosscutting Security and Audit modules that ensure that applications complain with ONO's IT security policies and with legal audit requirements.

Ono.CDI also includes prescriptive architectural guidelines and blueprints that govern the architecture of ONO's intranet applications. Therefore the application is completely aligned with ONO's development policies lowering its technology transfer and maintenance cost (the maintenance and the development team are formed by different people) and ensuring a quality minimum based proven architectural and design patterns (Fowler, 2002; Gamma, Helm, Johnson

Figure 4. ONO's development framework high level architectural view



& Vlissides, 1995) and development practices.

Financial Planning: More with Less

One of the main goals of the project was delivering within a low budget but still with high level of functionality and quality. On top of our deliverables and technology planning, we created a financial plan to make our developments feasible on time and on schedule.

Our existing software and hardware corporate assets (Ono.CDI development framework and Ono's Intranet deployment infrastructure) set the tooling and deployment costs to the minimum (we where working on proven, reliable, and well-known-to-us components and deploying to a controlled and well known production environment). This also saved us from acquiring new licenses and/or buying new products, since we where building upon (reusing) our in-house IT assets (software and hardware).

The iterative incremental plan allowed us to get a tighter control on budget, having a revision stage at the end of each iteration. Therefore we didn't pay upfront for something that we couldn't plan in advance and made a more efficient use of money, spending smaller amounts in exactly what we need when we need it.

Unfortunately we cannot disclose economic information (prices of products and services) from our partners and supplier proposals. However, what we can certainly state is that we've built our system within a range of 12.5% and 27% (approximate) of the cost the proposals that we had received.

The HR Workflow System

The result of the process is the HR Workflow System that provides support for the employee life-cycle management processes. In a previous section we established a set of requirements that needed to be fulfilled by the system. Following, we will explain how these requirements have been met by the tool.

HR budget control. The system provides a detailed HR budget control. It has been designed and built to keep track of the budget of each organizational unit on a process instance basis. For example, when a manager requests HR to hire a new employee the system checks if his area is above or below its headcount budget. If the area is above its headcount budget the request is annotated with special information (as shown in Figure 5) in order to make this situation easily noticeable to the rest of the participants in the flow. The application also includes reports to provide the HR experts and the upper management with high level real-time views of the overall HR budget evolution in the organization.

Support for the approval workflow. The system used ONO's Workflow Engine capabilities to model the approval workflows. A workflow definition has been created for each of the flows associated with each type of employee life-cycle process (hiring, change of working conditions, and dismissal). Figure 3 (before in this paper) shows the flows for hiring and dismissing employees are shown. Figure 6 shows the flow for changing the conditions of an existing employee within the company (this flow is more complex than the other, with more participants and business rules).

Enabling collaborative work. The HR Workflow System notifies each employee of new events via e-mail whenever a process needs his partici-

Figure 5. Fragment of the "hiring request" screen. Notice how the headcount indicators ("HC presupuestado" and "HC real" fields) are highlighted in red. This means that this request is above the headcount of the requesting organizational unit

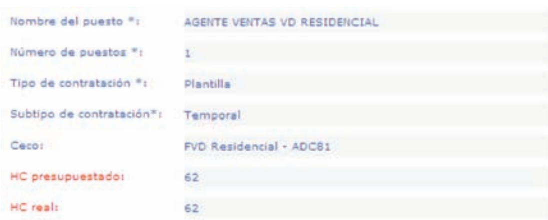
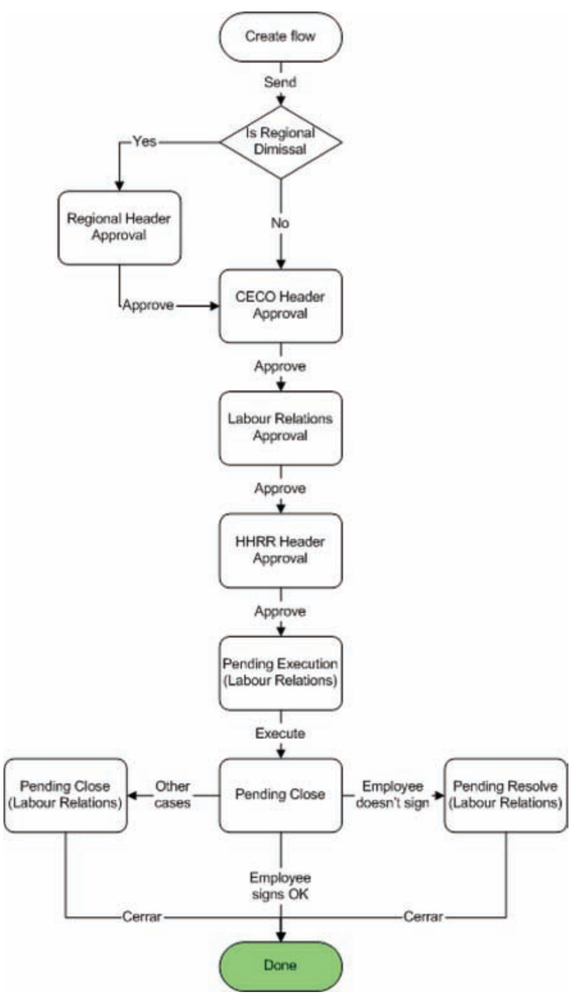


Figure 6. Change conditions of employee flow (promotions, change of salaries, change of departments,change of location, etc.)



pation. This enhances the users experience since the users don't need to be polling the application periodically to check if there is something that requires their attention. Additionally, the system implements an "Inbox" metaphor (similar to the one used in the e-mail systems) that gathers the requests that need participation of the user. When a user enters in the application his inbox is displayed. Additionally a summary of it (Figure 7) is present during all his session within the application.

Automatic reporting. The system generates automatically the necessary reports for HR control and for the upper management control. They are generated dynamically and can be requested at any moment providing an up-to-date picture of the overall active employee life-cycle processes in the organization.

Traceability and auditability. Every action in the approval workflow of the employee life-cycle process instances is recorded in a history log that is displayed within each process instance (Figure 8). Additionally, access information is recorded but not shown (this is used for privacy and access control audits)

Reducing manual work. All the notifications to participants, reporting, and archival of finished processes is done automatically, eliminating the most tedious and error prone manual tasks.

Input consistency. The input screens reduce the work to be done by the users. Each participant only needs to complete a very concrete (and ideally small amount of) information in each step of the flow. The employee and organizational related information is extracted directly from ONO's IT infrastructure, and the rest of the fields are parameterized lists (whenever is possible) simplifying considerably the creation of new requests (Figure 9).

Reduce bottlenecks in the processes. The application provides "delegation" functionality that allows an employee to "delegate" in another employee his functions. There are also "super-users" (HR members) that can act in any request at any time. Therefore, except in the case rare

Figure 7. Inbox summary. This widget is displayed in the left menu bar of the screen and is always visible. This widget shows the total number of requests of each type that requires the participation of the logged user. When the user clicks in any of the request types the complete inbox is displayed

Figure 8. History of a process. Each action performed on the request is recorded (including the execution user, date, target state, and observations)

Historia de la Solicitud

Acción	Usuario	Fecha	Estado	Observaciones
Crear	LEON EZEQUIEL WELICKI	09/07/2007 11:12:55	Pendiente Validar Responsable	
Enviar a Área	LEON EZEQUIEL WELICKI	09/07/2007 11:14:29	Pendiente Validar Director Área	
Enviar a RRHH	LEON EZEQUIEL WELICKI	09/07/2007 11:15:27	Pendiente Validar Director RRHH	
Aprobar	LEON EZEQUIEL WELICKI	09/07/2007 11:19:20	En Proceso	
Terminar	FRANCISCO JAVIER PIQUERES JUAN	09/07/2007 12:31:36	Selección Terminada	

Figure 9. Input screen for working conditions change. After an employee is selected (first field) all his information is retrieved from the corporate SAP HR database and displayed in read only fields. The remaining input fields in the form are parameterized drop down lists

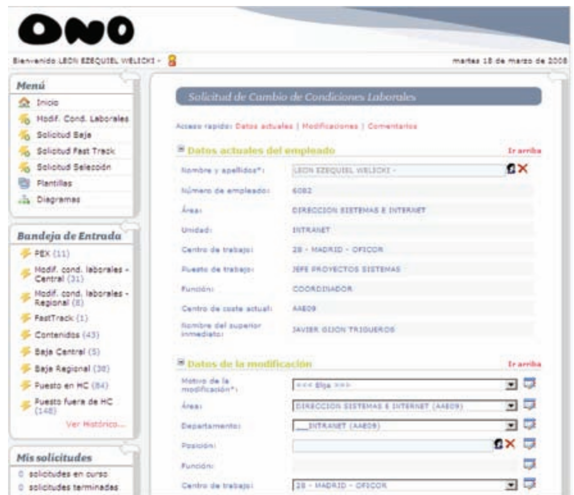


Figure 10. UI consistency and simplicity: the HR Workflow System (right side of the figure) is similar to the corporate intranet (left side of the figure)

situations that need a careful analysis of a senior manager, no employee is a bottleneck for a process instance in the system.

Ease of use and far reach. We based our user interface (UI) in our corporate Intranet (OnoNET) that is well known among all ONO employees (Figure 10). We also used all the UI widgets of our development framework having consistency with the existing applications in the intranet. The use of web-based technologies simplifies the deployment and reach of the tool: any user with a browser can access and use the HR Workflow System. Our UI foundation follows the usability guidelines presented in Nielsen (1999) and Krug (2000).

Achieving Strong Employee Life-Cycle Management

ONO HR Workflow System provides support for automated, formally defined, input consistent, predictable, and repeatable employee process life-cycle management. Additionally, all reporting is done automatically at run-time, eliminating manual consolidations.

Another very important outcome is that since we have automated the consolidation of information a single repository for employee corporate information can be easily established and feed with the results of the processes. This repository (shown in Figure 11 as "Employees Digital Identity Database") is an authoritative up-to-date source of information for all other e-HRM applications, avoiding the need of manual or semi formal small integrations, enhancing the operational quality and information of eHRM within the organization. As soon as a change in an employee's information or hierarchical structure is introduced in the system (through the execution of an instance of an employee life-cycle process) it is available to all applications without the need of performing any manual integration task. Moreover, employee and organizational concepts are normalized within all eHRM applications in the company allowing easier integrations (they all share the same employee and organizational abstractions).

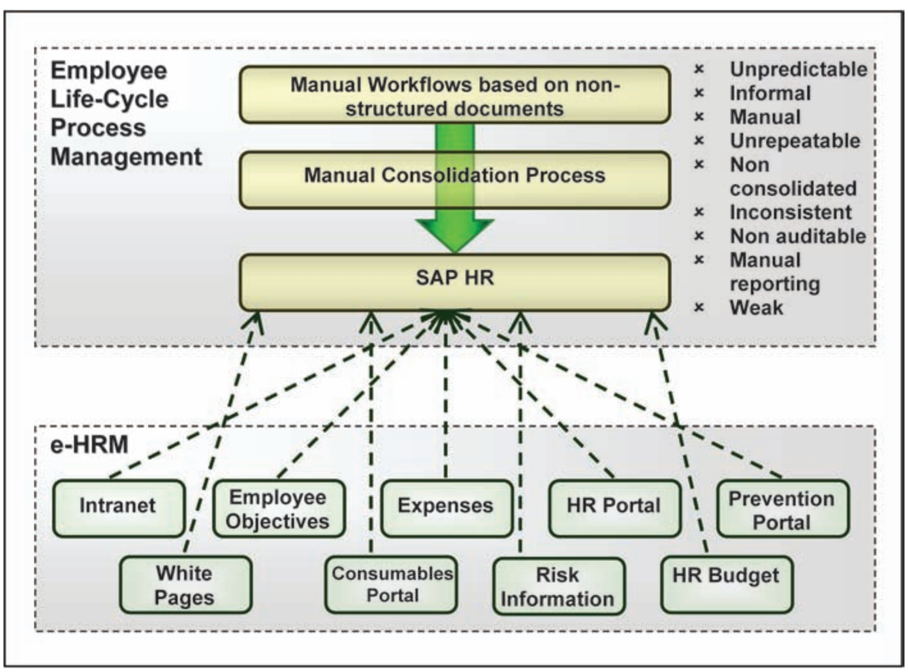
Ono HR Workflow System has all the necessary features to provide strong employee life-cycle process management (Figure 11).

Future Improvements

We improved and normalized the employee life-cycle support significantly, but our work isn't finished yet. Currently, we are working on the following enhancements to the system:

- **Dynamic HR budget with SAP Business Warehouse:** implement SAP Business Information Warehouse (SAP Corporation,

Figure 11. Ono's current "strong employee life-cycle process" model. The process instances go through a formal approval workflow in a web-based application. The results of the processes feed ONO's SAP HR and in return this feeds a centralized corporate information (identity) repository. Applications have a unique source of information for organizational data, reducing the integration and maintenance burden within the whole e-HRM ecosystem



2008) to provide dynamic and up-to-date information on the headcount budget. Currently the budget is loaded in the system upon HR's request. Using SAP's Business Intelligence suite can provide also analytical capabilities when looking at the future (e.g. forecasting budgets) and when looking at the past (e.g. analyzing real versus forecasted)

- **Improving the support for the selection process:** currently some parts of the selection process (for hiring new employees) are done offline. The system doesn't provide support for managing the interviews with candidates and managing offer letters (this two issues need very special attention of ONO's legal department according to Spanish laws on information protection).
- **Automatic creation and archival of formal letters:** Automatically create formal letters within each process and archive them in the process instance. This will eliminate the remaining paperwork associated with the process instances.

CONCLUSION

The workflow system has been in production for more than a year. It has successfully managed a big number of requests (we cannot disclose that information here), bringing reliability, traceability, and audatibility to the employee life-cycle management processes. It has become one of the core systems for supporting HR operations.

The resolution of each employee life-cycle process went down from weeks (in the manual case) to just days or hours (according the complexity of the request). Each participant in the process has to just perform a very concrete action and work with very concrete set of data. Additionally, he is notified via mail every time he needs to perform an action on a request.

Reporting is done automatically. The HR staff doesn't have to spend time on creating reports and the upper management has real-time in information on the on-going employee life-cycle processes.

Everyone in the company knows the processes (they are unambiguous, well documented, and accessible to everyone in the organization) and the processes are always the same for every employee in the organization. They are no longer dependent on the participants. At the same time the processes model the reality of our company and had been tailored to provide as much value as possible to it.

We could summarize the main benefits of the system in these three items:

1. **Reliable information:** The data sources are accurate based on normalized input and each process instance is auditable.
2. **Agile information:** The employees are notified only when they have to participate in a flow eliminating unnecessary and unproductive "polling" in an application.
3. **Improved information management:** The employee process life-cycle processes information is centralized, easy to access, and normalized, making it easy to know about the state of any on going or finished process in a uniform and simple way.

REFERENCES

Agile Alliance. (1999) Retrieved September 1, 2008, from, http://www.agilealliance.com/

Beck, K. (1999). *Extreme Programming: Embrace the Change*. USA: Addison-Wesley Professional.

Brooks, F. (1995). *The Mythical Man-Month: Essays on Software Engineering, Anniversary Edition*. Addison-Wesley Professional.

Corporation, S. A. P. (2008). Components & Tools of SAP NetWeaver: SAP NetWeaver Business Intellegence. Retrieved September 1, 2008, from http://www.sap.com/platform/netweaver/components/bi/index.epx

Davenport, T., & Prusak, L. (2000). *Working Knowledge: How Organizations Manage what they Know*. Harvard Business School Press.

Fowler, M. (2002) *Patterns of Enterprise Application Architecture*. Addison-Wesley Professional

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design Patterns: Elements of Reusable Object-Oriented Software*. Reading, MA: Addison-Wesley Professional.

Krug, S. (2000). *Don't Make Me Think!* USA: New Riders Press.

Martin, R. (2002). *Agile Software Development, Principles, Patterns, and Practices*. USA, Prentice Hall

Nielsen, J. (1999). *Designing Web Usability*. Peachpit Press.

Nonaka, I., & Takeuchi, H. (1995). *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. USA: Oxford University Press.

Ruël, H. J. M., Bondarouk, T. V., & Looise, J. C. (2004). E-HRM: innovation or irritation? An exploration of web-based human resource management in five large companies. *Management Review*, *15*, 3.

Schwaber, K., & Beedle, M. (2000). *Agile Software Development with SCRUM*. USA: Prentice Hall

Seely Brown, J., & Duguid, P. (2002). *The Social Life of Information*. Harvard Business School Press.

Welicki, L. (2006). The Configuration Data Caching Pattern. *Proceedings of the 13th Conference on Pattern Languages of Programs (PLoP 2006)*. Portland, Oregon, USA, October 2006.

Welicki, L., & Sanjuan Martinez, O. (2007). Improving Performance and Server Resource Usage with Page Fragment Caching in Distributed Web Servers. *Proceedings of International Workshop on Scalable Data Management Applications and Systems*. Las Vegas, Nevada, USA.

Workflow Management Coalition (1995). *The Workflow Reference Model*.

This work was previously published in Handbook of Research on E-Transformation and Human Resources Management Technologies: Organizational Outcomes and Challenges, edited by T. Bondarouk; H. Ruël; K. Guiderdoni-Jourdain; E. Oiry, pp. 335-350, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 6.4
Some Key Success Factors in Web-Based Corporate Training in Brazil:
A Multiple Case Study

Luiz Antonio Joia
Brazilian School of Public and Business Administration of Getulio Vargas Foundation and Rio de Janeiro State University, Brazil

Mário Figueiredo Costa
Brazilian School of Public and Business Administration of Getulio Vargas Foundation, Brazil

ABSTRACT

Brazilian companies are increasingly turning to Web-based corporate training by virtue of the fact that they need to train their employees within tight budget constraints in a country of continental dimensions. However, most of these companies do not know what the critical success factors in these endeavors are. Therefore, this article seeks to investigate some key success factors associated with such digital enterprises. In order to achieve this, the multiple case study method is used, whereby two cases, both conducted within the same Brazilian company, leading to opposite outcomes—a success and a failure—are analyzed in depth. The conclusions reached in this article were that goal orientation, source of motivation, and metacognitive support were the three critical dimensions in these two Web-based corporate training programs under analysis. Lastly, some managerial implications of these results are outlined.

INTRODUCTION

Nowadays, market dynamics are becoming increasingly intense due to new strategic orientations and the pressing need for businesses to adapt themselves to new business models and regulatory frameworks. For this reason, it is of paramount importance for companies to become agile, as well as achieve low costs and high returns on investment associated with their employee training programs. On the other hand, the increasing speed of obsolescence in training content, plus the high costs of face-to-face training programs, as well as the logistic hurdles linked with their deployment—mainly in firms operating in countries of continental dimensions (like Brazil)—are major barriers to the implementation of such face-to-face training programs.

Another aspect is that information technology (IT) is changing the way people search, locate, access, and retrieve available knowledge, as well as altering the learning process and the way training is conducted (Hodgins, 2000). While employees take charge of their own learning process and professional development, the employers face new challenges in training and retaining teams with in-depth knowledge about their business (Hodgins, 2000).

It is in this context of rapid change, with massive information loads and the search for training programs, that Web-based corporate distance training comes into its own. Information technology can solve most of the problems associated with the hitherto existing employee training undertakings, enabling the implementation of corporate distance training programs (Rosemberg, 2001).

Despite being a key factor for developing feasible training programs, information technology per se is not a guarantee of success for these endeavors. Most of the time, it must be linked to pedagogical and didactical issues related to them. The specific characteristics of each training program must be analyzed in depth and considered as relevant as the implementation costs throughout the decision-making process (Clark, 1983).

The structuring of Web-based training programs is no easy task, as according to several scholars, various critical success factors must be taken into consideration (see, for instance, Carey, Mitchell, Peerenboom, & Lytwyn, 1998; Penuel & Roschelle, 1999). In line with this, this article seeks to investigate what these critical factors are through the analysis of two distinct Web-based training programs conducted within the same Brazilian company. Hence, the research question in this article is: "What are the critical success factors associated with the implementation of these two Web-based corporate training programs?"

In order to achieve this goal, this work is structured as follows. First, the first section addresses the theoretical references used in this article. Then, the research method is outlined. After that, the two cases under analysis are described, and in the next section, the results accrued from them are compared. In the last section, the authors present some final comments.

Theoretical References

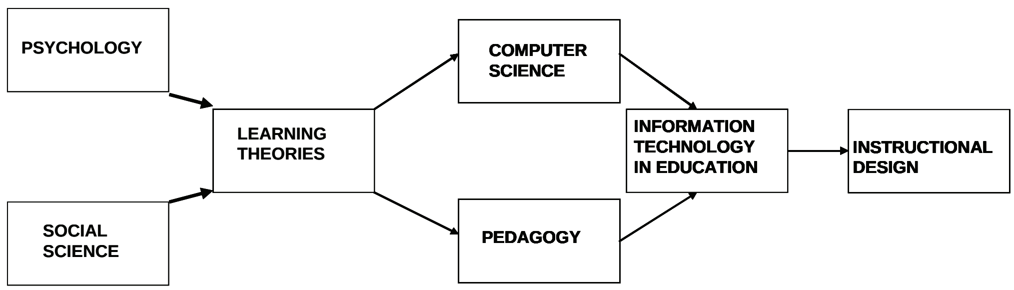
According to Wilhelmsem (2005) and Huitt and Hummel (2003), there are four knowledge fields associated with distance training, namely psychology, social science, pedagogy, and computer science. Figure 1 depicts how these four areas are interlinked, pointing to the crucial importance of social science and psychology—and learning theories—as the theoretical support for the areas of pedagogy and computer science, in order to enable the development of an instructional design aiming to apply information technology in education.

Behaviorism, cognitivism, and constructivism are all theories addressing the learning process, as well as the nature of knowledge and its main facets (Wilhelmsem, 2005). For behaviorists, knowledge is characterized as a passive process. Learning is explained without reference to the mental processes, as its focus is on observable behavior and in the way individuals adapt themselves to the environment. For cognitivists, the learning process molds the individual's mental construction. Finally, for constructivists, knowledge is seen as relative and socially built, varying according to time and space (Wilhelmsem, 2005).

Behaviorism

For behaviorists, psychology is the science of behavior, rather than the science of the mind. Behavior is correlated to external factors—the environment—instead of internal factors (Campos, 1982). The theory of classic conditioning, developed by Ivan Petrovich Pavlov (1849–1939), has a psychologi-

Figure 1. The creation process of instructional design in distance training (adapted from Wilhelmsem, 2005)



cal basis, as learning is developed via the linkage between situation, stimulus, and reaction. As the organism stores a repository of answers, trends, and reactions, it is able to act and produce an answer through an adequate natural and not conditioned stimulus (Campos, 1982).

Edward Lee Thorndike (1874-1949), in his connectionist learning theory, pointed out that the connection arises from the association between the impression of the senses and the impulses for action, namely from the linkage between the stimulus and the corresponding answer (Reinemeyer, 1999). Thorndike defined psychology as the science that studies the intellect, the behavior, and the characteristics of animals as well as human beings. Human education is concentrated on the emergence of certain changes in intellect, characteristics, and behavior in a composition based on four topics: objective, content, meaning, and method (Thorndike, 1911).

John B. Watson (1878–1958) presented a work in 1912 that is the basis for the development of behaviorism. He stated that humans and animals are complex machines that react to situations based on conditioned experiences, rather than on hereditary factors (Watson, 1929).

For Skinner (1904–1990), man is neutral and passive with behavior that can be described from a mechanistic standpoint. In his theory of operant conditioning, it is important that the stimulus follows the answer, which is named "reinforcement" by Skinner (Graham, 2005). Skinner stated that tools must be used to control the human learning process better. Thus, he recommended the use of programmed teaching, namely a process that allows the students to go further via a string of stages developed according to the student's own pace and reinforced immediately after each stage (Campos, 1982).

Cognitivism

In order to explain the development of cognitive processes, Piaget (1896–1980) revealed how organisms adapt themselves to their environments. This adaptation to the environment is controlled by the mental organization, namely schemes, which individuals use to represent a word and an established action. This adaptation is ruled by biological impulses in order to achieve equilibrium between the schemes and the environment (Huitt & Hummel, 2003).

Through his psychology of development of the intelligence, Piaget established a theory about the four stages whereby individuals—from birth to adulthood—acquire skills for logical reasoning. The rationale in this theory is that thinking is not the automatic outcome of reflections and intuitions, but rather a flexible operation developed by processes of trial and error (Piaget, 1964).

On the other hand, according to Lev Vygotsky (1896–1934), the impact of the external world on individuals' internal world, based on their interaction with reality, must be examined. Thus, social interaction is of paramount importance in a person's cognitive development. According to Vygotsky's principles, the origins of the changes that occur in these persons lie in their interaction with society, culture, and their particular story, that is, in the social theory of learning (Huitt & Hummel, 2003).

To the cognitivists, the thinking and previous knowledge of the students must be taken into account in the development of a course syllabus. Moreover, for them, the students move to new learning objectives in an increasing order of complexity, from the simpler to the more complex (Campos, Rocha, & Campos, 1998).

Constructivism

Studies in constructivist theory started with Jean Piaget, based on both an epistemological focus and an interdisciplinary perspective. His main question was: "How does one go from less developed knowledge to more developed knowledge?" A theory was then elaborated addressing the cognitive mechanisms of human beings (Huitt & Hummel, 2003).

For Piaget, intelligence is an active and organized assimilation process. When exposed to a challenge or stimulus, the individual experiences a disturbance in intellectual equilibrium, becomes curious, defied, and motivated. Then, through assimilation and accommodation, the individual recovers the former equilibrium, which is always dynamic as enabled by physical and/or mental actions. This equilibrium process is the central concept in Piaget's constructivist theory (Wilhemsen, 2005).

Lev Vygostky is another person who effectively contributed to building the conceptual basis of constructivism. He stressed the importance of language as a thinking tool, enabling the restructuring of several psychological functions, such as memory, attention, and the development of concepts. In this way, language acts decisively on the structure of thinking and is the basic tool for the construction of knowledge (Huitt, 2003).

Vygotsky (1935) also argued that knowledge is built from the action of individuals vis-à-vis reality and that individuals are not just active persons but also interactive and social beings, as they develop their knowledge, which is in turn developed by intra and interpersonal relationships. This process goes from a social dimension based on interpersonal links to an internal dimension based on intrapersonal links, such that the subjects effectively take part in the construction of their own culture and story, which changes them while also provoking changes in the individuals who interact with them.

Based on Piaget's research, Bruner (1985) argued that learning is an active process, whereby students build new ideas and concepts based on their own knowledge, select and transform information, develop hypotheses, and make decisions in order to establish their own cognitive structure (Huitt, 2003).

Teaching and Learning Theories

The constructivist approach requires that the aims and targets of training must be clearly defined, as well as the process for measuring the outcomes accrued from it (Mergel, 2005). On the other hand, a course based on the cognitivist model must consider the lessons accrued from the student's previous knowledge so as to achieve new learning objectives. This paradigm does not consider that the students will have the same prior experiences they once had, nor that they will learn in the same way they once did (Houser, 2005).

There are concepts that are common to both constructivism and cognitivism, since both are based on the assumption that new knowledge is built upon prior knowledge. For the constructivists, the student rather than the content, program format, and instructor is emphasized in a learning process. The latter is no longer the center of the learning process, as the students play this role (Wilson, Jonassen, & Cole, 1993).

In constructivism, knowledge is perceived as relative—nothing is absolute—varying according to time and space (Wilhelmsen, Stein, & Øyvind, 2005). According to Jonassen (1981), the cognitive tools are mental and computational devices that support, lead, and broaden the thinking process. In other words, the mind is in charge of knowledge acquisition via a linked cognitive process, and information plays the role of a stimulus that is perceived and recorded in the mind.

Instructional Design

Instructional design is the association of distinct learning theories with the development of pedagogical content, conveying to a certain training program. Its main objective is the application of learning theories in order to set up a concrete path that enables learning processes (Wilson et al., 1993). Rieber (1992) argues that there is no conflict between instructivism—based on behaviorism—and constructivism, such that training can incorporate features accrued from both paradigms without conflict. Rieber (1992) also states that the basic principles for the development of instructional design should include setting up the challenge to be proposed to the student correctly, making it neither too easy nor too difficult. It should also offer elements of surprise in order to arouse the curiosity of the student, as well as provide a context that intrinsically supports motivation and autoregulation of the learning process.

Malone (1999) argues that three characteristics increase the motivation and autoregulation of the student. The first is to provide a context whereby the students can enter into dialogue with their imagination and develop a personal state of fascination and intrigue. The second is to develop a context that arouses the students' curiosity, and the third is to set up a pattern that allows the students to travel from the "known to the unknown."

Assessment of Web-Based Corporate Training Programs

In many cases, the departments of a company need to develop corporate distance training programs via the Web. More often than not, these programs are oriented by technical imperatives, namely the obligation to use Internet technology. In some organizations, the Web-based training programs were designed specifically to justify the costs of the corporate intranet (Powell, 2000). However, the use of technology per se cannot be considered a justification for implementing any kind of training, as stated by Rosemberg (2001), Bregman and Jacobsen (2000), Bates (1995), and Kay (1970), to name a few.

In order to develop a comparative analysis between Web-based training programs, it is necessary to adopt a specific framework. In this article, the model proposed by Reeves and Reeves (1997) will be applied to identify and evaluate the distinct dimensions involved in Web-based training as explained later. This model has applications in the research, implementation, and evaluation of Web-based training programs such as those analyzed in this article.

It is important to stress that the model developed by Reeves and Reeves (1997) does not propose to evaluate either the outcome of a Web-based training program, or its success or failure. Indeed, the overriding purpose of this model is to assess the different aspects and facets of this kind of program (Reeves, 1997). The adopted model includes 10 dimensions of interactive learning on the World Wide Web, namely: (1) pedagogical philosophy, (2) learning theory, (3) goal orientation, (4) task orientation, (5) source of motivation, (6) teacher role, (7) metacognitive support, (8) collaborative learning, (9) cultural sensitivity, and (10) structural flexibility.

Each of the 10 dimensions in this model is presented as a two-ended continuum with contrasting values at either end. Needless to say, the world is rarely dichotomous, and there is more complexity involved in training than any of these dimensions suggest. However, the individual dimensions themselves are not as important as the interplay among the 10 dimensions that represent the instructional designs of various Web-based training programs. These dimensions are detailed below.

a. **Pedagogical Philosophy (Instructivist <=> Constructivist):** The debate over instructivist and constructivist approaches to teaching and learning persists to this day (Kafai & Resnick, 1996). Instructivists stress the importance of objectives that exist separately from the learner. Little emphasis is placed on learners themselves, who are viewed as passive recipients of instructions or treated as empty vessels to be filled with learning (Sherry, 1996). By contrast, constructivists emphasize the primacy of the learner's intentions, experience, and cognitive strategies. According to constructivists, learners construct different cognitive structures based upon their previous knowledge and what they experience in different learning environments. It is of paramount importance for constructivists that learning environments be as rich and diverse as possible. Instead of an empty vessel, the learner is regarded as an individual replete with pre-existing motivations, experiences, aptitudes, and knowledge. Tasks to be accomplished and problems to be solved must have personal relevance to the learner. The constructivists believe that what we know is constructed—both individually and socially—based on prior experience.

b. **Learning Theory (Behavioral <=> Cognitive):** According to behaviorists, the critical factor in learning is observable behavior, and instruction involves shaping desirable behavior through the arrangement of stimuli, responses, feedback, and reinforcement. A stimulus is provided (e.g., a short presentation of content), then a response is elicited—often via a question. Feedback is given as to the accuracy of the response, and positive reinforcement is given for accurate responses. Inaccurate responses result in a repetition of the original stimulus, and the cycle begins again. Cognitive psychologists place more emphasis on internal mental states than on behavior. Cognitive taxonomy of internal learning states includes simple propositions, schema, rules, skills, mental models, and so forth. They claim that a variety of strategies—including memorization, direct instruction, deduction, drill and practice, and induction—are required in any learning environment, depending upon the type of knowledge to be created by the learner.

c. **Goal Orientation (Sharp <=> Broad):** The goals for education and training can range from sharply focused goals to general higher order goals. Hence, the goal orientation of Web-based training systems varies in degree of focus from sharp to broad (Cole, 1992).

d. **Task Orientation (Academic <=> Authentic):** The context of learning is enormously important to adults (Giardina, Oubenaissa, & Bhattacharya, 2002; Merriam, 1993). Academic design depends heavily on having the learners carry out traditional academic exercises, whereas authentic design engages adults in practical activities such as preparing job applications, thereby situating practice and feedback within realistic scenarios. If knowledge, skills, and attitudes are learned in a practical context, they will be used in that context in similar situations.

e. **Source of Motivation (Extrinsic <=> Intrinsic):** Motivation is a primary factor in any theory or model of learning (Amabile, 1993). All new educational technology promises to be intrinsically motivating. This dimension ranges from extrinsic (i.e., outside the learning environment) to intrinsic (i.e., integral to the learning environment). Motivation instruction is intrinsically elusive, irrespective of the delivery system.

f. **Teacher Role (Didactic <=> Facilitative):** The teacher role continuum ranges from didactic to facilitative. In the former role, the teacher presents information and asks learners to memorize information and recall it later in tests. The latter role assigns cognitive responsibility to the learners, for them to be responsible for recognizing and judging patterns of information, organizing data, constructing alternative perspectives, and presenting new knowledge in meaningful ways with the teachers as tutors of this process.

g. **Metacognitive Support (Unsupported <=> Integrated):** Metacognition refers to a learner's awareness of objectives, ability to plan and evaluate learning strategies, and capacity to monitor progress and adjust learning behavior to accommodate needs (Flavell, 1979). The metacognitive support dimension is unsupported at one end of the continuum and integrated at the other. Recapitulation of the students' strategies at any point in the problem-solving process, as well as construction of Web-based portfolios (Nevado, Basso, & Menezes, 2004), are examples of how support for reflection, and metacognition might be provided in Web-based corporate training.

h. **Collaborative Learning Strategies (Unsupported <=> Integral):** The collaborative learning dimension ranges from a complete lack of support for collaboration to the inclusion of collaborative learning as an integral feature. Cooperative and collaborative learning refers to instructional methods in which learners work together in pairs or small groups to accomplish shared goals (Kirschner, Stribos, Karel Kreijns, & Beers, 2004).

i. **Cultural Sensitivity (Insensitive <=> Respectful):** All instructional systems have cultural implications. In an insensitive approach the training is developed irrespective of the culture and diversity of the learners it is intended to address. On the other hand, a respectful approach is based on the diversity in the populations in which the system will be used so that the overall learning environment is enhanced. It is unlikely that Web-based training can be designed to adapt to every cultural norm, but sites should be designed to be as culturally sensitive as possible (Brown & Voltz, 2005).

j. **Structural Flexibility (Fixed <=> Open):** "Fixed" systems, still dominant in education, are usually limited to specific places, for example, a classroom or laboratory, at specific times, for example, 50-minute class period. Irrespective of time and/or location constraints the learner can use "open" systems. The World Wide Web provides opportunities for more asynchronous (open) learning, although some Web-based learning tools are temporally fixed (synchronous), such as chats, video-conferences, and so forth.

Table 1 depicts the 10 dimensions defined for analyzing Web-based training programs, as supported by Reeves and Reeves (1997). For each dimension (in the central column of the table), the opposite poles of the adopted ratio scale (ranging from 0 to 10) are described and their meanings explained.

Research Method

The multiple case study method as described by Yin (1994) was adopted in this research, in which two Web-based distance-training programs developed within the same Brazilian company were analyzed in-depth. Case studies are particularly suitable for answering "how" and "why" questions, and are ideal for generating and building theory in an area where little data or theory exists (Yin, 1994), as in this knowledge field. It also enables researchers to use "controlled opportunism" to respond flexibly to new discoveries made while collecting new data

Table 1. Dimensions to evaluate the characteristics of Web-based distance training (adapted from Martin, 1998, and Joia, 2001)

0←	Dimension	→10
Instructivist Knowledge is imparted by the instructor	**Pedagogical Philosophy** 0—10	**Constructivist** Knowledge is constructed—both individually and socially—by the students
Behavioral Emphasis on observable behavior	**Learning Theory** 0–10	**Cognitive** Emphasis on internal mental states
Sharp Direct instruction focusing on desired behavior	**Goal Orientation** 0–10	**Broad** Simulations encompassing more than just a solution for the problem
Academic Emphasis on traditional academic exercises	**Task Orientation** 0–10	**Authentic** Emphasis on practical activities
Extrinsic Motivation lies outside the learning environment	**Source of Motivation** 0–10	**Intrinsic** Motivation lies in the student and the learning environment
Didactic The teacher is considered to be a knowledge repository	**Teacher Role** 0–10	**Facilitative** The teacher is a mentor and tutor for the students
Unsupported There are no student progress tracking mechanisms or adjustments to individual needs	**Metacognitive Support** 0–10	**Integrated** Student progress tracking mechanisms are implemented, as well as adjustments to individual needs
Unsupported Students work alone	**Collaborative Learning** 0–10	**Integrated** Students work together in pairs or in small groups
Insensitive Training is prepared regardless of the culture and diversity of the learners it seeks to address	**Cultural Sensitivity** 0–10	**Respectful** Training is based on the diversity of the populations where the system will be used
Fixed Program limited to specific places at specific times	**Structural Flexibility** 0–10	**Open** Program independent of time and/or location constraints

(Eisenhardt, 1994), as was done and is presented below in this work.

Notwithstanding having a major exploratory facet, this study also presents explanatory characteristics, as a causal relationship between the dimensions of the programs analyzed (Reeves & Reeves, 1997) and the respective outcomes are pursued. Yin (1994, p. 46) argues that in the multiple case study method, each case must be carefully selected, so as to generate either similar or opposing results. In line with this, a Brazilian company was chosen (the identity of which is confidential), and two Web-based training programs it developed and staged were selected, each one generating contrasting final results.

The first case—hereinafter referred to as Program A—was considered a success as it achieved its main objectives. The second case—hereinafter named Program B—developed by the same company, was considered a failure, as most of its targets were not accomplished. In order to comply with Yin's (1994) ideas necessary to validate this case study method, the following four issues were cautiously taken into consideration, namely: construction validity, internal validity, external validity, and reliability, as revealed below.

Construct Validity

In order to validate the "key success factors in Web-based corporate training" construct, multiple data sources were used, and also a chain of evidence related to research questions was pursued. The existing records associated with these projects were analyzed in-depth. The managers of both programs were located in the company and interviewed—there was a single manager for the first case (Program A) and two managers for the second case (Program B).

Questionnaires were circulated among the training users. These questionnaires sought to establish their perceptions relating to the 10 dimensions proposed by the Reeves and Reeves (1997) model. In addition to this, the users also revealed their perceptions about the rate of accomplishment of objectives of each program vis-à-vis the actual objectives proposed for the programs in their initial designs.

In line with the ideas proposed by Reeves and Reeves (1997), the minimum value of the scale (0) indicates that a dimension is fully aligned with the instructivist/behaviorist paradigm, whereas the maximum value of the same scale (10) proves that a dimension is fully aligned with the constructivist/cognitivist paradigm (Joia, 2001). Moreover, the maximum value of the scale (10) associated with the "accomplishment of training objectives" indicates user perception of complete success for the training program, whereas the minimum value (0) points to user perception of total failure for the training program. The aforementioned questionnaires were answered by 32 users of the first case analyzed (Program A) and 31 users of the second case (Program B).

Internal Validity

With a clear exploratory approach, this work addressed some explanatory elements used to verify the possible causal effects between the dimensions of the theoretical model and the training outcomes. This was done to support the internal validity of this research, in accordance with the recommendations of Morra and Friedlander (1999).

The first analysis conducted sought to compare user perceptions about the rate of accomplishment of objectives for the two programs, in order to verify whether or not the respective average of these grades could be considered statistically distinct. Once the difference between user perceptions regarding the rate of accomplishment of objectives for each program was recorded, a statistical comparison of user perception averages associated with each dimension of the theoretical model applied was performed. Since it had already been seen that the two programs presented statistical differences with respect to their outcomes, namely success and failure, the dimensions that did not present statistically significant differences within the two programs were discarded as not being critical success factors.

Thus, from this prior comparison, two dimensions of the Reeves and Reeves (1997) model were removed, leaving eight dimensions to be analyzed further. In order to achieve this, a multivariate linear regression was used, where the rate of accomplishment of training objectives was the dependent variable while the grades given by the users to each of the eight remaining dimensions of the model served as the independent variables. The significance level of each coefficient associated with these dimensions (independent variables) was then calculated and analyzed, while the dimensions whose coefficients did not present evidence of linear correlation with the dependent variable (accomplishment of objectives) were discarded.

The above procedure highlighted three dimensions, which could be considered critical success factors for the training programs analyzed.

Lastly, as a final quantitative validation, a simple linear regression was performed on each dimension removed from the study for not being related to the accomplishment of training objectives. The simple regressions supported that these factors did not possess a fair linear correlation with the objectives of the training programs.

External Validity

The external validity addresses whether or not the findings accrued from this research can be generalized for other similar cases not yet studied (Yin, 1994, p. 35). This work investigated the same factors related to two distinct cases developed by the same company so as to support the external validity of this research, enabling the results to be applied in other cases within the same firm analyzed.

Reliability

A protocol for documentation of the adopted procedures was developed to guarantee the reliability of this study. A digital data repository was also created to store all information gathered during the data collection stage. This repository stores the data set acquired during the field research for this investigation, as well as all the results accrued from the statistical analysis performed.

CASE DESCRIPTION

The Company

The company under analysis is a major Brazilian firm in the information technology industry. It has more than 30,000 employees with offices throughout Brazil.[1] In 2003, the company posted total revenue of US$865 million and net income of US$76 million. Due to its nationwide presence, this company faces an ongoing challenge to implement face-to-face corporate training programs, due to budget constraints. So, it is in this context that the two training programs, namely Program A and Program B were envisaged and implemented. The name of the company, as well as further details about it, are kept confidential, as agreed with its top executives.

Program A

Program A, considered a successful case by the company, is a mandatory corporate distance training program for all managers, namely its main target audience. Any employee who is promoted to a managerial function is obliged to take this course within a maximum timeframe of one year.

This training program lasts 9 months and consists of three distinct stages that encompass distance and face-to-face training. The focus of this program lies in the development of leadership skills. Accordingly, the following issues are addressed: the attributes that make an effective leader; the different kinds of leadership styles that are best used under certain conditions; the various theories of leadership practice and the pros and cons of each; and the leadership responsibilities related to administrative and management tasks.

The training program is based on the premise that, rather than being an isolated event, learning is a continuous process throughout the professional's lifetime. Program A uses several information technology tools, such as intranet that is heavily deployed to provide information considered essential for the managers of the company.

Stage I of this program (prelearning laboratory) is developed online in a distance-based training format. This stage lasts from 5 to 6 months and is an individual activity that demands between 48 and 56 hours of study.

Stage II of this program (learning laboratory) is a face-to-face experience lasting 5 days. The professionals must have successfully completed Stage I before embarking on this second stage. This learning laboratory takes place in the Global Learning Center of the company, in the city of São Paulo.

Stage III of this program (postlearning laboratory), like Stage I, is developed on a distance-training basis. This stage focuses on collaborative learning via the company's intranet, as well as public forums and tools like instant messaging.

Throughout the duration of the course, a mediator is previously assigned and available to take part in the program, both in person and online, in order to resolve any doubts the professionals may have, to supply the students with suggestions, and to help them solve general problems.

According to an interview with the manager of Program A, this program is considered a success, having fully achieved its targets. Furthermore, 32 users of Program A, who attended the program during 2005, answered the questionnaire developed for this research and evaluated their participation on this training program as a highly positive experience (average of 8.5 and standard deviation of 1.32 on a ratio scale ranging from 0 to 10). Therefore, it may be considered that the objectives were achieved. All of the 32 respondents were managers of the company.

Program B

Program B started at the beginning of 2004, initially as an effort to provide and make information about the company's productive and administrative processes available to employees located in the various offices of the company nationwide. The design and development of the program was organized by the company's IT team, supported by the basic premise of using the corporate intranet to publish all the content considered relevant.

The first version of the program gathered and consolidated the wealth of information about the company's processes already published in the intranet under a single site with a unique index for conducting searches. For this purpose, a team of five employees from two different business units was formed to assist the IT area in the identification and classification of information.

Once the information had been duly identified and classified, the IT area began to configure the program, so as to feature distinct courses categorized by subject. These courses could then be accessed by any employee via the intranet. Consequently, for each course implemented, a program manager was chosen to be in charge of developing the assessment questions (multiple-choice based), having privileged access to the answers given by the students.

After an initial test period—based on just one course developed for a specific group of employees—three distinct courses were made available—two of them focusing on specific working processes of the firm (order fulfillment and customer service), and the third addressing administrative content (employee performance assessment and promotion).

The main target of this training program was to reduce the costs involved in corporate training, as well as to speed up the adaptation and training time for newly hired professionals to become accustomed to the processes used by the organization. After less than one year, having failed to achieve its objectives, the program was redesigned.

Thirty-one users of Program B who attended the program during 2005 answered the questionnaire distributed by the researcher. In essence, they evaluated the experience of taking part in this program as negative since the aims were not achieved (average of 4.52 and standard deviation of 1.15 on a ratio scale ranging from 0 to 10). This evaluation from these employees was tallied with the opinion of the program managers, as they stressed that the objectives of this program were not achieved.

Comparison of Results

Initially, it is necessary to analyze the differences singled out by both the program managers and users concerning the achievement of objectives of the training programs. According to the assessment of the manager of Program A, the objectives of the training were fully achieved, and in his general evaluation, the program was rated as "very good." Conversely, the managers of Program B realized that the main targets of this program were not achieved, which led the program to be redesigned. Thus, according to the managers' perceptions, the difference related to achievement of objectives between the two programs becomes clear.

In order to analyze user perceptions related to the programs, it is necessary to evaluate the difference between the average grades given by the students to each one of the programs. The average user evaluation grade regarding the achievement of objectives in Program A was 8.5 (s=1.32; n=21, on a ratio scale of 0 to 10), whereas the same value concerning Program B was 4.52 (s=1.32; n=32, on a ratio scale of 0 to 10). This difference between the averages seems to tally with the opinion of the program managers. However, it is necessary to apply a statistical test (t-test) to compare the average of each program, so as to establish whether or not they can be considered different according to a statistical level of significance.

Table 2 depicts the results accrued from the comparison of employee evaluation averages related to the achievement of objectives of the training programs. From the results presented in Table 2, it is clear that there is a significant statistical difference between user perception averages related to the achievement of objectives of the training programs ($p < 5\%$). Furthermore, it can be observed that the interval of confidence does not encompass zero; that is, it is all positive. Thus, it is possible to support with a 5% level of significance that the averages are different and the average of Program A is greater than the average of Program B (Sincich, 1995, p. 532). It can be argued that with respect to "achievement of objectives," Program A achieved better results than Program B.

On the basis of this, the factors that influenced these results were researched, based on the theoretical model adopted in this article. Consequently, the evaluation averages of each dimension of the Reeves and Reeves' (1997) model were analyzed in order to find out which ones actually had an impact on the results depicted above. Similarly, the dimensions that presented statistical significant differences in the sample averages for each program were examined, as these are the dimensions that can be considered to be influential in the achievement of objectives of each Web-based corporate training program analyzed. Table 3 compares the averages related to each dimension of the programs under analysis, according to the framework of Reeves and Reeves (1997).

As can be seen in Table 3, there is no difference in the pedagogical philosophy and structural flexibility dimensions in the two cases, with a 5% level of statistical significance ($p>0.05$). Hence, these dimensions can be disregarded as critical success factors in Web-based corporate training. Based on this result, a multiple linear regression between the achievement of objectives (dependent variable) and the eight dimensions that presented significantly distinct averages (independent variables) was run. The intention was to verify which variables could be considered truly influential in the outcomes achieved. It is important to stress

Table 2. Comparison of averages related to "achievement of objectives" according to the users of the training programs

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig.	T	df	Sig. (2-tailed) (p)	Mean Difference	Std. Error Difference	Lower	Upper
Achievement of Objectives	.202	.655	12.752	61	.000	3.98	.31	3.36	4.61

that this regression seeks to verify the impact of each dimension on the outcomes of the programs under analysis, rather than to predict the outcomes of similar programs based on the dimensions of the model proposed by Reeves and Reeves (1997).

Table 4 depicts the summary of results and the statistical values accrued from this multiple regression. This summary supports the validity of using the eight dimensions of the theoretical model (predictors) to forecast the achievement of objectives for each case studied (in the summary, the "R" column represents the correlation coefficient and the "R Square" column represents the determination coefficient). From these data, it can be argued that nearly 70% (0.675) of the variance of the "achievement of objectives" variable can be explained by the dimensions included in this regression.

After validation of the model, an attempt was made to verify which coefficients, namely the dimensions of the model applied, actually influenced the achievement of objectives of Web-based training programs. Table 5 presents the summary of the statistics related to the coefficients of the regression model. From the results depicted in Table 5, it can be deduced that with a 5% level of significance, the learning theory, task orientation, teacher role, collaborative learning, and cultural sensitivity dimensions did not reveal evidence of any statistically significant linear relationship with "achievement of objectives" (Sig. > .05).

In order to strengthen the results accrued from this multiple linear regression, with respect to the lack of evidence of any linear relationship of the learning theory, task orientation, teacher role, collaborative learning, and cultural sensitivity variables, simple linear regressions of each of these variables vis-à-vis the "achievement of objectives" were performed. Table 6 presents the summary of the results accrued from these five simple regressions, which was drawn up separately from Table 5 to make it easier for the reader to fully understand the influence of each discarded dimension in the "achievement of objectives."

As can be observed from analysis of the correlation coefficient (column "R") and the determination coefficient (column "R Square") of the five simple regressions, these variables did not effectively have any bearing on the "achievement of objectives" variable ("R Square" smaller than 0.3).

Table 3. Comparison of the averages of the samples' dimensions of the model

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Pedagogical Philosophy	.010	.919	.511	61	.611	.11	.23	-.34	.56
Learning Theory	55.065	.000	2.470	61	.016	.52	.21	.09	.94
Goal Orientation	4.285	.043	6.239	61	.000	1.36	.22	.92	1.79
Task Orientation	16.813	.000	4.963	61	.000	1.03	.21	.61	1.44
Source of Motivation	8.686	.005	4.951	61	.000	1.15	.23	.68	1.61
Teacher Role	28.837	.000	6.790	61	.000	2.56	.38	1.81	3.31
Metacognitive Support	68.946	.000	9.747	61	.000	1.94	.20	1.54	2.33
Collaborative Learning	129.092	.000	3.760	61	.000	.78	.21	.37	1.20
Cultural Sensitivity	20.583	.000	7.756	61	.000	1.00	.13	.74	1.26
Structural Flexibility	.943	.335	-.751	61	.455	-.19	.26	-.71	.32

Table 4. Summary of the linear regression of the dimensions of the model

Model Summary (sample = 63 respondents; p-value=0.001)				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.847(a)	.717	.675	1.34
a Predictors: (Constant), Cultural Sensitivity, Learning Theory, Source of Motivation, Goal Orientation, Teacher Role, Task Orientation, Collaborative Learning, Metacognitive Support				

Table 5. Analysis of the statistical significance of the coefficients of the linear regression of the dimensions of the model

Coefficients									
	Unstandardized Coefficients		Standardized Coeficients	t	Sig.	95% Confidence Interval for B		Colinearity Statistics	
Model	B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
(Constant)	2.160	.547		3.950	.000	1.063	3.256		
Learning Theory	-4.589E-02	.230	-.017	-.200	.843	-.507	.415	.727	1.376
Goal Orientation	.486	.211	.226	2.299	.025	.062	.910	.541	1.849
Task Orientation	-.215	.256	-.088	-.839	.405	-.729	.299	.475	2.105
Source of Motivation	.845	.209	.388	4.046	.000	.426	1.263	.571	1.753
Teacher Role	.100	.124	.084	.805	.424	-.149	.349	.486	2.058
Metacognitive Support	.645	.228	.342	2.833	.006	.189	1.101	.359	2.785
Collaborative Learning	.108	.271	.042	.399	.691	-.436	.652	.478	2.090
Cultural Sensitivity	.288	.384	.087	.750	.457	-.481	1.057	.387	2.587
Dependent Variable: Achievement of Objectives									

Lastly, a final statistical analysis was performed. Analyzing the results of the multiple linear regression of the three variables selected as being influential in the achievement of objectives of the training programs—goal orientation, source of motivation, and metacognitive support—it can be seen that this model is very similar to the former multiple regression model (Table 4) which took eight variables into consideration. Table 7 portrays a summary of this model.

CONCLUSION, RESEARCH LIMITATIONS, AND FURTHER STUDIES

Concluding Remarks

Despite the fact that distance education has been around for over a century, the development of training programs has not achieved its full potential within organizations (Berge, 2002). Different technologies have been used since the creation of

Table 6. Summary of the models of simple linear regression of the variables discarded in the multiple linear regression

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1(a)	.291(a)	.085	.070	2.27
2(b)	.494(b)	.244	.232	2.06
3(c)	.524(c)	.275	.263	2.02
4(d)	.462(d)	.213	.200	2.11
5(e)	.514(e)	.265	.253	2.04
(a) Predictors: (Constant), Learning Theory (b) Predictors: (Constant), Collaborative Learning (c) Predictors: (Constant), Task Orientation (d) Predictors: (Constant), Teacher Role (e) Predictors: (Constant), Cultural Sensitivity				

Table 7. Summary of the linear regression of the "metacognitive support," "source of motivation," and "goal orientation" dimensions

Summary of the Regression (sample=63 respondents; p-value= 0.000)				
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
1	.838(a)	.703	.688	1.32
a Predictors: (Constant), Metacognitive Support, Source of Motivation, Goal Orientation				

the first distance training program, though Web technology is considered a watershed in this realm. While the technological progress has been impressive, the implementation of Web-based distance training has only increased at a slow pace (Cross, 2004). Besides the hype around Internet technology and its use in the business arena, the first trials using the Internet in corporate training arose at the end of the 20th century. However, most of these initial applications either failed or fell short of the expected outcomes (Cross, 2004).

For over a century, society has been trying to understand precisely how human beings learn. As with most problems in the social sciences, there is no single answer. However, it is clear that some rationale behind this research question must be developed. It must be remembered that Western society (mainly the USA) has been heavily influenced by the instructivist/behaviorist paradigm, upon which its educational system was designed (Criswell, 2000).

On the basis of theoretical references and case research analysis, it became clear that the deployment of Web-based training programs is not merely a technological issue. As in any training program, the inherent objectives and characteristics that it is seeking to achieve must be analyzed by the designers, so as to permit selection of the most adequate learning theory and define the instructional design, as well as develop and deploy the training program adequately.

Based on the comparison of averages, it was concluded with 5% level of statistical significance that there was no difference between the pedagogical philosophy and structural flexibility dimensions in the two cases analyzed. The sample averages of the former dimension (1.96 for Program A and 1.85 for Program B) indicate that both programs were highly instructivist, namely most of the knowledge is imparted by the training, rather than constructed by the students themselves. In other words, most of the learners' prior experiences were not taken into consideration in either case. This tallies with some authors who reveal the hurdles in developing a constructivist Web-based corporate training program in an environment where efficiency is pursued in order to be attained in a short time frame (see, for instance, Criswell, 2000; Joia, 2001; Joia & Casado, 2007). Likewise, the sample averages of the latter dimension (2.69 for Program A and 2.88 for Program B) pointed to the fact that "fixed" training programs are still dominant in corporate training, as in neither of the programs could the learners use the systems irrespective of time and/or location.

Thereafter, applying a linear multiple regression between the dimensions of the model developed by Reeves and Reeves (1997) and the achievement of objectives of both training programs, it can be seen that five out of the eight remaining dimensions of the theoretical model did not have a significant influence on the results of either program. Actually, the dimensions that effectively had a major impact on the outcomes of training Programs A and B were goal orientation, source of motivation and metacognitive support.

The low averages observed for the goal orientation dimension (2.94 for Program A and 1.58 for Program B) indicate that the objectives of both programs were more specific than generic. However, it is important to note that Program A aimed at achieving somewhat higher-order goals (namely leadership skills) than Program B. Conversely, Program B set out to address sharply focused goals (namely the firm's processes). In other words, with respect to this dimension, Program A was less instructivist/behaviorist than Program B.

This result duly corroborates the ideas of several authors who argue the need for a broader orientation for the success of a distance training program, that is, one that elicits more than the mere solution of specific problems (see, for instance, Dick & Carey, 1996; Kay et al., 1970; Mager, 1972; Sancho, 1998, to name just a few).

Program B—with an average of 1.06—had hardly any metacognitive support, whereas Program A—with an average of 3.00—revealed a certain level of implementation of this dimension. Once again, based on data collected from informal interviews, the users of Program B declared that there was no tool for students to track their progression during this training program. Moreover, regarding metacognitive support, the actual description of the features available in Program B to students, from the program managers' perspective, namely access via the intranet and multiple choice questionnaires, reveals and supports the lack of means for users to assess their learning strategies in a timely manner.

On the other hand, Program A did indeed provide some opportunities for students to develop the kind of assessment addressed above. The tool upon which this program was built allowed the users to track their outcomes at each stage of training, as well as the percentage of total time available to complete the course, and the estimated total time necessary to accomplish each stage of the program. Furthermore, Program A allowed the students to check back on content they had already studied on the course, thereby enabling them to control their learning process, as suggested, for instance, by Nevado et al. (2004), Campbell, Strijbos, Karel Kreijns, and Beers (2000), and Costa, Fagundes, and Nevado (1998).

Lastly, Program B users' assessment concerning the source of motivation dimension produced an average of 1.26, indicating that the source of motivation was mostly extrinsic. On the other hand, in Program A (average of 2.41), it becomes clear that there was at least some prior intrinsic source of motivation during the training program per se, probably due to the fact that these employees had just been promoted to managers. Thus, it can be considered that more than being motivated by the course, the students were supposed to be motivated by the company and their careers—a claim supported by interviews developed with five users of Program A. Conversely, the users of Program B did not appear to be motivated to take part in the training program, except for external motivation based on the mandatory nature of the program. Interestingly, this result complies with the ideas of Carroll (1968), Amabile (1993), and Keller and Suzuki (2004) about the importance of taking intrinsic motivation into account in any pedagogical model.

Hence, from the comparison of the two cases, the following items can be considered key success factors in these Web-based training programs:

- Clear definition of training content, target employees and objectives of the program, seeking more than merely the solution of specific problems;
- Development of a source of intrinsic, as opposed to extrinsic, motivation;
- Implementation of Web-based metacognitive support.

Figure 2 depicts the inter-relationship and influence of the critical success factors in the development of Web-based distance training programs in a concise way. As suggested in Figure 2, the three key success factors accrued from the analysis of the results of this research vis-à-vis the theoretical background enable the selection of the learning theory, the instructional design and the technologies to be used in this endeavor. It is also important to highlight that there is a feedback process in the model of Figure 2, namely the factors interact among themselves during the life cycle of the training program.

It is interesting to note that according to Ertmer and Newby (1993) and Conole, Dyke, Oliver, and Seale (2004), the selection of a specific learning theory is not a key success factor by itself. Moreover, the realization that this dimension did not directly influence the outcomes accrued from selected programs A and B (as both presented instructivist/behaviorist characteristics) complies with Reeves' (1997) frame, as it does not support the allegation that an instructivist/behaviorist program is necessarily better than a constructivist/cognitivist one and vice-versa.

However, this is a point that must be the subject of in-depth investigation in future research addressing training in virtual environments. Program A presented a more constructivist/cognitivist approach than Program B, as witnessed by the fact that the averages of the three relevant dimensions in the former program were higher than the corresponding dimensions in the latter program. This tallies with some authors who have argued that the constructivist/cognitivist approach is best suited for Web-based distance training (see, for instance, Costa et al., 1998).

Research Limitations

As with all research, this project has a few limitations now set forth. First of all, the number of respondents—32 users of Program A and 31 users of Program B—led to a sample size limitation, preventing the authors from running one multiple linear regression for each training program. According to Hair, Anderson, Tatham, and Black (1998, p. 166), there should be at least five observations for each independent variable. As there were eight remaining variables, a sample of at least 40 respondents for each training program was required. Accordingly, a linear multiple regression adding a dummy variable for Program A and Program B had to be run. The outcomes of this latter regression have shown the difference between the degree of accomplishment of objectives of either program (Hair et al., 1998, pp. 167–168). Moreover, as programs A and B are not exactly equal, some other factors associated with their corresponding content and modus operandi, just to name two aspects, can also have had an influence on their respective outcomes. For instance, in the second program, the fact that the IT team designed much of the content, a task that should more appropriately have been performed by domain knowledge specialists, may also have affected the outcome.

Besides, there may have been motivational differences between the participants of the two programs, as well as differences regarding the participants' IT literacy, since the degree of general familiarity with these technologies may be an important factor in determining success or otherwise. Furthermore, this article attempted to establish the value perceptions of the employees regarding the outcomes of the two Web-based training programs analyzed. There are some limitations in this approach, as some of the variables derived from the Reeves and Reeves (1999) model are not such simple variables as to be clearly understood by the respondents beyond all reasonable doubt, even after various meetings with the author. Indeed, a certain degree of subjectivity and bias from the employees may have occurred (Scandura & Williams, 2000).

Lastly, this is not a cross-cultural research project. Therefore the aspect of whether or not there is any influence accruing from the Brazilian setting in the outcomes of this research is not analyzed. The reason for this lies in the very fact that there are as yet very few works about Web-based corporate training in Brazil in existence. In order that one can develop cross-cultural studies, it is important to have information about what is supposed to be compared.

Further Studies

This article naturally does not claim to be the ultimate research in this knowledge field. The subject deserves a great deal more study and investigation. Research involving a larger number of companies and focusing on each specific dimension involved in the development of Web-based distance training programs might reveal other important issues related to this realm, in order to allow the organizations to better understand, improve and measure the outcomes of these endeavors.

Lastly, as has been said earlier, it is important to fully understand whether or not the Brazilian environment influenced the outcomes presented. Moreover, it is also important to verify whether there are differences between Web-based corporate training programs conducted in developing countries (such as Brazil) and developed countries. Thus, there is still much ground to be covered in this area.

REFERENCES

Amabile, T. (1993, Fall). Motivating creativity in organizations. *California Management Review*, *40*(1), 39–58.

Bates, A. W. (1995). *Technology, open learning and distance education*. London/New York: Routledge.

Berge, Z. L. (2002). Obstacles to distance training and education in corporate organizations. *Journal of Workplace Learning*, *4*(5), 182–189.

Bregman, P., & Jacobson, H. (2000). Searching for answers: Yes, you can measure the business results of training. *Training*, *38*(8), 68–72.

Brown, A. R., & Voltz, B. D. (2005). Elements of effective e-learning design. *The International Review of Research in Open and Distance Learning*, *5*(1). Retrieved June 24, 2008, from http://www.irrodl.org/index.php/irrodl/article/viewArticle/217/300

Bruner, J. (1985). Models of the learner. *Educational Researcher*, *14*(6), 5–8.

Campbell, D. M., Melenyzer, B. J., Nettles, D. H., & Wyman, R. M., Jr. (2000). *Portfolio and performance assessment in teacher education*. Allyn & Bacon.

Campos, D. S. (1982). *Psicologia da aprendizagem*. Editora Vozes Ltda, Petrópolis, RJ, Brazil

Campos, F. C. A., Rocha, A. R. C., & Campos, G. H. B. (1998). *Design instrucional e construtivismo: Em busca de modelos para o desenvolvimento de software*. Retrieved June 24, 2008, from http://www.niee.ufrgs.br/ribie98/TRABALHOS/250M.pdf

Carey, T., Mitchell, S., Peerenboom, D., & Lytwyn, M. (1998). *Effectiveness of learning technologies: The costs and effectiveness of technology-based approaches to teaching and learning.* University of Guelph.

Carroll, J. B. (1968). On learning from being told. *Educational Psychologist*, *5*, 4–10.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research*, *53*(4), 445–459.

Cole, P. (1992). Constructivism revisited: A search for common ground. *Educational Technology, 32*(2), 27–34.

Conole, G., Dyke, M., Oliver, M., & Seale, J. (2004). Mapping pedagogy and tools for effective learning design. *Computers & Education*, *43*(1–2), 17–33.

Costa, I. T., Fagundes, L. C., & Nevado, R. A. (1998). Projeto TEC-LEC Modelo de uma Nova Metodologia em EAD Incorporando os Recursos da Telemática. *Proceedings of the Informática na Educação: Teoria e Prática, 1*(1), 83–100. Porto Alegre.

Criswell, E. (2000). The humanistic tradition: A vision for the future. *Journal of Humanistic Psychology, 40*(3), 74–82.

Cross, J. (2004). An informal history of eLearning. *On the Horizon*, *12*(3), 103–110.

Dick, W., & Carey, L. (1996). *The systematic design of instruction*. New York: HarperCollins Publisher.

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, *14*(4), 532–550.

Ertmer, P. A., & Newby, T. J. (1993). *Behaviorism, cognitivism, constructivism: Comparing critical features from a design perspective*. Performance Improvement Quarterly, 6(4), 50–72.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American.

Giardina, M., Oubenaissa, L., & Bhattacharya, M. (2002). Designing a framework for the implementation of situated online, collaborative, problem-based activity: Operating within a local and multi-cultural learning context. *International Journal on E-Learning*, *1*(3), 41–46. Norfolk, VA: AACE.

Graham, G. (2005). Behaviorism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Fall 2002 ed.).* Retrieved June 24, 2008, from http://plato.stanford.edu/entries/behaviorism/#1

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis (5th ed.).* Prentice Hall.

Hauser, L. (2005). *Behaviorism*. *In The Internet encyclopedia of philosophy.* Retrieved June 24, 2008, from http://www.utm.edu.research/iep/b/behavior.htm

Hodgins, H. W. (2000). *Into the future: A vision paper.* Commission on Technology and Adult Learning. Retrieved June 24, 2008, from http://www.learnativity.com/download/MP7.pdf

Huitt, W. (2003). *Constructivism*. *Educational Psychology Interactive*. Retrieved June 24, 2008, from http://chiron.valdosta.edu/whuitt/col/cogsys/construct.html

Huitt, W., & Hummel, J. (1998). *An overview to the behavioral perspective*. Retrieved June 24, 2008, from http://chiron.valdosta.edu/whuitt/col/behsys.html

Joia, L. A. (2001). Evaluation of hybrid socio-constructivist model for teacher training. *Journal of Technology and Teacher Education*, *9*(4), 519–549.

Joia, L. A., & Casado, N. (2007, September 23–26). Fatores criticos de sucesso em treinamentos corporativos a distância via Web: Evidências empírico-exploratóriasa partir de um estudo de caso. In Proceedings of the 31st EnANPAD (Brazilian Academy of Management Meeting), Rio de Janeiro.

Jonassen, D. H. (1981). *What are cognitive tools?* University of Colorado, U.S.A. Retrieved June 24, 2008, from http://www.cs.umu.se/kurser/TDBC12/HT99/Jonassen.html

Kafai, Y., & Resnick, M. (Eds.). (1996). *Constructionism in practice: Designing, thinking, and learning in a digital world*. Lawrence Erlbaum Associates.

Kay, H., Dodd, B., & Sime, M. (1970). Iniciação à instrução programada e às máquinas de ensinar. São Paulo: IBRASA.

Keller, J., & Suzuki, K. (2004). Learner motivation and e-learning design: A multinationally validated process. *Learning, Media and Technology, 29*(3), 229–239.

Kirschner, P., Strijbos, J. W., Karel Kreijns, K., & Beers, P. J. (2004). Designing electronic collaborative learning environments. *Educational Technology Research and Development, 52*(3), 47–66.

Mager, R. F. (1972). Objetivos para o ensino efetivo. Rio de Janeiro: SENAI.

Malone, J. C., Jr. (1999, January). Operants were never "emitted." Feeling is doing, and learning takes only one trial: A review of B. F. Skinner's recent issues in the analysis of behavior. *Journal of the Experimental Analysis of Behavior, 1*, 115–120. University of Tennessee.

Martin, K. (1998). WBI or not WBI? *Issues of Teaching and Learning, 4*(7). Retrieved June 24, 2008, from http://www.catl.uwa.edu.au/NEWSLETTER/issue0798/dimensions.html

Mergel, B. (1998). *Instructional design and learning theory*. University of Saskatchewan. Retrieved June 24, 2008, from http://www.usask.ca/education/coursework/802papers/mergel/brenda.htm

Merriam, S. B. (1993). *Adult learning: Where have we come from? Where are we headed?* In An update on adult learning theory (pp. 5–14). San Francisco: Jossey-Bass.

Morra, L., & Friedlander, A. C. (1999, May). Case study evaluations (Working Paper Series No. 2). OED (Operations Evaluation Department), World Bank.

Nevado, R. A., Basso, M. V. A., & Menezes, C. S. (2004). Webfólio: Uma proposta para avaliação na aprendizagem: Conceitos, Estudos de Casos e Suporte Computacional. In Anais do Simpósio Brasileiro de Informática na Educação, Manaus-AM, Brazil.

Penuel, B., & Roschelle, J. (1999). *Designing learning: Cognitive science principles for the innovative organization*. Center for Technology in Learning, SRI International.

Piaget, J. (1964). *The construction of the reality in the child*. Retrieved June 24, 2008, from http://www2.cddc.vt.edu/marxists/reference/subject/philosophy/works/fr/piaget2.htm

Powell, G. C. (2000). *Are you ready for WBT?* Wayne State University. Retrieved June 24, 2008, from http://it.coe.uga.edu/itforum/paper39/paper39.html

Reeves, T. C. (1997). *A model of the effective dimensions of interactive learning on the World Wide Web*. University of Georgia. Retrieved June 24, 2008, from http://it.coe.uga.edu/~treeves/WebPaper.pdf

Reeves, T., & Reeves, M. (1997). *Effective dimensions of interactive learning on the World Wide Web*. In B. Khan (Ed.), Web-based instruction (pp. 59–66). Englewood Cliffs, NJ: Educational Technology Publications.

Reinemeyer, E. (1999). Edward Lee Thorndike. *Psychology History*. Retrieved June 24, 2008, from http://carbon.cudenver.edu/~mryder/itc_data/idmodels.html

Rieber, L. P. (1992). *Computer-based microwords: A bridge between constructivism and direct instruction*. Educational Technology, Research, and Development. Retrieved June 24, 2008, from http://www.sll.stanford.edu/projects/Wallenberg/ateam/abstracts/Rieber.html

Rosemberg, M. J. (2001). *E-learning: Strategies for delivering knowledge in the digital age*. New York: McGraw-Hill.

Sancho, J. M. (1998). *Para uma tecnologia educacional* (B. A. Neves, Trans.). Porto Alegre: ArtMed.

Sherry, L. (1996). Issues in distance learning. *International Journal of Distance Education, 1*(4), 337–365.

Sincich, T. (1995). *Business statistics by exampl*e. Prentice Hall.

Thorndike, E. L. (1911). *Animal intelligence*. Macmillan.

Vygotsky, L. (1935). *Thougth and languages*. Retrieved June 24, 2008, from http://marxists.org/archive/vygotsky/works/words/vygotsky.htm

Watson, J. B. (1929). *Behaviorism: The modern note in psychology*. Retrieved June 24, 2008, from http://psychclassics.yorku.ca/Watson/Battle/watson.htm

Wilhelmsen, S., Stein, I. Å., & Øyvind, M. (1998). Psychological theories: *A brief survey of the changing views of learning*. Retrieved June 24, 2008, from http://www.uib.no/People/sinia/CSCL/Web_struktur-4.htm

Wilson, B., Jonassen, D., & Cole, P. (1993). *Cognitive approaches to instructional design*. University of Colorado at Denver. Retrieved June 24, 2008, from http://carbon.cudenver.edu/~bwilson/training.html

Yin, R. (1994). *Case study research: Design and methods (2nd ed.)*. Sage Publications Inc.

ENDNOTE

[1] Brazil is the largest country in South America and the fourth largest in the world in coterminous area, ranking after Russia, Canada, and China (the U.S. is larger with Alaska, Hawaii, and the dependencies included). Occupying nearly half of the South American continent, it covers an area of 8,511,965 sq km (3,286,488 sq mi), extending 4,395 km (2,731 mi) N–S and 4,320 km (2,684 mi) E–W.

This work was previously published in International Journal of Web-Based Learning and Teaching Technologies, Vol. 3, Issue 4, edited by L. Esnault, pp. 1-28, copyright 2008 by IGI Publishing (an imprint of IGI Global).

Chapter 6.5
Multi-Tier Framework for Management of Web Services' Quality

Abdelghani Benharref
Concordia University, Canada

Mohamed Adel Serhani
United Arab Emirates University, UAE

Mohamed Salem
University of Wollongong, Dubai, UAE

Rachida Dssouli
Concordia University, Canada

ABSTRACT

Web services are a new breed of applications that endorse large support from main vendors from industry as well as academia. As the Web services paradigm becomes more mature, its management is crucial to its adoption and success. Existing approaches are often limited to the platforms under which management features are provided. In this chapter, we propose an approach to provide a unique central console for management of both functional and nonfunctional aspects of Web services. In fact, we aim at the development of a framework to provide management features to providers and clients by supporting management activities all along the lifecycle. The framework allows/forces providers to consider management activities while developing their Web services. It allows clients to select appropriate Web services using different criteria (e.g., name, quality, etc.). Clients also make use of the framework to check if the Web services they are actually using or planning to use are behaving correctly. We evaluate the Web services management features of our framework using a composite Web service.

DOI: 10.4018/978-1-60566-042-4.ch002

INTRODUCTION

Web services standard is a recent paradigm of emerging Web components. It combines a set of technologies, protocols, and languages to allow automatic communication between Web applications through the Internet. A Web service is any application that exposes its functionalities through an interface description and makes it publicly available for use by other programs. Web services can be accessed using different protocols, different component models, and running on different operating systems. They usually use hypertext transfer protocol (HTTP) (W3C, 1999) as a fundamental communication protocol, which carries exchanged messages between Web services and their clients. Web services use extensible markup language (XML)-based (W3C, 2006) messaging as a fundamental means of data communication.

Research on Web services has focused more on interfacing issues, that is, simple object access protocol (SOAP) (W3C, 2004), Web services description language (WSDL) (WSDL, 2001), and universal description, discovery, and integration (UDDI) (OASIS, 2005). Until recently, considerable efforts have been conducted to address the issues of management of Web services in service-oriented architecture (SOA).

Web services management is among the hot issues that are not yet mature. Ongoing research from academia and industry are still emerging. Management of Web services is critical for their success because they are being actually used in a wide range of applications, ranging from entertainment, finance, and healthcare to real-time critical applications. Management issues in Web service can be divided into two dimensions: (1) management of functional aspects, namely fault management, and (2) management of nonfunctional aspects such as quality of service (QoS). Quality of a Web service, referred to as QoWS in this chapter, reflects the quality of a Web service, both in terms of correctness of functional behaviour and level of supported QoS. A Web service supporting QoWS is said to be QoWS-aware.

Nowadays, management of Web services is highly platform-dependent which implies the following limitations: (1) management features are usually available to Web services providers but often not to other partners (e.g., clients, third parties); (2) management solutions are usually restricted to only one management aspect, functional or nonfunctional; and (3) most of management solutions require considerable amount of computer and network resources to be deployed and used.

The first limitation restricts the utilization of management information to providers who are using it to assess the QoWS of their Web services. However, other entities involved in SOA industry might need to use this information as well. Clients can use this information during discovery and selection of Web services so they can figure out those with desirable QoWS. Moreover, many providers are likely to offer Web services providing similar functionalities but with quite different QoWS. In such a competitive market, attraction and loyalty of clients are primarily based on high standards of provided QoWS.

In SOA, a significant amount of work is taking place to allow both Web services providers and their clients to define and concisely use QoWS during publication, discovery, and invocation of Web services. For example, to select from a set of potential Web services, the one which is mostly available, and has a low response time and/or an acceptable utilization fee is preferable.

This chapter presents our approach for management of Web services. This approach provides a unique central environment for management of both functional and nonfunctional aspects of Web services. In fact, we aim at the development of a framework to provide management features to Web services providers and clients by supporting management activities all along the lifecycle of a Web service, from specification to invocation. The framework allows/forces providers to consider management activities while developing their Web services. In fact, the provider should concisely and precisely describe QoWS factors during design and implementation of the Web service. These factors will/shall be used latter by clients to select appropriate Web services during the discovery and selection operations. Clients also make use of the framework to check if the Web services they are actually using or planning to use are behaving correctly in terms of functional and nonfunctional facets.

The concepts presented all along this chapter will be illustrated in details through a case study. A Web service example will be used to show how different phases of the Web service development lifecycle must be conducted while promoting good practices for advanced management activities. At each phase, information and documents required by the management framework are produced and their impact on management is thoroughly discussed.

This book chapter is organized as follows. The next section provides background information required for nonexpert readers to follow the flow of ideas in following sections. The following section discusses related work in management of Web services and their limitations. A composite Web service used to illustrate our management framework is introduced. The subsequent section details, in a step-by-step tactic, how different management activities proposed by our framework can be conducted at each phase during the development of a Web service. We show then promising experimental results while using the framework to manage the Web service introduced hereafter. We close the chapter by presenting conclusions and insights for ongoing and future work.

BACKGROUND

Web services are a new variant of Web applications. It is a new paradigm which allows different applications to communicate automatically with each other over the Internet. They are self-contained, self-describing, modular applications that can be published, located, and invoked across the Internet (Wahli, 2002). The endeavor of this new paradigm is to allow applications to be delivered over the Internet and to run across all kinds of computers and platforms.

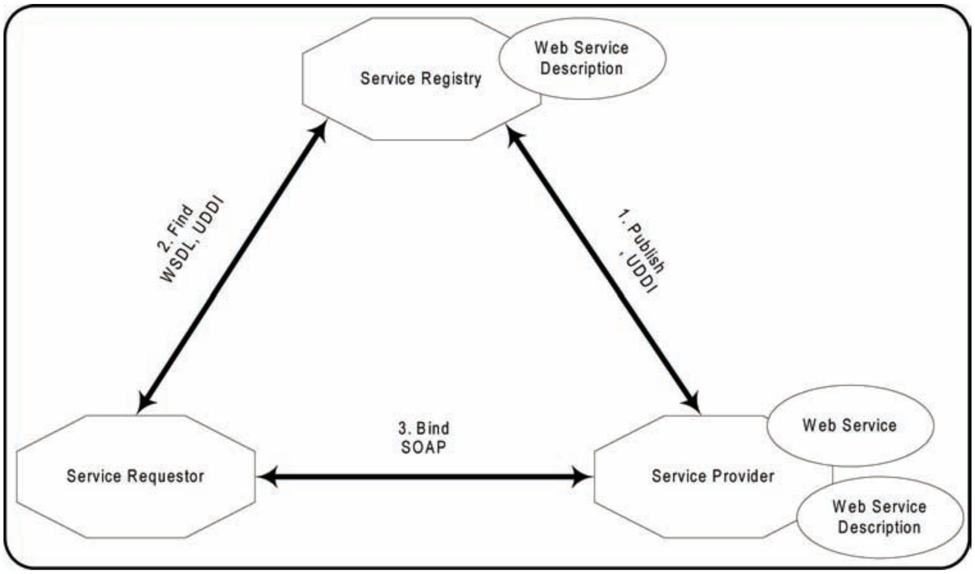
A Web service is any application that can be published, located, and invoked through the Internet. Each Web service has a Web service description language document (WSDL)(W3C, 2001), which consists of an XML (W3C, 2006) document providing all required knowledge to communicate with the Web service, including its location, supported transport protocols, messages formats, list, and signatures of published operations.

A Web service can perform any kind of transactions that may range from getting a city's temperature to a more complicated transaction, like for instance, searching and/or building the best travel packages from specific travel agencies. The main objective of Web services is to allow, at a high level of abstraction, applications to be accessible over the Internet. They can be of great use, for instance, for 3G networks operators to expose their core network functionalities to third parties (3GPP, 2003) and for digital imaging where they can provide an important benefit to the digital photography industry. The common picture exchange environment (CPXe) (CPXe, 2005), a Web service business framework, will make transfer and printing of digital images as suitable as the use of films.

SOA defines three roles (i.e., provider, requester, and registry) and three operations (i.e., publish, find, and bind). The relationship between the roles and the operations are illustrated in Figure 1. Additional information on the Web services architecture can be found by Kreger (2001).

The starting point in Web services activities is the development, deployment, and publication of the Web service by its provider. When a requestor (client) needs a specific Web service,

Figure 1. Service oriented architecture



the client probes the registry for a list of potential Web services. The returned list contains matching records; each record contains required information to connect to the corresponding Web service. Based on a set of criteria (i.e., location, availability, etc.), the requestor selects a suitable Web service and binds to it.

Web services can be developed either from scratch or by composition. Composition of Web services is the process of aggregating a set of Web services to create a more complete Web service with a wider range of functionalities. This composition has a considerable potential of reducing development time and effort for new applications by reusing already available Web services.

Currently, there are standards or languages that help in building composite Web services, such as: Web services flow language (WSFL) (Leymann, 2001), DAML-S (Ankolekar, Burstein, Hobbs, Lassila, Martin, McDermott et al., 2002), Web services conversation language (WSCL) (Banerji, Bartolini, Beringer, Chopella, Govindarajan, Karp et al., 2002), Web services choreography interface (WSC) (Arkin, Askary, Fordin, Jekeli, Kawaguchi, Orchard et al., 2002), and business process execution language (BPEL) (Andrews, Curbera, Dholakia, Goland, Klein, Leymann et al., 2003). These languages make the Web services composition process easier by providing concepts to represent partners and orchestrate their interactions. BPEL, which represents the merging of IBM's WSFL and Microsoft's XLANG, is gaining a lot of interest and is positioned to become the primer standard for Web service composition.

RELATED WORK

Most works on Web services focus on their development and deployment. Management of Web services (W3C, 2002), and in particular fault and performance management, are not yet a well-studied area. However, some interesting works have to be cited.

Existing approaches for management of Web services include approaches from network management and those that have been developed specifically for Web services. The approaches that have been used for network management for a long time seem to be a candidate for the management of Web services. However, their main drawbacks are due to the major differences between Web services and network components, and the need for the participation of a component in its management. In fact, most network components (devices) run standardized protocols that have specific and known attributes to be managed. Manufacturers of components running proprietary/no-standard protocols and/or applications often provide their customers with specific management agents/applications or well-defined sets of APIs.

In network oriented approaches, simple network management protocol (SNMP) (Case, Fedor, Schoffstall, & Davin, 1990) is based on TCP/IP and the client/server communication mode. In this approach, an agent associated with a management information base (MIB) (Perkins & McGinnis, 1997), communicates with a management station by processing ***get*** (report the value of an attribute) and ***set*** (modify the value of an attribute) messages and generating ***trap*** messages (unsolicited notification). Thus, SNMP management system requires a SNMP agent, a MIB, and a management station (manager).

The common management information protocol (CMIP) (ISO/IEC, 1998) fulfills in the OSI reference model protocol stack (ISO/IEC, 1989), a role similar to that of SNMP in TCP/IP. CMIP has many advantages compared to SNMP, including the number of available commands and the possibility to operate over TCP/IP. However, complexity and long development time, especially CMIP over TCP/IP (CMOT) (Warrier, Besaw, LaBarre, & Handspicker, 1990), have kept its adoption pervasively.

A considerable amount of work in the Web services community is dedicated to the determination of the requirements and the definition of specific approaches for Web services management. These approaches can be divided into two main groups: approaches based on active testing and approaches requiring the Web service (architecture) to support management interfaces. The World Wide Web (W3) Consortium presents a set of requirements that Web services management architectures should satisfy to provide management features (W3, 2004). This includes the definition of standard metrics, management operations, and methodologies for accessing management capabilities. The complying architectures must provide a manageable, accountable, and organized environment for Web services operations. It must support at least resource accounting, usage auditing and tracking, performance monitoring, availability, configuration, control, security auditing and administration, and service level agreements. Another approach in which the Web service provides specific interfaces for management is presented by Farrell and Kreger (2002). The developer is supposed to supply commands and APIs for management operations that are invoked by the management system.

Casati, Shan, Dayal, and Shan (2003) classify the management of Web services into three levels: infrastructure-level, application-level, and business-level. The infrastructure-level deals with the Web service platform while the application-level focuses on the Web services themselves. The business-level takes into consideration the conversations between a Web service and its client

Management approaches presented by W3 (2004), Farrell and Kreger (2002), and Casati et al. (2003) assume that the Web service will provide management operations that one can invoke. Developers of Web services have then to develop and deploy these operations in addition to the core business operations the Web service is offering.

A couple of management tools to be integrated into Web services environment are already available. Hewlett Packard's Web service management engine (HP, 2007) is a collection of software components that enables some management features, including the definition and the enforcement of service level agreement (SLA). Parasoft (2006) provides a set of tools (e.g., SOAPTest, .TEST, WebKing) to assist during the lifecycle of a Web service. These tools have to be installed and configured, thus requiring extra resources and introducing new cost for Web services providers.

There has been a considerable amount of work on testing Web services in the last couple of years. The work can be divided into two main groups: works targeting functional aspects of Web services and works tackling nonfunctional. The first group is concerned with the correctness of interactions between Web services and their clients while the second group is concerned with QoS management of Web services.

Functional Management

The majority of work on functional management is based on active testing where appropriate test cases have to be carefully generated, executed, and their results analyzed. This unavoidable phase of active testing has, however, practical limitations. First of all, exhaustive testing is impractical for quite large Web services. In fact, test cases can not cover all possible execution scenarios that a Web service will have to handle while serving clients' requests. The size of test cases is bounded by the cost a Web service's provider is willing to spend on testing activities. Usually, active testing stops whenever developers are confident that the Web service is good enough to be put into the market.

Many recent results were published lately describing test cases generation methods for Web services; they are mainly based on static analysis of WSDL documents. Xiaoying, Wenli, Wei-Tek, and Yinong (2005) present a method for test data generation and test operation generation based on three types of dependencies: input, output, and input/output. Jiang, Xin, Shan, Zhang, Xie, and Yang (2005) propose a method for test data generation in which a set of tests is randomly generated based on the WSDL document. ChangSup, Sungwon, In-Young, Jongmoon, and Young-Il (2006) combined both EFSM models and WSDL documents to generate test cases.

QoS Management

QoWS management includes definition of QoS attributes, QoS publication, discovery, validation, and monitoring. Existing approaches for QoS management can be classified into two groups: one based on extending related technologies including WSDL and UDDI to support QoS and the other mandating independent entities to perform some or all of QoS management tasks.

In the first category, W3C (2003) extends SOAP header to include QoS information. WSDL is also extended to describe QoS parameters, their associated values, computation units (e.g., millisecond, request/second), and so forth. UDDIe, a UDDI extension, consists of extending the current UDDI data structure with QoS information (ShaikhAli, Rana, Al-Ali, & Walker, 2003). The aim of these extensions is to allow QoS-based publication and discovery of Web services.

In the second group, solutions are presented for one or more of the following QoS management operations:

- **QoS attributes:** The first step in QoS management is the definition of evaluation's criteria and attributes. A set of attributes have been defined, studied, and used in software engineering for a long time (Fenton & Pfleeger, 1997; Gray & MacDonell, 1997; Salamon & Wallace, 1994).
- **QoS publication and discovery** (Kalepu, Krishnaswamy, & Loke, 2004; Ran, 2003; Serhani, Dssouli, Hafid, & Sahraoui, 2005): This operation allows providers to include QoS information in WSDL. This information is then used by requestors when selecting the appropriate Web service in terms of functional and QoS requirements.
- **QoS verification** (Kalepu et al., 2004; Serhani et al., 2005; Tsai, Paul, Cao, Yu, & Saimi, 2003): This operation allows the provider to certify that the QoS claimed by the Web Service is accurate.

- **QoS negotiation** (Serhani et al., 2005): If the available published QoS requirements do not satisfy a client's needs, negotiation operations and strategies can be followed to reach an agreement on different QoS attributes.
- **QoS monitoring** (Benharref, Glitho, & Dssouli, 2005; Benharref, Dssouli, Glitho, & Serhani, 2006; Ho, Loucks, & Singh, 1998; Schmietendorf, Dumke, & Reitz, 2004; Yuming, Chen-Khong, & Chi-Chung, 2000): Performs monitoring of Web services during interactions with clients to assess if the QoS attributes agreed upon in previous points are delivered.

Discussion

All the solutions presented above fit in one or more of the following categories:

1. Platform-dependent
2. Assume that a Web service will participate in its management by providing specific interfaces (e.g., W3C architecture)
3. Are based on active testers

The usage of platform-dependent management approaches is restricted to the targeted platform. When management features are embedded to the hosting platform, they are only available to the provider and cannot be used by clients or third party certification entities. A client might need management information for two tasks: (1) during discovery and selection to select the appropriate Web service, and (2) during invocation to assess the quality of the interactions. The client must rely on management information made available by the Web service provider and has no mean of verifying it. Moreover, information used in assessing the behavior is taken from one location, that is, at the provider's side. There are many situations, in composite Web service for example, where this information should be gathered from different sources and locations.

The Web services architecture becomes more complex if it has to support management features in addition to its basic functions. The performance of the Web service and its hosting platform is also degraded due to these additional features. Moreover, developers of Web services have to also implement the needed interfaces and APIs to support management. Since these features will be used somehow sporadically, the return on investment of their development and deployment might be relatively low.

Once a Web service is interacting with clients, active testing cannot be used to monitor, in real time, the correctness of interactions. Moreover, application of generated test cases consumes resources and may disturb the Web service.

Since management of Web services is somehow at its earlier stages, related work usually concentrates more on provision of management features without evaluating the overhead they generate. In order to select the appropriate management approach, a potential user must be able to evaluate it in terms of usefulness and associated cost.

Furthermore, most of the existing work on management of Web services does not tackle management issues at the earlier phase of their development. However, management features need to be addressed as early as possible in the development process, especially during the design and implementation phases. For example, design for manageability will describe manageability scope and functions. Moreover, it will expose a Web service as a manageable entity providing some of the following capabilities (Farrell & Kreger, 2002): operations, events, interfaces, status, configuration, and metrics that can be used for managing and controlling Web services.

To solve some of the limitations of related work cited above, this chapter presents a novel framework for management of Web services. This framework considers QoWS management issues from earlier phases of the development lifecycle of a Web service. These issues are specified during specification and design, verified and certified before deployment, published with the WSDL document, used during discovery and selection, and passively monitored during invocation of the Web service.

To illustrate the applicability of our approach for management of Web services, we will be using a case study all along the chapter's sections to demonstrate how each management task, at each development phase, can be achieved. Introduction to this case study and its utilization context are given in the next section.

CASE STUDY

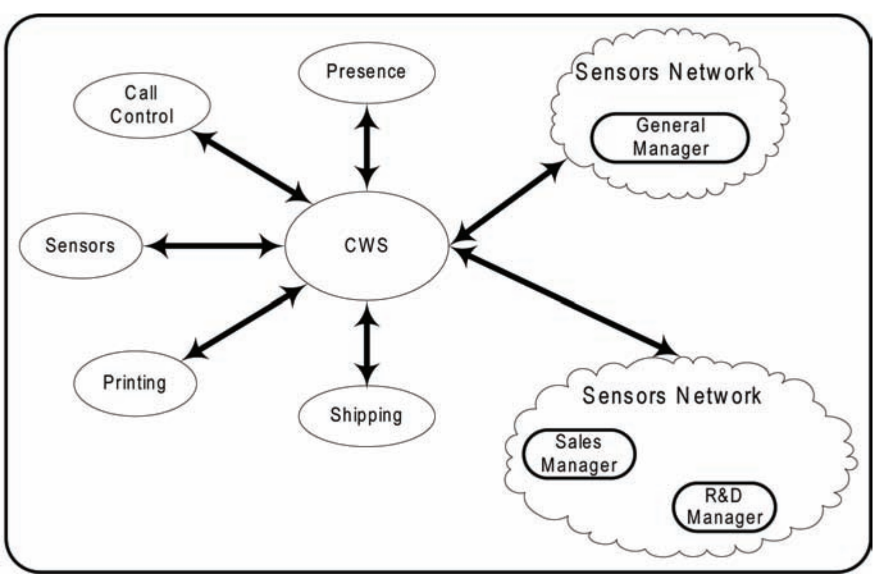
For the end of year meetings, a general manager has to meet with managers from different departments (e.g., Sales, R&D, etc.). Managers are located in different locations and, because of their busy timetables, they cannot meet in a single location. A practical alternative is to conduct these meetings in a series of teleconferences. Only mangers are concerned and only those of them who are in their offices can join a conference. This is implied by security issues since confidential information will be exchanged during meetings and communication between different locations is secured (e.g., VPN). At the end of each meeting, meetings' reports must be printed and distributed among all participants.

The manager decides to use available conferencing Web services. Such Web services should allow creation of conferences, and the addition and removal of participants to conferences depending on their profiles and physical locations. At the end of each meeting, the Web service should be able to submit produced reports for printing and deliveries.

The general manager is highly concerned with the environment in which meetings will be carried out using Web services. A thorough QoWS-based discovery and selection operation had lead to the utilization of "conferencing Web service" (CWS), a QoWS-aware composite Web service, which performs all of the required tasks. The manager decides to make use of the monitoring feature of the management framework to assess the QoWS of the CWS.

To perform all these tasks, the CWS is a composition of the following basic Web services (Figure 2):

Figure 2. Composite/basic Web services



- **Presence:** This Web service contains information on managers' profiles (e.g., name, address, location, status, position, and availability).
- **Sensors:** This Web service detects the physical location of managers.
- **Call Control:** This Web service creates and manages multiparty conferences (e.g., initiates conferences, adds/removes participants, and ends conferences).
- **Printing:** At some points during conferences or later on, managers may want to print documents (e.g., meeting reports, etc.). The printing Web service will print these documents and keeps them for shipping.
- **Shipping:** Documents printed during and after the conference should be distributed among participants at different locations. The CWS informs the shipping Web service of the location of the documents to be shipped and their final destinations.

WEB SERVICES MANAGEMENT FRAMEWORK

For a better exploitation of Web services, a Web services lifecycle is supposed to integrate features such as QoWS factors' precise definition, QoWS specification, QoWS-based discovery and selection, and QoWS monitoring. This implies that these features need to be addressed in earlier phases through Web services development process, especially during the design phase, then ultimately in the implementation phase, and possibly during selection and invocation of the Web service. QoWS management information, for example, is first specified then published to be later on discovered by clients using QoWS-aware discovery. Our approach is to investigate possibilities to augment the development process of Web services with the above important features.

In the subsequent sections, we will describe how management of QoWS should be supported during development, publication and deployment, discovery, selection, and invocation.

During Development of the Web Service: Behavior's Specification

The first step in a QoWS-aware development and utilization of Web services is the selection and concise definition of factors that will characterize the quality of a Web service. As cited above, these factors are divided into functional and nonfunctional aspects.

Functional Aspects

Definition of functional attributes specifies the correct functional behavior of the Web service for each invocation of each published operation. This covers the content of invocations, their responses, and their sequence.

Two main ways for functional behavior's description have been studied in the literature, concentrating on formal models and knowledge bases/expert systems (Vijayananda & Raja, 1994). Formal models have many advantages over expert systems. First of all, expert systems rely on human expertise and are more appropriate for systems that have been encountered previously. Second, formal models can be useful for automatic generation of source code and executable test cases. For these reasons, finite state machines (FSM) (Dssouli, Saleh, Aboulhamid, En-Nouaary, & Bourhfir, 1999), a widely known formal model, will be used to specify expected functional behaviors of Web services in this chapter.

A finite state machine M is defined as a tuple $(S, S_0, X, Y, D_S, \delta, \lambda)$, where

- S is a set of states,
- $S0 \in S$ is the initial state,

Figure 3. XML representation of an FSM machine

```
<fsm name="Conferencing Web Service ">
 <state name="Init" initial="YES">
          <transition ID="t1"          input="Config_Valid" output="True" next="Ready"/>
          <transition ID="t2"          input="Config_Invalid" output="False" next="Init"/>
 </state>
 <state name="Ready" initial="NO">
<transition ID="t3"          input="CreateConf_Valid" output="True" next="ConfCreated"/>
 </state>
 <state name="ConfCreated" initial="NO">
          <transition ID="t4"          input="AddUser"output="True" next="ConfCreated"/>
 </state>
…
</fsm>
```

- X is a finite set of inputs,
- Y is a finite set of outputs,
- $D \subseteq S \times X$ is the specification domain,
- $\delta: D^S \rightarrow S$ is the transfer function, and
- $\lambda: D^S \rightarrow Y$ is the output function.

The machine starts at S_0. Whenever an input is received, λ computes the corresponding output and δ determines the corresponding next state(s).

An FSM can be represented by an XML document as illustrated in Figure 3, which gives a partial overview of the FSM machine of the CWS. The root of the document (***fsm***) has an attribute (***name***) and a set of children which represents states. The name is a textual description of the Web service. Each child has a name, the attribute (***initial***), and a set of ***transitions***. The name is a textual description of the state while the attribute "initial," if set to YES, indicates that this is the initial state of the machine. A transition has four attributes: ***ID***, ***input***, ***output***, and ***next***. The first attribute is a textual description of the transition, the second attribute identifies the event that triggers this transition if the machine is in the associated state, the third attribute is the output generated when firing that transition, and the last attribute specifies the state that the machine will reach after firing the transition.

Nonfunctional Aspects: QoS

QoS consists of a set of factors or attributes such as response time, reliability, availability, accessibility, and so forth. Information on QoS attributes can be specified in many different ways. It can be described in a separate document, embedded within the description of functional behavior, or as an extension to WSDL document. However, to allow QoWS-aware discovery and selection of Web services, QoWS attributes should be available within the WSDL document. The client indicates preferences in terms of QoWS when probing the registry. The registry returns then a list of available Web services providing required operations with requested QoWS.

The first step in extending SOA with QoS is the definition of QoS attributes. In this chapter, we will focus on the following attributes:

- **Processing Time (PT):** This is a measure of the time a Web service takes between the time it gets a request and the moment it sends back the corresponding response. PT is computed at the Web service's provider side.

- **Maximum Processing Time (MxPT):** This is the maximum time the Web service should take to respond to a request.
- **Minimum Processing Time (MnPT):** This is the minimum time the Web service should take before responding to a request. Unlike PT, which is a dynamically computed attribute, MnPT and MxPT are statically defined and MnPT ≤ PT ≤ MxPT.
- **Response Time (RT):** It consists of the time needed between issuing a request and getting its response. It is measured at the client's side to include the propagation time of requests and responses.
- **Maximum Response Time (MxRT):** This is the maximum accepted time, for the client, between issuing a request and getting its response.
- **Minimum Response Time (MnRT):** This is the minimum time, for the client, between issuing a request and getting its response. This attribute in unlikely to be used since the client is usually more interested in MxRT. For the client, RT ≤ MxRT must always be satisfied.
- **Availability:** This is a probability measure that indicates how much the Web service is available for use by clients. It can also consist of the percentage of time that the Web service is operating.
- **Service Charge (SC):** It defines the cost a client will be charged for the Web services utilization. SC can be estimated by operation, type of requests, period of utilization, session, or by volume of processed data.
- **Reputation:** This is a measure of Web services' credibility. It depends basically on previous end users' experiences while using the Web service. Different users may have different opinions on the same Web service. The reputation value can be given by the average ranking given to the Web service by several users.

MnPT, MxPT, availability, and SC are related to profiles of users of the Web service. This profiling is based on the type of subscriptions of clients and/or the QoWS they are willing to pay for. For

Figure 4. QoWS in WSDL document

```
<?xml version="1.0" encoding="UTF-8"?>
 <wsdl:portType name="ConferenceService">
  <wsdl:operation name="addUser" parameterOrder="userAddresse callID">
   <wsdl:input message="intf:addUserRequest" name="addUserRequest"/>
   <wsdl:output message="intf:addUserResponse" name="addUserResponse"/>
            <Profile name="GOLD">
                  MnPT = NULL
                  MxPT = 10ms
                  SC= "$10"
            </Profile>
            <Profile name="SILVER">
                  MnPT = 10ms
                  MxPT = 30ms
                  SC= "$5"
            </Profile>
  </wsdl:operation>
 </wsdl:portType>
```

example, a gold-subscribed user must be served quicker (MnRT = 0) than a bronze-subscribed user (MnRT > 1ms).

Figure 4 illustrates embedded QoS attributes in the definition of the operation tag within the WSDL document of the CWS.

Before Deployment and Publication: QoWS Verification and Certification

Once a Web service is developed, it must be tested to verify whether it is correct with regards to the behavior's specification document produced during preceding development phases. The management framework has features that a Web service's developer can use to verify and certify the Web service's behavior. This certification information is then published with the WSDL description of the Web service so potential clients will use it.

Verification and certification procedures enable providers to evaluate QoWS of their Web services prior to the publication. Our approach consists of a two-phase verification and certification technique, which is conducted by a verifier Web service and a certifier Web service (Figure 7). The first phase consists of verifying the WSDL document, including the QoWS parameters description. The second phase consists of applying a measurement technique to compute the QoWS metrics stated in the Web service interface and compares their values to those claimed in the WSDL document. This is used to verify the conformity of a Web service to its description from a QoWS point-of-view (QoWS testing). Therefore, a set of QoWS test cases are defined and used as input to QoWS verification. The configuration and generation of these test cases is described in detail by Serhani et al. (2005). Once the Web service passes the verification tests, the certifier issues a conformance certificate to certify that QoWS claims are valid. This certificate will be considered as a key differentiator between Web services offering similar functionalities. The verifier and certifier perform the following tasks:

- It asks for information about the provider and its Web service (e.g., servers' resources capacity, connections used, Network information, etc.).
- It checks the WSDL files of the target Web services (e.g., location, interface, and implementation description)
- It makes sure that all published operations are available.
- It verifies the QoWS described in WSDL. The QoWS verifier can initiate, if necessary, additional tests to validate other information provided in the WSDL document. This information concerns QoWS attributes classification (e.g., definition, computation logic, and upper and lower bounds).
- It stores the verification report in a specific-purpose database.

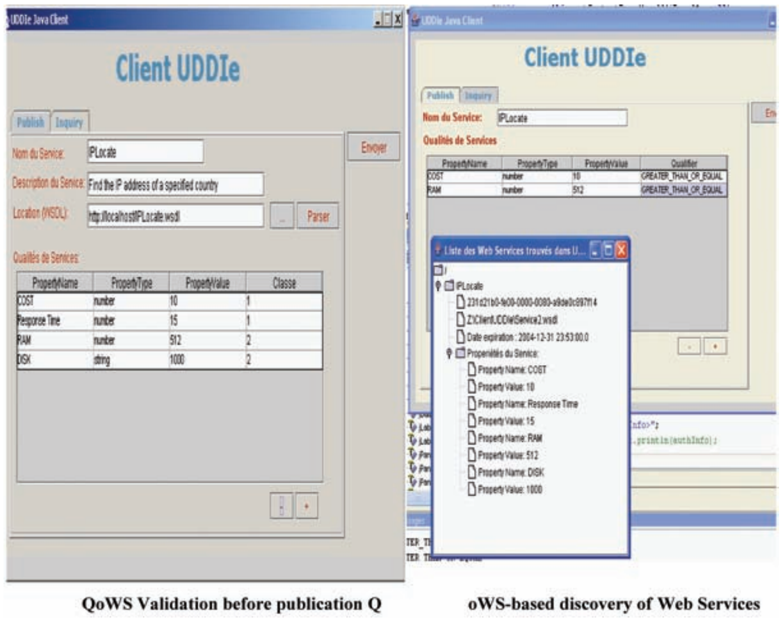
During Discovery and Selection: QoWS-based Discovery and Selection

In standard SOA, the find operation is based on the name of the Web service. A client is likely to get a list of Web services following a basic find operation issued to a UDDI registry. Alternatively, an intelligent find operation must consider, in addition to the name, QoWS information verified and certified in the previous phase, so the returned list of Web services is somehow short and concise.

Our framework defines how an intelligent discovery operation should look like and how it can be used. Discovering a Web service will be based on its functionalities as well as its QoWS. We automated the processes of publication and discovery of Web services based on QoWS using a supporting application (Figure 5), which allows the following tasks:

- **Publication:** In order to publish their Web services using our application, providers

Figure 5. UDDIe client's application for QoWS-driver publication and discovery



should supply the Web service name, description, and the location of its WSDL document. This document is then parsed to validate its content and to display the list of QoWS information. The validation process verifies the correctness of published operations in addition to the QoWS information. The provider can add/remove/modify QoWS attributes before publication. At this stage, the WSDL document is validated and the provider publishes the QoWS-enabled WSDL document.

- **Discovery:** The application allows the user to query the registry while specifying the Web service name and the set of required QoWS attributes and their related values. The list of Web services descriptions that fulfill the client's requirements is displayed via the application interface. The frame contains corresponding Web services and their associated QoWS information.

During Invocation: QoWS Monitoring

During interactions between a Web service and its client, it might be necessary to assess if the Web service is behaving as initially specified and claimed in previous subsections. This assessment will require a continuous online monitoring of interactions between the Web service and its client.

Online monitoring of Web services requires passive observers (Benharref et al., 2005, 2006). A passive observer receives a copy of all exchanged messages between a Web service and its client and checks their validity. Passive observation of systems modeled as FSM is usually performed in two steps (Lee, Netravali, Sabnani, Sugla, & John, 1997):

1. **Passive Homing (or state recognition):** In this step, the observer is brought to a state equivalent to the one that the Web service might be in. If no such state is found, a fault

is immediately reported. The set of messages leading to this state is known as the homing sequence. This first step is required if observation starts while the Web service and its client already exchanged some messages. These messages will not be available to the observer but the latter can figure out a homing sequence to determine the appropriate state.

2. **Fault Detection:** Starting from the state identified in the previous step, the observer checks the observed behavior against the system's specification. If an observed event is not expected then a fault is immediately reported.

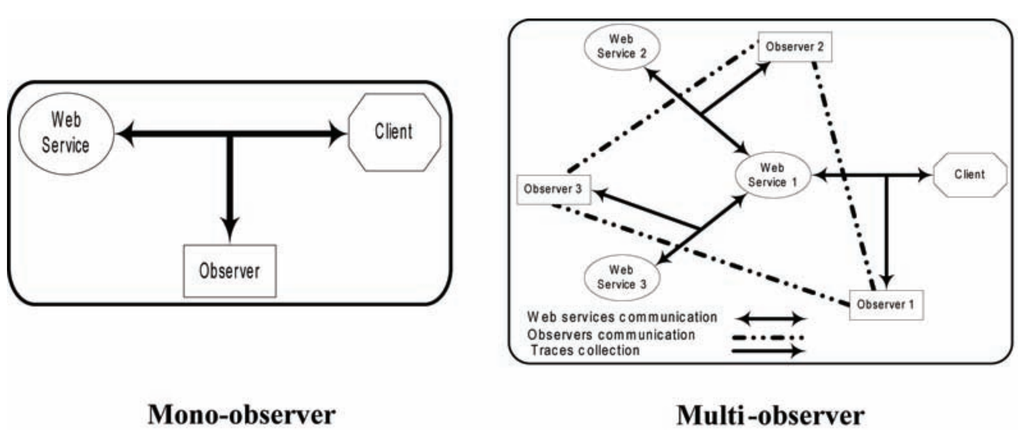
The observation in distributed architectures requires the selection of a number of observers and their best locations (where to get copies of exchanged messages). The number and location of the points of observation affect significantly the detection capabilities of the observation architectures. For example, if the observed Web service is a composite Web service, it might be more interesting (in terms of misbehavior detection) to consider a network of observers, that is, an observer for each Web service rather than a unique observer for the composite Web service. In such architectures, cooperation of all observers can generate pertinent information for Web services management. The consideration of a global observer (for the composite Web service) and local observers (for composing Web services) presents a framework where this cooperation can be orchestrated for the benefit of better misbehavior detection.

Our Web services management framework offers two monitoring architectures as depicted in Figure 6, which shows a mono-observer architecture (Benharref et al., 2005) and multiobserver architecture (Benharref et al., 2006).

Three types of interactions are illustrated in Figure 6. Web services communication refers to the SOAP-based communication between Web services and their clients. Traces collection consists of forwarding messages exchanged between the observed Web service and its client to local observers. Observers communication conveys information between observers. This information is divided into three categories:

3. **Configuration information:** During configuration of different observers, local observers must indicate to the global observer which Web service they are observing and where they are located. The global observer needs this information to identify observers and associates the traces it will receive to

Figure 6. Monitoring architectures



appropriate observers/Web services.

4. **Traces from local observers to the global observer:** Whenever a local observer gets a trace, it sends it to the global observer.
5. **Notifications of faults:** If the global observer detects a fault, it informs other local observers. In the case where a local observer detects a fault, it informs the global observer. The latter informs remaining local observers that misbehavior has been observed elsewhere and they should be aware of some specific traffic/actions.

Traces collection mechanisms studied by Benharref et al. (2005) show that mobile agents present the least overhead. Whenever an entity wants to use monitoring architectures, it invokes a Web service observer (WSO) that generates a set of mobile observers and sends them to locations specified during invocation.

Except specification of expected behaviors of Web services which has to be done by the provider, other management operations presented in previous sections are performed by invoking the verifier, the certifier, and the observer, three Web services provided by the management framework. These components and their associated operations are illustrated in Figure 7.

Different steps discussed above have been applied to a set of Web services to illustrate their applicability. The next section shows an example of application to the CWS introduced earlier in this chapter.

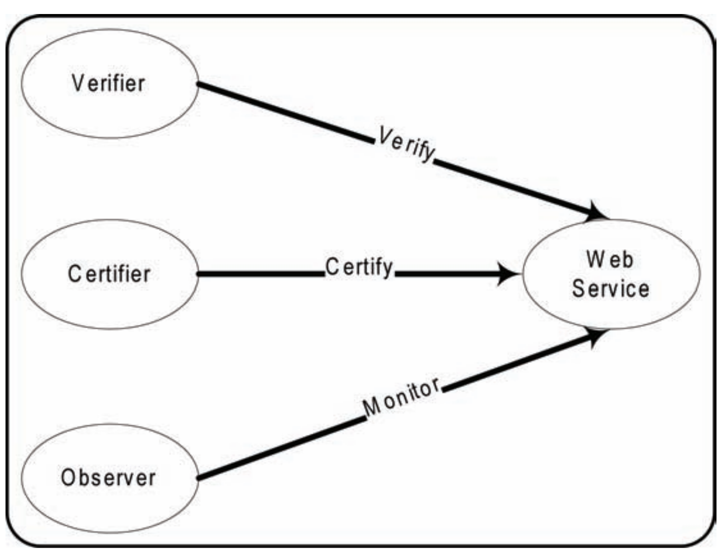
EXPERIMENTATION AND RESULTS

As indicated earlier, CWS is a composite Web service for management of teleconferences, printing, and shipping of teleconferences reports. Experimenting with the framework using this Web service required implementation and deployment of a dozen of Web services.

Implementation Issues

All Web services, including the WSO, are implemented in BEA WebLogic (BEA, 2004). Mobile observers get traces using SOAP handlers, which are available within the BEA platform. A SOAP handler, a specific-purpose Java class, intercepts a request/response to/from a Web service before it

Figure 7. Management framework components and operations



gets to the core Web service or the client respectively, and can also perform operations on it. In our case, the SOAP handler sends each intercepted request or response in a user datagram protocol (UDP) datagram to the concerned mobile observer all along with the date at which this event occurred to allow the observer to assess QoS attributes such as response time. The mobile observer checks this trace and forwards it to the global observer.

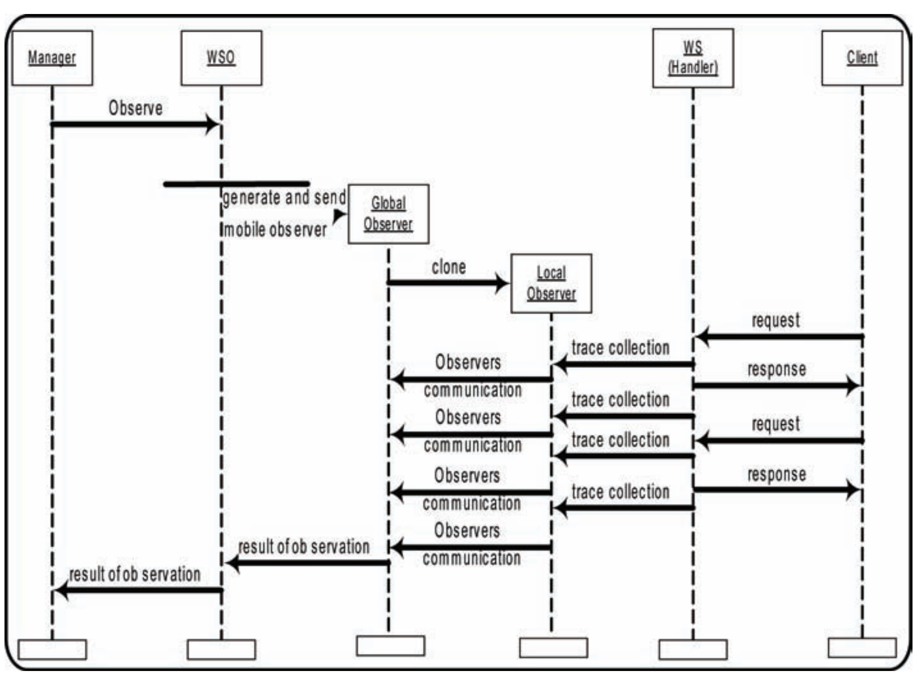
To be able to detect lost UDP datagrams, a sequence number field is used. When a mobile observer detects a lost datagram (wrong/not expected sequence number), it suspends the misbehavior detection and reperforms the homing procedure. It restarts the detection once this procedure is achieved correctly. Since the behavior/operation of SOAP handlers within all observed Web services is similar, a unique (generic) SOAP handler is developed and then distributed to all providers participating in the observation.

The overhead of the management framework can be quantitatively evaluated with regards to required computer resources and generated network overhead. Both analytical analysis and experimentations showed that most of the overhead is related to the online monitoring. In fact, the verification and certification operations are straightforward and usually conducted off-line, that is, before the Web service is made available to clients. Moreover, all required resources are located at the verifier and certifier Web services providers. For these reasons, overhead analysis presented in upcoming subsections will concentrate on online monitoring, especially traces collection.

Monitoring

In addition to the observation of the CWS, the manager wants to make sure that all the steps are performed according to the agreed on contract and QoWS. Fortunately, all the providers accept to participate, to some extent, in the monitoring. The provider of the CWS will host all mobile observers using the Jade platform (Jade, 2007). This provider will also supply WSDL documents and

Figure 8. Multiobserver deployment



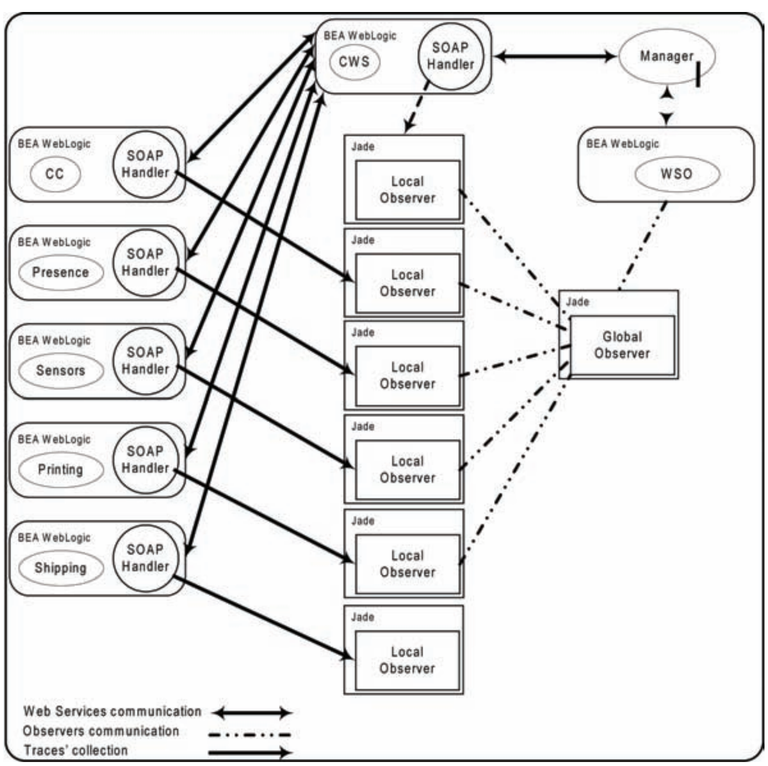
FSM models of each of the basic Web services. Basic Web services providers will configure SOAP handlers for traces collection and forward.

The observation procedure of CWS is performed following the steps detailed below and illustrated in Figure 8. To keep the figure simple, just one Web service handler and one Web service client[1] are depicted in the figure.

1. The manager invokes the WSO, providing different locations of mobile observers.
2. The WSO generates a mobile agent and sends it to one of the destinations submitted during invocation in Step 1.
3. Once the mobile agent gets into its destination, it clones itself as many times as required to observe all Web services.
4. The mobile agent observing the CWS becomes the global observer; other mobile observers are local.
5. SOAP handlers forward traces to appropriate mobile observers.
6. Local observers analyze these traces and forward them to the global observer.
7. Whenever misbehavior is detected (by global or local observers), correlation then fault location is initiated by the global observer to find the faulty Web service.
8. The global observer reports to the WSO.
9. The WSO reports to the manager.

Each local observer is listening to a UDP port to receive events from SOAP handlers. The global observer is listening to two different UDP ports: one to receive events (request or response) from local observers and another port to receive information on detected misbehaviors by local observers. The SOAP handler sends each event between a client and its Web service to the attached local observer. The latter forwards this event to

Figure 9. Multiobserver configuration for monitoring of CWS



the global observer and checks the validity of this event. If misbehavior is detected, the local observer notifies the global observer. Figure 9 shows the overall configuration of interacting client, Web services, mobile observers, and communication between these entities.

Processing CPU and Memory Utilization

Computer resources (CPU and memory) used by traces collection entities are somehow insignificant with regards to the minimal standard configuration of actual personal desktops/laptops. Except for the mobile agent approach, CPU and memory utilization are so low that they are even difficult to precisely evaluate.

For mobile observers, CPU and memory utilization on a laptop equipped with an AMD Athlon 64/3000+ processor and 512MB RAM, CPU and memory utilization are as follows:

- **Hosting a mobile platform:** If the mobile agent administration interface is located on the laptop, the CPU utilization varies between 2% and 4%. For memory, it uses around 30 Megabytes.
- **Joining a mobile platform[2]:** If the mobile agent platform is running on a remote computer, joining it requires 12 MBytes memory at the laptop and around 2 MBytes on the host running the administration interface. For CPU, there is almost no impact at both sides.
- **Receiving a mobile observer:** When a mobile observer is received, it requires around 27 MBytes of memory. For CPU, there is a high utilization during 1 to 2 seconds while initializing and displaying the graphical interface of the mobile observer, then the CPU utilization goes back to previous level
- **Processing traces:** Even in the worst case where traces are received with a very small delay, the CPU used by the mobile observer for analyzing them is around 2%. However, there is no additional memory utilization.

Network Load

The network load introduced by the observation is classified into load due to the deployment of mobile agents and load due to the traces collection process.

Deployment Load

Since all observers are located at the composite Web service provider's side, only one mobile agent is generated by the WSO. The size of the traffic to move a mobile agent is around 600 Kilobytes (600 KB).

Traces Collection Load

Generally, for each interaction between a Web service and its client, 2 UDP datagrams are generated: a first datagram from the SOAP handler to the local observer, and a second datagram from this local observer to the global observer. Whenever a local observer detects misbehavior, a third datagram is sent (fault notification). The average size of a datagram is 150 bytes. So, each response/request pair introduces 4 datagrams if everything goes fine, 5 datagrams if one of the events is faulty, or 6 datagrams if both are faulty. We suppose that faults will not occur very often, and then few fault notifications will be generated. This assumption is realistic since all Web services are supposed to undergo acceptable testing process before their deployment. The traces collection load then is reduced to the forward of events, that is, 4 datagrams for a request/response pair. This represents a load of 600 bytes.

Table 1. Some of the executed scenarios

Target Web Service	Fault description	Comments
CWS	Submit a printDocument request before creating a conference	Fault detected by local and global observer
Call Control	Add a user before creating a conference	Fault detected by local and global observer
Presence	Try to add a user to the conference that is not recognized by the Presence service	Fault detected by local and global observer
Shipping	Request shipping of a document that has not been submitted for printing	Fault detected by local and global observer
Shipping	A trace collection event (shipDocument response) from a handler to the local observer is lost (Figure 10)	Neither the local observer nor the global observer will detect the fault.
Shipping	A trace collection event (shipDocument response) or a fault notification from a local observer to the global observer is lost (Figure 11)	The global observer will not be able to detect the fault or process the notification (correlation)

Results and Analysis

To illustrate the detection capabilities of our architecture, we injected faults to some Web services and/or in the network and monitored the behavior of observers (Table 1). The observers have been able to detect most of the injected faults.

A fault that cannot be detected occurs when the last event in a communication between a Web service and its client is lost (see Figures 10 and 11). As discussed earlier, traces are sent as UDP packets. To be able to detect lost packets and recover the observation, a sequence number attribute is used. An observer detects a lost packet if the sequence number of the following packet is different than expected. When a lost packet carries the last event in a communication, observers will not be able to detect this incident since no future packets will arrive. Table 1 shows brief descriptions of some of the executed scenarios and the reactions of observers (both local and global) to the fault.

CONCLUSION

Web services are a new generation of Web applications. This new paradigm of communication puts more emphasize on business-to-business

Figure 10. Trace lost before getting to local observer

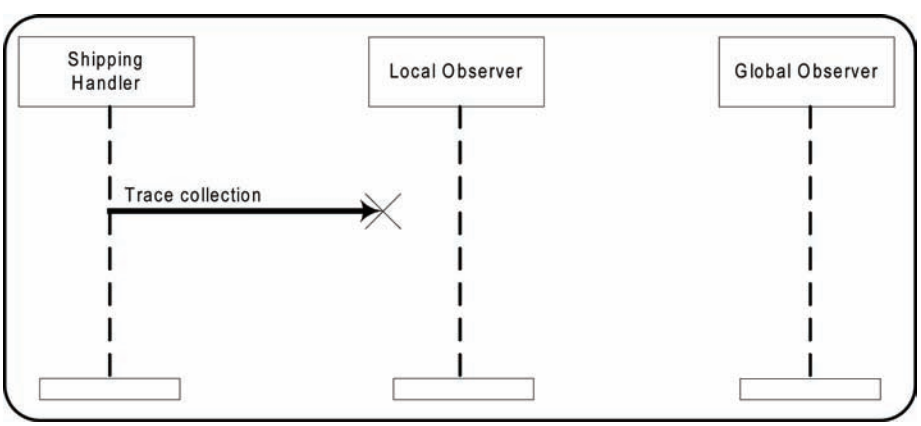
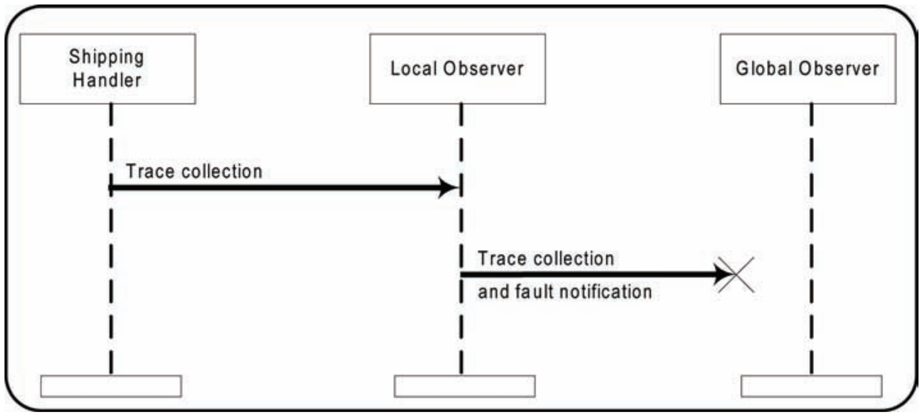


Figure 11. Trace or fault notification lost before getting to global observer



interactions rather than the business-to-consumer transactions model that the Internet was largely providing. Management of Web services is of prime importance for all entities involved in service oriented architecture. In an environment where the interacting components are not known a priori, can be on different operating systems and platforms, and coded in different programming languages, the management of Web services is very challenging compared to the management of traditional distributed systems.

In this chapter, we have developed a framework to provide management features to Web services providers and clients by supporting management activities all along the lifecycle of a Web service, from development to invocation. The framework encourages providers to consider management activities while developing their Web services by specifying QoWS attributes. It allows clients to select appropriate Web services using QoWS information published previously by providers. Clients also make use of the framework to check if the Web services they are actually using or planning to use are behaving correctly in terms of functional and nonfunctional facets. A prototype of the framework has been developed and used in management of a set of Web services. In this chapter, a conferencing Web service has been used to evaluate the effectiveness and the overhead of the framework.

As a future work, we intend to enhance the framework by providing support to other management operations, such as fault isolation and repair. Work is in progress to implement the remaining components of the framework and to evaluate the overall features of the framework on a large scale Web services environment.

REFERENCES

W3. (2004). *Web services endpoint management architecture requirements*. Retrieved May 22, 2008, from http://dev.w3.org/cvsWeb/2002/ws/arch/management/ws-arch-management-requirements.html?rev=1.7

W3C. (1999). *HTTP*. Retrieved May 22, 2008, from http://www.w3.org/Protocols/

W3C. (2001). *Web services description language*. Retrieved May 22, 2008, from http://www.w3c.org/TR/wsdl

W3C. (2002). *Web services management concern* (white paper). W3C Consortium.

W3C. (2003, November). *QoS for Web services: Requirements and possible approaches*. Author.

W3C. (2004). *Simple object access protocol*. Retrieved May 22, 2008, from http://www.w3c.org/TR/soap

W3C. (2006). *eXtensible markup language*. Retrieved May 22, 2008, from http://www.w3c.org/XML

Andrews, T., Curbera, F., Dholakia, H., Goland, Y., Klein, J., Leymann, F., et al. (2003). *BPEL4WS version 1.1 specification*. Retrieved May 21, 2008, from ftp://www6.software.ibm.com/software/developer/library/ws-bpel.pdf

Ankolekar, A., Burstein, M., Hobbs, J. R., Lassila, O., Martin, D., McDermott, D., et al. (2002). DAML-S: Web service description for the Semantic Web. In S. Verlag (Ed.), *First International Web Conference* (pp. 348-363). Sardinia, Italy: Springer Verlag.

Arkin, A., Askary, S., Fordin, S., Jekeli, W., Kawaguchi, K., Orchard, D., et al. (2002). *Web service choreography interface (WSCI)*. Retrieved May 21, 2008, from http://www.w3.org/TR/wsci/

Banerji, A., Bartolini, C., Beringer, D., Chopella, V., Govindarajan, K., Karp, A., et al. (2002). *WSCL: The Web services conversation language*. Retrieved May 21, 2008, from http://www.w3.org/TR/wscl10/

BEA. (2004). *WebLogic plateform*. Retrieved May 21, 2008, from http://www.bea.com

Benharref, A., Dssouli, R., Glitho, R., & Serhani, M. A. (2006). *Towards the testing of composed Web services in 3rd generation networks*. Paper presented at the IFIP International Conference on Testing of Communicating Systems (TestCom) (Vol. 3964, pp. 118-133). New York: Springer Verlag.

Benharref, A., Glitho, R., & Dssouli, R. (2005). *Mobile agents for testing Web services in next generation networks*. Paper presented at MATA 2005 (Vol. 3744, pp. 182-191). Montreal, Canada: Springer Verlag.

Casati, F., Shan, E., Dayal, U., & Shan, M.-C. (2003). Business-oriented management of Web Services. *Communications of the ACM*, *46*(10), 55–60. doi:10.1145/944217.944238

Case, J., Fedor, M., Schoffstall, M., & Davin, J. (1990). S*imple network management protocol (SNMP)* (RFC 1157).

ChangSup. K., Sungwon, K., In-Young, K., Jongmoon, B., & Young-Il, C. (2006). *Generating test cases for Web services using extended finite state machine*. Paper presented at the IFIP International Conference on Testing of Communicating Systems (TestCom) (pp. 103-117). New York: Springer Verlag.

CPXe. (2005). *I3A standards - initiatives - CPXe*. Retrieved May 21, 2008, from http://www.i3a.org/i_cpxe.html

Dssouli, R., Saleh, K., Aboulhamid, E., En-Nouaary, A., & Bourhfir, C. (1999). Test development for communication protocols: Towards automation. *Computer Networks*, *31*(17), 1835–1872. doi:10.1016/S1389-1286(99)00063-8

Farrell, J. A., & Kreger, H. (2002). Web services management approaches. *IBM Systems Journal*, *41*(2), 212–227.

Fenton, N. E., & Pfleeger, S. L. (1997). *Software metrics: A rigorous and practical approach* (2nd ed.). Boston: PWS Pub.

3GPP. (2003). *Open service architecture workgroup*. Retrieved May 21, 2008, from http://www.3gpp.org/TB/CN/CN5/CN5.htm

Gray, A. R., & MacDonell, S. G. (1997). Comparison of techniques for developing predictive models of software metrics. *Information and Software Technology*, *39*(6), 425–437. doi:10.1016/S0950-5849(96)00006-7

Ho, S., Loucks, W. M., & Singh, A. (1998). Monitoring the performance of a Web service. In *IEEE Canadian Conference on Electrical and Computer Engineering* (Vol. 1, pp. 109-112). Waterloo, Ontario, Canada: IEEE Press.

HP. (2007). *Open view*. Retrieved May 21, 2008, from http://www.managementsoftware.hp.com

ISO/IEC. (1989). *7498, information processing systems -- open systems interconnection – basic reference model.* Author.

ISO/IEC. (1998). *9596, information technology -- open systems interconnection -- common management information protocol.* Author.

Jade. (2007). *Java agent development framework.* Retrieved May 21, 2008, from http://jade.tilab.com

Jiang, Y., Xin, G.-M., Shan, J.-H., Zhang, L., Xie, B., & Yang, F.-Q. (2005). A method of automated test data generation for Web services. *Chinese Journal of Computers*, *28*(4), 568–577.

Kalepu, S., Krishnaswamy, S., & Loke, S. W. (2004). Verity: A QoS metric for selecting Web services and providers. In *Fourth International Conference on Web Information Systems Engineering Workshops* (pp. 131-139). Rome, Italy: IEEE Computer Society.

Kreger, H. (2001). *Web services conceptual architectures (WSCA 1.0*) (white paper). IBM Software Group.

Lee, D., Netravali, A. N., Sabnani, K. K., Sugla, B., & John, A. (1997). *Passive testing and applications to network management*. Paper presented at the International Conference on Network Protocols (pp. 113-122). Atlanta: IEEE Computer Society.

Leymann, F. (2001). *Web service flow language (WSFL) 1.0.* Retrieved May 21, 2008, from http://www-4.ibm.com/software/solutions/Webservices/pdf/WSFL.pdf

OASIS. (2005). *Universal description, discovery, and integration.* Retrieved May 21, 2008, from http://www.uddi.org/

Parasoft. (2006). *SOATest.* Retrieved May 21, 2008, from http://www.parasoft.com/jsp/products/home.jsp?product=SOAP

Perkins, D., & McGinnis, E. (1997). *Understanding SNMP MIBs*. Upper Saddle River, N.J.: Prentice Hall PTR.

Ran, S. (2003). A Framework for Discovering Web Services with Desired Quality of Services Attributes. In *International Conference on Web Services* (pp. 208-213). Las Vegas, Nevada: CSREA Press.

Salamon, W. J., & Wallace, D. R. (1994). *Quality characteristics and metrics for reusable software*. National Institute of Standards and Technology.

Schmietendorf, A., Dumke, R., & Reitz, D. (2004). SLA management - Challenges in the context of Web-service-based infrastructures. In *IEEE International Conference on Web Services (ICWS)* (pp. 606-613). San Diego, CA: IEEE Computer Society.

Serhani, M. A., Dssouli, R., Hafid, A., & Sahraoui, H. (2005). A QoS broker based architecture for efficient Web services selection. In *International Conference on Web Services (ICWS)* (Vol. 2005, pp. 113-120). Orlando, FL: IEEE Computer Society.

ShaikhAli. A., Rana, O. F., Al-Ali, R., & Walker, D. W. (2003). UDDIe: An extended registry for Web services. In (pp. 85-89). Orlando, FL, USA: IEEE Computer Society.

Tsai, W. T., Paul, R., Cao, Z., Yu, L., & Saimi, A. (2003). Verification of Web services using an enhanced UDDI server. In *Eighth IEEE International Workshop on Object-Oriented Real-Time Dependable Systems* (pp. 131-138). Guadalajara, Mexico: IEEE.

Vijayananda, K., & Raja, P. (1994). *Models of communication protocols for fault diagnosis*. Swiss Federal Institute of Technology.

Wahli, U. (2002). *Self-study guide: WebSphere studio application developer and Web* services (1st ed.). Retrieved May 22, 2008, from http://www.books24x7.com/marc.asp?isbn=0738424196

Warrier, U., Besaw, L., LaBarre, L., & Handspicker, B. (1990). *The common management information services and protocols for the Internet (CMOT and CMIP)* (RFC 1189).

WSDL. (2001). *Web services description language.* Retrieved May 22, 2008, from http://www.w3c.org/TR/wsdl

Xiaoying, B., Wenli, D., Wei-Tek, T., & Yinong, C. (2005). WSDL-based automatic test case generation for Web services testing. In *International Workshop on Service-Oriented System Engineering* (pp. 207-212). Beijing, China: IEEE Computer Society.

Yuming, J., Chen-Khong, T., & Chi-Chung, K. (2000). Challenges and approaches in providing QoS monitoring. *International Journal of Network Management*, *10*(6), 323–334. doi:10.1002/1099-1190(200011/12)10:6<323::AID-NEM382>3.0.CO;2-K

This work was previously published in Managing Web Service Quality: Measuring Outcomes and Effectiveness, edited by K. Khan, pp. 23-47, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 6.6

On the Management Performance of Networked Environments Using Web Services Technologies

Lisandro Zambenedetti Granville
Federal University of Rio Grande do Sul – Porto Alegre, Brazil

Ricardo Neisse
Federal University of Rio Grande do Sul – Porto Alegre, Brazil

Ricardo Lemos Vianna
Federal University of Rio Grande do Sul – Porto Alegre, Brazil

Tiago Fioreze
Federal University of Rio Grande do Sul – Porto Alegre, Brazil

ABSTRACT

The management of telecommunication and data networks has been based on standards defined in historical contexts quite different than the current times. As a consequence, traditional management technologies are not able to address important challenges posed by the modern infrastructures. Web Services technologies enable the proper communication of processes deployed on quite hostile environments such as the Internet. The use of Web Services for management allows the integration of low-level activities (e.g., retrieving monitoring information from gateways) with high-level business processes (e.g., creating a new product and its marketing strategy.) Despite clear advantages, Web Services-based management does not come for free; since Web Services are based on XML documents, its performance, compared with traditional management technologies, may represent an important drawback. This chapter covers the aspects of using Web Services for management focusing on the different interactions between managers and devices and the performance associated with it.

DOI: 10.4018/978-1-60566-194-0.ch046

INTRODUCTION

For more than ten years the Simple Network Management Protocol (SNMP) (Case, 1990) has been used to manage networks and services. Standardized by the Internet Engineering Task Force (IETF), SNMP is not only a protocol itself but a whole management framework widely recognized and accepted by both academia and industry. Despite its evolution through the definition of SNMPv2 (Presuhn et al., 2002) and SNMPv3 (Harrington et al., 2002), the SNMP framework still has restrictions that prevent its integration with other critical disciplines, such as e-Business, e-Learning, and e-Government. For example, since SNMP traffic is normally blocked by Internet firewalls, is it not possible for different companies to exchange management information via SNMP. Also, SNMP data is encoded following rules quite different than those based on XML (eXtensible Markup Language), normally used by e-Business solutions. Thus, although SNMP could technically be used in other fields, the framework restrictions make SNMP feasible almost exclusively for network management.

Recently, the Web Services (WS) technology has emerged as an interesting and promising management alternative that could overcome some of the SNMP problems. However, since Web Services are younger than SNMP, investigations are being carried out in order to understand the difficulties and the impact in adopting Web Services for management. First investigations in this field were more focused on the network bandwidth consumption (Neisse et al., 2004) because Web Services, which are based on XML, intuitively would consume more bandwidth than SNMP, which is a binary protocol with messages supposedly smaller. Next, response time and other performance aspects such as memory consumption and processing have been investigated as well (Dreves et al., 2004) (Pavlou et al., 2004), again because in comparison to SNMP, Web Services potentially would require more memory and processing power to store and parser XML structures.

Technically, Web Services could completely replace SNMP, but that is not actually feasible because network operators would not instantaneously upgrade or replace the already deployed SNMP-enabled devices and services just because a new management framework, based on Web Services, is available. However, solely using SNMP would not allow the integration of network management-related tasks with other tasks required by those disciplines cited before. Therefore, an intermediate approach is required in order to integrate "legacy" devices and services into Web Services-based systems. That can be successfully accomplished by the use of Web Services gateways.

Gateways have been around in the network management field almost since the beginning of SNMP. CMIP (Common Management Information Protocol) to SNMP (Saydam et al., 1998) and CORBA (Common Object Request Broker Architecture) to SNMP (Aschemann et al., 1999) are examples of gateways investigated in the past whose objective was to integrate SNMP with other technologies, in this case, CMIP (OSI, 1991) defined in the ISO/OSI management framework, and CORBA (Orfali et al., 1998). The interesting point regarding gateways for Web Services integration is that they can be designed and built using different approaches, and each design approach impacts not only on the gateway building process itself, but also on the performance of the underlying managed network and associated management system.

In this chapter we present and discuss the different approaches for Web Services for management integration, namely protocol-level, object-level, and service-level gateways. We show how SNMP-enabled devices can be effectively integrated into Web Services-based systems in a feasible fashion. The discussion about the approaches for Web Services to SNMP gateways is also presented. The gateways approaches are evaluated considering a set of evaluation parameters, such as gateways' ease of use and gateways' response time.

BACKGROUND

As mentioned before, gateways in network management have been around since the SNMP beginning. Gateways based on XML and Web Services technologies, however, are more recent.

Oh *et al*. (2002) defined SNMP to XML gateways and three methods for interactive translation: DOM (Document Object Model)-based translation, HTTP (HyperText Transfer Protocol)-based translation, and SOAP (Simple Object Access Protocol)-based translation. In the DOM-based translation, an XML-based manager calls a DOM interface that resides in the gateway. Such call is then translated to SNMP operations between the gateway and the target device. With the HTTP-based translation the gateway receives XPath and XQuery expressions defined by an XML-based manager. Such expressions are then translated to SNMP requests. Finally, in the SOAP-based translation the gateway exports more sophisticated services accessed by the XML-based manager. With these services the manager can look up information with XPath or proceed with complex queries through XQuery expressions.

Strauss and Klie (2003) have proposed an SNMP to XML gateway similar to the HTTP-based translation method of Oh *et al*. The gateway accepts HTTP messages with XPath expressions in the URL. The XPath expressions are then verified to be translated to SNMP messages. DOM is then used to access the XML documents inside the gateway, reducing the data transferred between the XML-based manager and the gateway. To cope with configuration management, HTTP POST messages are translated to SNMP Set requests.

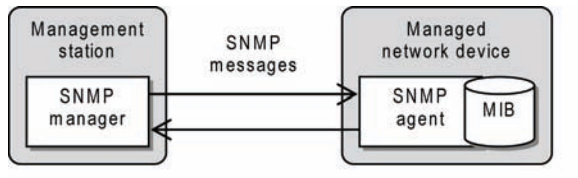
Neisse *et al*. have implemented a system that automatically creates SNMP to XML gateways given an SMI (Structure of Management Information) MIB (Management Information Base) file (Neisse et al., 2003). The gateways created retrieve information from target SNMP devices and generate XML documents sent back to an XML-based manager to be further parsed and analyzed. As the work of Strauss and Klie before, the translation is executed through the help of the smidump tool (Strauss 2003).

The work that has been done up to today uses just a subset of the facilities found in the WS architecture. As verified, gateways are created to access SNMP-based devices and export management information on XML documents (real WS, based on SOAP, are barely used). Besides, the WS description through WSDL (Web Services Description Language) and its registration in UDDI (Universal Description, Discovery and Integration) are not addressed in the work done so far. In addition, these investigations presented in this section do not address the diversity in building and using the developed gateways, which is critical for their proper use in real environments.

THREE APPROACHES FOR WEB SERVICES GATEWAYS

The typical management framework (which SNMP follows) is basically composed by four main elements: manager, agent, protocol, and management information base. The **manager**, from a management station, accesses the management information base (**MIB**) of a managed device contacting, via the management protocol, a protocol **agent**, usually located inside the target device. SNMP defines protocol operations to retrieve (*e.g.*, Get, GetNext) and modify (*e.g.*, Set) management information. Management information is defined in **MIB modules** written in plain text files according to the SMI (Structure of Management Information) (Mccloghrie 1999) specification, which is a subset of the ASN.1 (OSI 1987) language. The management information defined in a MIB module, and the way such information must be handled, defines the

Figure 1. An SNMP manager contacts an SNMP agent via SNMP messages to read or modify the managed network device MIB



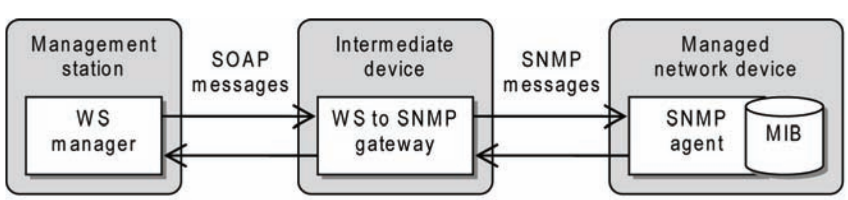
management services offered by the managed device. Figure 1 shows a conceptual view of the general SNMP framework. Further details about SNMP are available in IETF documents at http://www.ietf.org.

Two aspects of the classical management framework are important for Web Services (WS) integration. First, the SNMP framework explicitly separates the definition of the management information from the protocol that manipulates it. Second, the SNMP agent is usually just a means to access devices' MIBs through the management protocol, but it is not an active entity in the sense that it does not take any management decision based on the values of the management information. Rather, the manager is the entity responsible for analyzing the management information of a device in order to proceed with management actions, *i.e.*, management processing power is heavier at the manager side than at the agent side. The introduction of an intermediate gateway between manager and agent does not only provide the required WS to SNMP translations, but also allows balancing the processing power needed to manipulate the management information between the manager and the new introduced gateway.

In order to accomplish a WS-based management solution, the traditional SNMP manager is replaced by a WS manager, which is a WS-based client application that needs to retrieve management information from SNMP-enabled devices. A WS to SNMP gateway is placed between the new WS manager and the traditional SNMP agent in order to translate the WS manager requests to SNMP requests, and to translate back the SNMP agent replies to WS replies sent to the WS manager. In addition, WS interactions between manager and gateway are accomplished by SOAP (Simple Object Access Protocol) (Gudgin et al., 2003), which is currently the most used protocol for WS communication. Figure 2 presents the management scenario where a WS to SNMP gateway is deployed.

The creation of WS to SNMP gateways can be accomplished via different approaches, as pointed before. In the following subsections we present three main approaches. In one of them, we assume the use, at the managed device, of the Script MIB specification as an example of a more sophisticated management service. The Script

Figure 2. A WS manager contacts, via SOAP messages, a WS to SNMP gateway in order to access the SNMP agent of the end managed device



MIB (Schoenwaelder et al., 2000) is a MIB module that defines management information required to allow the transferring and remote execution of management scripts on target devices. With such MIB module a manager can send a management script to a remote device and request it to execute the transferred script. The results of the execution are then later retrieved by the manager also using the Script MIB.

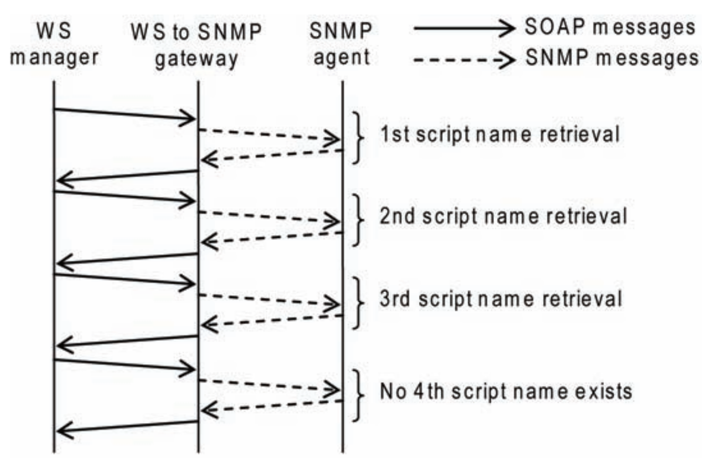
Protocol-Level Gateways

Protocol-Level gateways (Schoenwaelder et al. 2003) provide communications between WS managers and SNMP agents via protocol translations, *i.e.*, SNMP messages are directly mapped to WS operations. For example, the SNMP messages Get, GetNext, and Set are mapped to exactly other Get, GetNext, and Set WS operations. In such a mapping, most data required to build up an SNMP request is provided in input arguments of the WS gateway operations. On the way back, most data carried by an SNMP response become either the WS operation results or output parameters.

A protocol-level gateway receives SOAP requests from a WS manager and translates it to SNMP requests forwarded to the target devices. After processing the request, the target device sends back to the gateway SNMP replies that are translated to SOAP replies forwarded to the WS manager. Figure 3 shows the interactions in a scenario where a protocol-level gateway is deployed and a WS manager needs to know the name of the management scripts available in the target device. In this example, the end device hosts three management scripts. Since SNMP discovers the end of a list receiving an inadequate reply after the last element, to retrieve the name of all management scripts (*i.e.*, three names), four interactions are required.

In SNMP, each management object has an object identifier (OID), which is a sequence of integers used to address the information. For example, the smScriptName object of the Script MIB has OID “1.3.6.1.2.1.64.1.3.1.1.2.” OIDs are used by SNMP managers to inform SNMP agents about the internal data to be retrieved or modified. In the case of the protocol-level gateway, the WS manager still deals with OIDs to manage the target devices. Hence, OIDs are carried by SOAP messages from the WS manager and gateway, and by SNMP messages between gateway and SNMP agent.

Figure 3. In protocol-level gateways, each SNMP message corresponds to another SOAP message



Object-Level Gateways

Instead of translating protocol operations, object-level gateways (Neisse et al., 2004) map management information to a WS operation. For example, instead of exposing a GetNext operation, an object-level gateway for the Script MIB would expose a GetSmScriptName operation to list the scripts available for execution. The GetSmScriptName operation is a mapping of the smScriptName object defined in the Script MIB.

It is important to notice that in this case the WS manager does not need to deal with SNMP OIDs anymore. The object-level gateway implementation stores the required OIDs in order to use them once the WS manager requests a device's information. Another important point is related to the retrieval of management information. In the case of the object-level gateway the gateway itself, and not the WS manager, controls the interactions with the SNMP agent required to retrieve all data from a list, builds up the SOAP reply, and sends it back to the WS manager. This interaction control, which would be performed by the manager in the protocol-level gateway, is then moved to the object-level gateway, introducing a certain level of control on the gateway side.

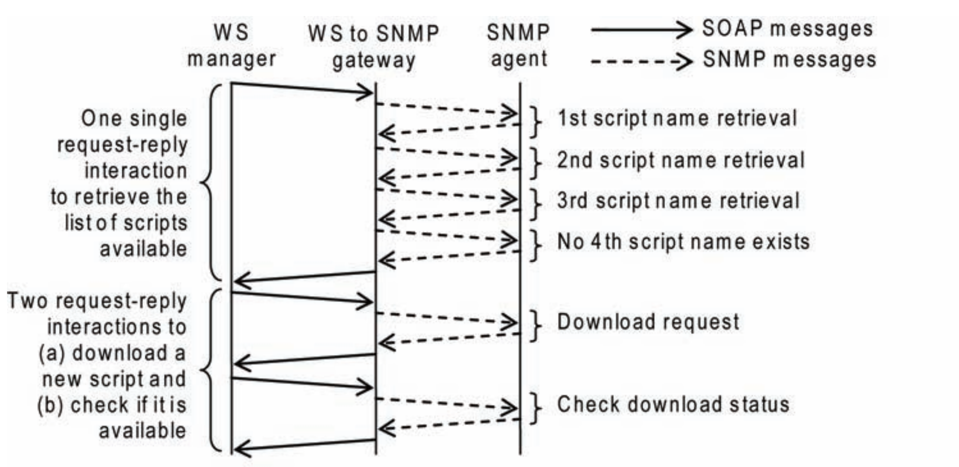
In comparison to Figure 3, Figure 4 shows that WS manager and gateway have fewer message exchanges when the object-level gateway is used. Figure 4 additionally shows the interactions needed to order the target device to download (from an external server not depicted in the figure) a management script, and associated retrieval of the download status. The example of Figure 4 deals with three different objects defined in the Script MIB: a list of scripts (smScriptName), an object to inform the management script location (smScriptSource) from where the target device must download it, and an object that reports the download status (smScriptAdminStatus).

Service-Level Gateways

As presented before, the set of management information defined in a MIB module ends up providing a management service available at target devices. Service-level gateways (Fioreze et al., 2005) are those that map the management services of a MIB module.

Let's consider again the Script MIB module, which defines a set of objects that need to be manipulated in order to transfer a management script from an external server to the managed device. Such objects need to be handled in a co-

Figure 4. In object-level gateways, one WS operation may be associated to several SNMP messages



ordinated way; otherwise the download operation may fail. Although a MIB module defines the managed information in a formal way in SMI, the management service exposed by a MIB module is defined informally in the comments clauses of each object. It means that the way and order in which each Script MIB object must be manipulated to properly download a script is informally defined in the Script MIB objects comments.

Service-level gateways are built in an empirical fashion because there is no concrete element (*e.g.*, protocol or management information) able to formally define the services exposed by a MIB module. For example, the set of Script MIB objects that supports the download request, checking, and execution of a management script could be mapped to a single WS gateway operation, as presented in Figure 5.

Although building up service-level gateways is not done on top of formal basis, its usage from the WS manager is easier because such manager does not need to know neither the managed device OIDs (as in the protocol-level gateways) nor how a set of objects needs to be manipulated in order to have a management services properly working (as in the case of the object-level gateways).

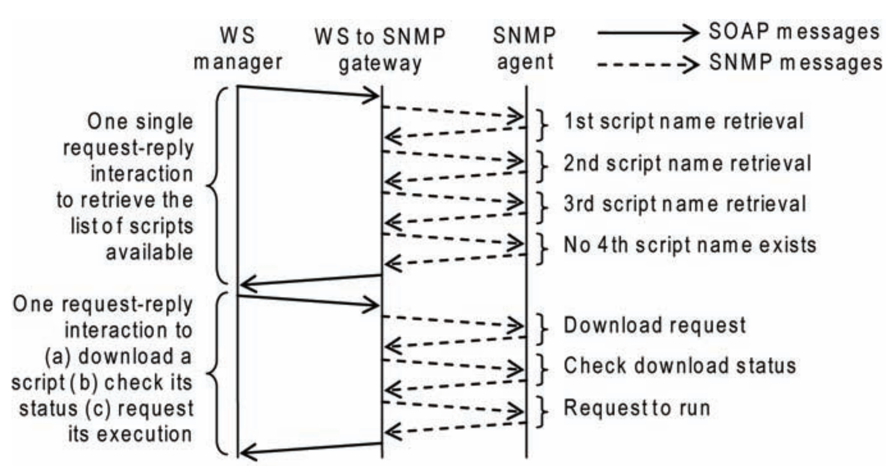
PERFORMANCE OF WEB SERVICES-BASED MANAGEMENT APPROACHES

In this section we present the performance of the three WS to SNMP gateways presented before in order to characterize them in terms of network bandwidth consumption, perceived execution time, effort to maintain, and ease of use. These parameters not only allow determining the gateways' behavior and characteristics but they also indicate the management scenarios that each gateway is more suitable for.

Consumed Network Bandwidth

Network consumption is critical for management because usually the bandwidth available for network users' traffic is shared with management traffic. If management traffic grows too high, less bandwidth is available to users, which is obviously inadequate. SNMP is a quite light protocol considering the bandwidth consumed to carry the management information. SOAP, however, has messages longer than SNMP because SOAP is based on XML. This difference could lead to the conclusion that WS manager and gateway inter-

Figure 5. In service-level gateways, one WS operation may be associated to several SNMP interactions and objects in order to have a management service accomplished



actions consume more bandwidth than gateway and SNMP agent interactions. That is true for the protocol-level gateways, where each SNMP message corresponds to exactly another SOAP message, but it is not always true for object and service-level gateways.

Let's observe the bandwidth consumed by the three approaches considering two management scenarios. In the first one, a WS manager interacts, in different moments, with a protocol-level and an object-level gateway in order to retrieve a list of integer objects defined only for the purpose of this test, *i.e.*, the meaning of the returned integers is not important, but the bandwidth consumed. We start with a single one integer object and progressively increase the testing list until 70 single integer objects are retrieved, and measure the network usage between the WS manager and the gateway, and between the gateway and the SNMP agent. The testing network in this first scenario is identical to the one presented in Figure 2, *i.e.*, protocol-level and object-level gateways are dedicated machines placed between manager and agent.

The graph from Figure 6 presents the network usage for the SNMP traffic and SOAP traffic running over HTTP and HTTPS (for the case of required security support). Since intuitively SOAP would consume more bandwidth than SNMP, compressed SOAP traffic has been also observed to check how much the bandwidth consumption could be decreased. As shown in the graph, SNMP traffic will always consume far less bandwidth than SOAP, even if SOAP compression is present. This situation gets even worse when the number of objects to be retrieved increase: since each SOAP message consumes more bandwidth than the associated SNMP message, the higher the number of messages, the greater the difference between SOAP and SNMP.

By replacing the protocol-level gateway for an object-level gateway, we have the bandwidth consumption presented in Figure 7. In this case, although initially SOAP traffic (over HTTP or HTTPS, and compressed and uncompressed) consumes more bandwidth than SNMP, if more objects are retrieved from the managed device SOAP traffic will eventually consumes less bandwidth

Figure 6. SNMP traffic always consumes less bandwidth than protocol-level gateway traffic, even when SOAP compression is used

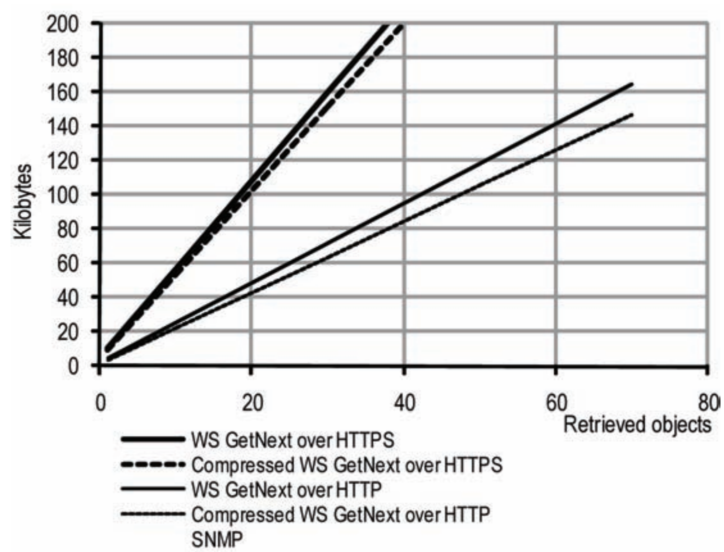


Figure 7. Since the object-level gateway can group SNMP retrieved information in a single SOAP reply, at some points total SOAP traffic will consume less bandwidth than SNMP traffic

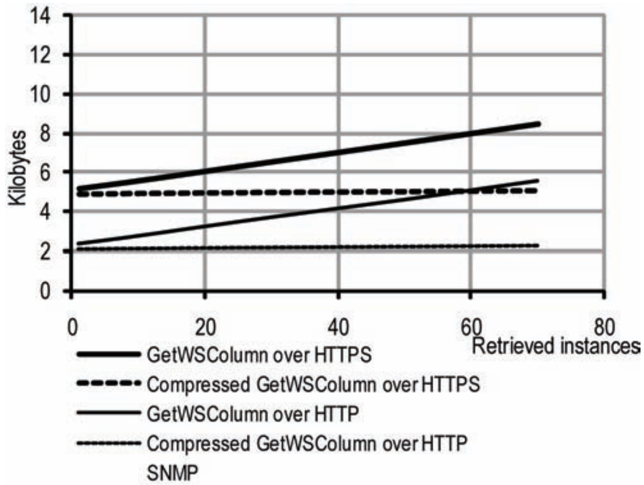
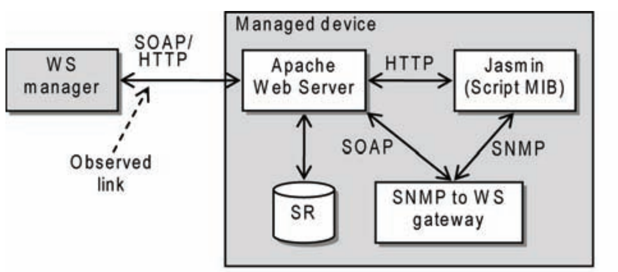


Figure 8. Scenario for the object-level and service-level gateways



than SNMP. That happens because the object-level gateway groups all SNMP retrieved information from the target device in a single SOAP reply. Thus, there is only one SOAP request and associated reply, while several SNMP interactions are present, as has been depicted in Figure 4.

By reviewing the graph from Figure 7 it is possible to conclude, for this first scenario, that uncompressed SOAP over HTTPS traffic consumes less bandwidth than SNMP if a list with more than 38 integers needs to be retrieved. Compressed SOAP over HTTPS traffic is "better" than SNMP if more than 26 integers are retrieved. Uncompressed SOAP over HTTP traffic consumes less bandwidth than SNMP after 17 integers, and compressed SOAP over HTTP after 11 integers. It is interesting to notice that for more than 68 integers, compressed SOAP over HTTPS traffic consumes less bandwidth than uncompressed SOAP over HTTP. In general, the introduction of compressing makes the SOAP traffic line angle to slow down, while the introduction of security support via HTTPS makes the SOAP traffic line initiate at a higher value.

In order to check the bandwidth consumed by service-level gateways, we put them in perspective with object-level gateways in a second scenario, where we turn back to our previous Script MIB cases. In this scenario, presented in Figure 8, we placed both gateways and SNMP agent inside the same target device. Although this configuration does not affect the bandwidth evaluation, it has important impact on the response time, to be presented in the next subsection.

In this scenario, the the WS to SNMP gateway is coded using the nuSOAP library for PHP. The target device is a Linux-based host running the Apache Web server, a local management script repository (SR), and the Jasmin tool. Jasmin (Braunschweig, 2003) is an implementation of the Script MIB developed by the Technical University of Braunschweig and NEC C&C Research Laboratories. Jasmin implements the Script MIB published in the RFC 2592, which was later updated by the RFC 3165. Jasmin supports the Java and TCL runtime engines, so that network managers can delegate Java and TCL management scripts to the Jasmin-enabled end device.

The interactions on this scenarios works like this. The WS manager builds up an SOAP/HTTP request sent to the target device's Apache server, which removes the HTTP information and delivers the SOAP request to the internal WS to SNMP gateway. The gateway contacts, using SNMP, the Jasmin agent to request Script MIB operations. One operation is the download of a management script from the Jasmin agent. Since our script repository (SR) is internal to the local device, once Jasmin is required to download a script it locally contacts the Apache Web server in order to download, via HTTP, a management script from the SR.

In this scenario the following management actions are executed:

- The manager orders the SNMP agent to download a management script named wait.jar;
- The manager blocks until the SNMP agent finishes the script download from the local repository SR;
- The manager blocks again until wait.jar is ready to run;
- The manager requests the execution of wait.jar passing as input argument the number of seconds wait.jar must wait until its end (0, 5, 10, and 15 seconds are considered);
- The manager blocks again until the execution of wait.jar is over;
- Finally, the manager retrieves the execution results from the target device.

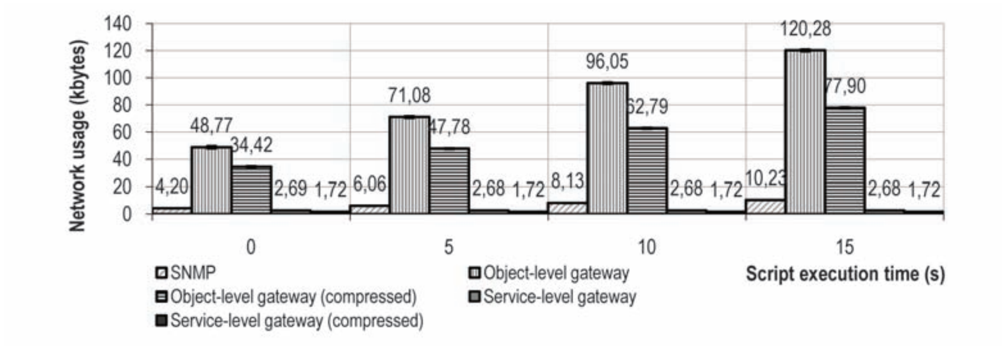
Although not explicitly shown on Figure 8, two different managers are in fact considered: a WS manager and an SNMP manager. The WS manager interacts with the WS to SNMP gateway, while the SNMP manager interacts directly with the Jasmin agent, skipping the WS to SNMP gateway. With these two managers it is possible to compare the bandwidth consumed by SOAP and SNMP to accomplish the management actions defined previously.

One important difference distinguishes the object-level and service-level gateways: in the object-level gateway the WS manager controls each step, while in the service-level gateway the gateway itself controls all steps and the WS manager just needs to order the executions of the steps to the gateway. If instead of using a WS manager one uses an SNMP manager, then the SNMP manager will behave like the WS manager in the object-level gateway, since the SNMP manager will be the one responsible for controlling the execution of the management steps.

In the second scenario, only compressed and non compressed HTTP traffic is observed; HTTPS is not considered because, in comparison, compression has a stronger impact than HTTPS on the final bandwidth consumed. Figure 9 shows the bandwidth consumed by SNMP in a direct SNMP manager-agent interaction, and by WS manager and WS to SNMP object-level and service-level gateways.

According to the previous testing steps, the (WS or SNMP) manager blocks in three moments: when it waits for the wait.jar script to be downloaded, when it waits for the script to be

Figure 9. Object-level gateway, for different actions, consumes more bandwidth than SNMP, while service-level gateway, that aggregates SNMP information in a single SOAP request-reply interaction, consumes less bandwidth



ready to run, and when it waits for the script to finish its execution. To implement such blocking, the SNMP manager for the Script MIB and WS manager for the object-level gateway loop until they get a proper confirmation from the target devices indicating that the script is downloaded, ready to run, or finished.

From Figure 9, SNMP consumes far less bandwidth than the object-level gateway, but more bandwidth than the service-level gateway. In the case of the service-level gateway, no loop is required in the manager because the gateway itself blocks until the script results are available. Thus, the testing loop has been moved from the WS manager to the service-level gateway, then consuming less bandwidth. Figure 9 also shows the bandwidth consumed when the SOAP/HTTP traffic is compressed. As observed, and intuitively believed, less network resources are consumed when compression is used because compressed SOAP/HTTP messages are smaller than the uncompressed ones.

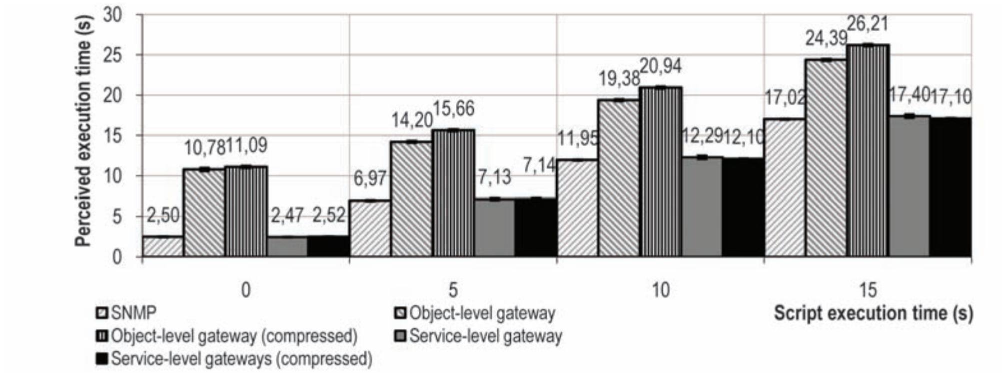
It is important to note that although object-level gateway consumes less bandwidth than SNMP in the first evaluation scenario, but more bandwidth than SNMP in the second scenario, this difference is not inconsistent. In the first scenario the object-level aggregates SNMP information in a single SOAP reply because the information being retrieved was a list of elements associated to the same object (a testing list of integers), while in the second scenario the SNMP information (objects) was different from each other, preventing the object-level gateway to aggregate them. That aggregation, however, was possible in the service-level gateway, which then consumed less bandwidth.

Perceived Execution Time

The perceived execution time is defined as the time when a manager perceives that a management task is over. That is different to the real execution time, since a management task may finish at a time, but perceived as so by the manager at a later instant. Perceived execution time is an important metric because managers take management actions normally based on the results of previous actions. If the perceived execution time is high, the management actions are delayed in the management station. Ideally, the perceived execution time should be as close as possible to the real execution time.

To check the perceived execution time, we have used the second evaluation scenario presented in the bandwidth evaluation. We have measured the execution time to download, execute, and retrieve the results of the wait.jar management

Figure 10. The execution time of actions executed through object-level gateways (with or without compression) take longer to be perceived than in the case of SNMP, while in for service-level gateway the perceived execution time is similar to SNMP



script using, again, an SNMP manager, and two WS managers for the object-level and service-level gateways. Figure 10 presents the measured perceived execution times.

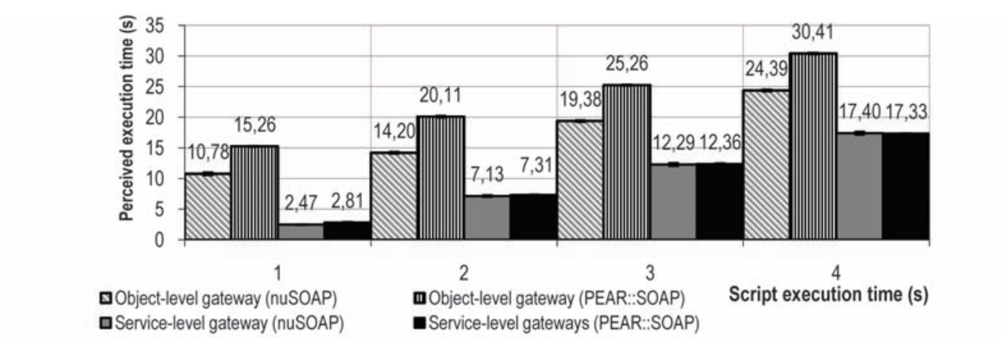
One aspect that influence in the perceived execution time is the set of software libraries used in the management system. As pointed before, we have been using the nuSOAP library to support Web Services in the PHP scripting language. In order to compare how SOAP libraries impact in the perceived execution time, we have re-implemented the service-level gateway now replacing nuSOAP by PEAR::SOAP, which is a second library for SOAP support in PHP. Figure 11 presents the results on using service-level gateways implemented using nuSOAP and PEAR::SOAP.

As can be observed, PEAR::SOAP tends to impose a greater perceived execution time, especially in the case of object-level gateways. This is so because PEAR::SOAP has a processing delay greater than nuSOAP, then taking more time to encode and deliver the SOAP/HTTP messages to the WS manager.

Maintenance Effort and Ease of Use

WS to SNMP gateways need to be maintained in order to properly follow the changes in the target device MIB structure. Gateways need also to be updated if they are supposed to access new

Figure 11. Generally, PEAR::SOAP, compared with nuSOAP, results in a greater perceived execution response time



target devices with different MIBs than the MIBs of previous managed devices. In addition, different gateways have different ease of use from the WS manager perspective, *i.e.*, WS managers face different level of easiness in dealing with different gateways. In this subsection we check the maintenance effort and the ease of use of the three WS to SNMP gateway approaches. Concerning the ease of use aspect, we take the approach of observing the gateways from a WS manager perspective, which is different than the perspective of an SNMP manager, since SNMP managers are aware of the SNMP technology, but WS managers are not.

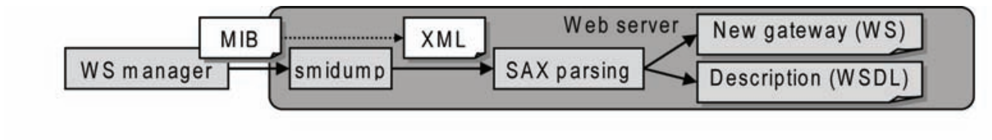
In the standard SNMP framework, managers require maintenance effort in order to keep in tune with the management information of target devices. SNMP manager maintenance is accomplished informing the manager about new management information using the MIB modules that describe this new information. Usually, when the manager needs to support a new device, the network human operator downloads, from the devices manufacturer Web site, MIB module files that specify the information available at the target device, and installs these modules in the manager. From this installation on, the SNMP manager is aware of the available information and can properly access and manipulate it on the target device.

Protocol-level gateway is not aware about the information being requested to the target device: it only translates requests from SOAP to SNMP and back. In this way, if new management information is available at the target device, actually the WS manager needs to be updated, not the protocol-level gateway. Thus, once built, protocol-level gateways require no maintenance effort, considering the changes in the MIB of target devices. On the other hand, protocol-level gateways are not easy to use to WS manager. Protocol-level gateways expose operations (*e.g.*, Get, GetNext, Set) that forces the WS manager to known the SNMP information peculiarities in order to deal with them. Thus, WS managers using protocol-level gateways are forced to be SNMP aware, which is not simple. Since dealing with SNMP information (*e.g.*, OIDs) is complex, protocol-level gateways are not easy to use too.

Object-level gateways, concerning their maintenance effort, do require to be updated as soon as new information is available on target devices. This is so because the set of management information defines the WS operations exposed by the object-level gateway. If new information is included in the target device, but the gateway is not updated, the information will be unavailable for the WS manager. In this case, the WS manager will be able to access only the information that has a corresponding WS operation on the object-level gateway.

Although object-level gateways require maintenance efforts, this can be automated via supporting software able to parse and process MIB module files and generate associated gateways. Since MIB modules are defined using the SMI precise rules, one can access its content and produce outputs results convenient for specific applications, that in our case corresponds to the creation or update of object-level gateways. Figure 12 presents the architecture of an object-level gateway building system developed for this investigation. The first step in a gateway creation is the transfer of a MIB module from the WS manager to the Web server trough HTTP or HTTPS. Internally on the server, the smidump tool checks the received MIB module and, if no inconsistencies are found, it generates an XML temporary document. The smidump tool (Strauss, 2003), developed at the Technical University of Braunschweig, is a whole MIB module checking and manipulating system able to translate MIB modules to other representations, including XML. The next SAX (Simple API for XML) parsing step, in turn, takes the XML temporary document to build up the new gateway. Each object in the original MIB module is transformed in operations of the new generated WS. These operations are instrumented with code

Figure 12. The object-level gateway building system receives a MIB module files and generates, after some intermediate steps, a new gateway and corresponding WSDL document



able to contact, via SNMP, a target device. The just created gateway is then stored in a standard directory in the Web server and available to be invoked just after its creation. At the same time the parsing step creates the code for the new WS, it also builds the WSDL document that describes the created WS. The WS-based manager that requested the creation of the new object-level gateway optionally informs the URL of a UDDI repository where the created gateway is registered. The original MIB module file provided by the WS manager is also stored in another standard directory for documentation purpose, as well as the intermediate XML document generated by the smidump tool prior to the parsing step.

With this gateway building process, new MIB modules can be easily added to a WS-based management environment. In order to let this process works properly, however, the original MIB module files should be correctly defined. It is not rare to find MIB module files that have definition problems (mainly old MIB files, since more recent MIB modules seem to be more properly defined). For example, according to the output of the smilint tool, the RFC1213-MIB module has 3 errors (Fenner, 2003). In the case of finding MIB definition problems, the intended WS to SNMP gateway will not be created, and the corresponding smidump message describing the errors found will be sent back to the WS-based manager. It is important to notice that the gateway building system itself is not a WS, but a Web application running at the same Web server that will host the gateways just created.

In terms of ease of use, object-level gateways are more interesting than protocol-level gateways. Since SNMP management information is mapped to WS operations, SNMP details are hidden by the object-level gateway from the WS manager. Although still dealing with management information at some level, WS managers do not need to know, for example, how to manipulate SNMP OIDs in order to retrieve data from the target devices. Given that the object-level building process presented before generates an associated WSDL document, WS managers can be aware of the new management information whenever the target device MIB is updated (and an associated gateway built). The important point is that the WS manager is not forced to deal with MIB modules files, *i.e.*, WS managers do not need to be SNMP aware anymore.

Service-level gateways, compared with protocol-level and object-level gateways, require far more efforts to be maintained. As presented before, service-level gateways are not built upon formal basis (protocol or MIB module definition), but rather based on the services exposed by a MIB module which are defined in the MIB module comment clauses. It means that for every new MIB module available that requires a service-level gateway, such gateway will be developed manually checking the MIB module content to produce a set of WS operations. The developer (or the designer) of service-level gateways is the one responsible for mapping the original MIB module into the gateway operations. That means that two different developers (or designers) will probably end up with different gateways for the same original MIB module.

Since the services are not formally defined, there is no current tool able to, given a MIB module, automatically produce a new service-level gateway, as it was the case for object-level gateways presented before. Thus, for each new MIB module available, the developer has to built the associated service-level gateways, which is a slower process if we compare with the object-level gateways (that are automatically created via a building system) and protocol-level gateways (that need to be implemented just once, regardless the changes in the MIBs of target managed devices).

However, regarding ease of use, service-level gateways are quite interesting because besides hiding SNMP details from the WS manager (as object-level gateways do), service-level gateways also hide the steps required to accomplish a specific management service defined in a MIB module. While in the object-level gateway the WS manager needs to know what WS operations must be called and in which order, in the service-level gateway the WS manager only needs to invoke a WS operation that corresponds to a whole management service. Surely, this ease of use depends on the quality of the mapping of the MIB service to the service-level gateway, but in general it is safe to say that even for bad mappings the service exposed by a service-level gateway is easier to use than the corresponding whole set of operations exposed by an object-level gateway.

In order to summarize the evaluation of the WS to SNMP gateways, table 1 puts the three approaches in perspective regarding the evaluation parameters consumed bandwidth, perceived execution time, maintenance effort, and ease of use.

CONCLUSION

In this chapter we have evaluated three different approaches for building Web Services (WS) to SNMP gateways. Such gateways are required for the integration of traditional network devices into WS-based systems, such as systems for e-commerce and e-government. Protocol-level gateways directly map SNMP messages to SOAP messages, while object-level gateways hide SNMP details from WS managers offering WS operations mapped from MIB modules. Finally, service-level gateways go further and expose management services instead of management information (like object-level gateways) or protocol mapping (like protocol-level gateways). However, service-level gateways are more difficult to maintain because they require the human interpretation of MIB modules to build up a new version of the gateways when the set of management information is changed. Object-level gateways, on the other hand, can be automatically built given a MIB module as input argument, while protocol-level gateways do not require maintenance efforts at all because their operation is not affected by changes in the management information available at the target devices.

Protocol-level gateways are interesting in

Table 1. Summary on WS to SNMP gateways evolution

Gateway	Bandwidth	Execution time	Maintenance	Ease of use
Protocol-level	Gateway with highest bandwidth consumption	Takes more time than SNMP to learn that a task is over	After being built, requires no maintenance at all	Difficult to use because forces WS managers to be SNMP aware
Object-level	Consumes less bandwidth than SNMP for a large number of objects	Takes far more time than SNMP to learn that a task is over	Easy to maintain given a gateway building system based on MIB modules	Hides SNMP details, but WS managers need to know how to use the service
Service-level	Consumes less bandwidth than SNMP and other gateways	Learns faster than SNMP that a task is over	Difficult to maintain and requires human interpretation of MIB services	Easy to use because service details are hidden from the WS manager

management environment where the set of management information changes quite frequently. Although they consume far more bandwidth than SNMP, protocol-level gateways may enable the communication of manager and agents located in different administrative domains because SOAP traffic tends to cross Internet firewalls easier than SNMP. However, the WS manager must be aware of the SNMP peculiarities in order to interact with the protocol-level gateway, which is an important restriction when dealing with systems of broader disciplines such as e-business.

Object-level gateways are ideal when the amount of information to be retrieved from managed devices is high. Since these gateways are able to group SNMP information in a single SOAP reply, object-level gateways tend to consume less bandwidth than SNMP for a large number of information. That enables protocol-level gateways to be used, for instance, in monitoring system composed of measurement elements deployed in networks located in different administrative domains, and bunches of collected data needs to be transferred from time to time to a central Web Services based analysis tool. Since the maintenance and ease of use of object-level gateways are not complex, the availability of new different management information on the managed network does not lead to complex updates of gateways and WS managers.

Finally, service-level gateways, besides consuming less bandwidth than SNMP and presenting response time close enough to SNMP, has the advantage of being easier to use. Service-level gateways are ideal for integrating SNMP into non network management systems, because these gateways not only hides the SNMP complexities but also exposes management services in a simpler way than the previous gateways. If we do not consider the maintenance effort, the results presented in this chapter help us to argue in favor of a service-oriented SNMP/WS integration, instead of information or protocol-oriented integration. The presented service-level gateway is easier to use (from the manager perspective), consumes less bandwidth than the other gateways, and presents a response time quite close to SNMP. It is important to highlight that these conclusions are based on the case of the Script MIB. However, although the SNMP framework is well known and accepted, designing WS management solutions taking the SNMP framework as a starting point is not a proper approach in modern times. Rather, services must be considered in the first place and SNMP "details" should be treated only when the protocol and related information issues begin to be relevant.

A WS feature that is often cited is that more complex and richer WS can be built from the combination and reuse of other simpler WS. This feature could be used to integrate network management services, or, more interestingly, to create new services that could not be easily built today except through heavy coding efforts. Integrating simpler WS to form richer network management WS is one of the interesting challenges to be addressed. Web Services coordination, orchestration, and choreography applied to the network management field are certainly aspects that must be investigated and considered in the network integration and management field. This is a feasible way to achieve a real and long awaited integration between network management and other Web Services-based disciplines.

REFERENCES

Aschemann, G., Mohr, T., & Ruppert, M. (1999). Integration of SNMP into a CORBA and Web-based Management Environment. In Steinmetz (Ed.), *KiVS'99* (pp. 210-221).

Braunschweig, T., & Europe, N. C. (2003). *Jasmin: A Script-MIB Implementation*. From http://www.ibr.cs.tu-bs.de/projects/jasmin

Case, J., Fedor, M., Schoffstall, M., & Davin, J. (1990). A Simple Network Management Protocol (SNMP). *Internet Engineering Task Force, Request for Comments: 1157 (STD: 15)*.

Drevers, T., va de Meent, R., & Pras, A. (2004). On the standardization of Web Services management operations. In *10th Open European Summer School and IFIP WG6.3 Workshop* (pp. 143-150).

Fenner, B. (2003). *Smilint output for all MIBs published in RFCs and Internet Drafts*. From http://www.icir.org/fenner/mibs/htmllint

Fioreze, T., Granville, L. Z., Almeida, M. J., & Tarouco, L. R. (2005). Comparing Web Services with SNMP in a Management by Delegation Environment. In *IFIP/IEEE International Symposium on Integrated Network Management* (pp. 601-614).

Gudgin, M., Hadley, M., Mendelsohn, N., Moreau, J.-J., & Nielsen, H. F. (2003). SOAP Version 1.2 Part 1: Messaging Framework. *W3C Recommendation*, from http://www.w3.org/TR/2003/REC-soap12-part1-20030624/

Harrington, D., Presuhn, R., & Wijnen, B. (2002). An Architecture for Describing Simple Network Management Protocol (SNMP) Management Frameworks. *Internet Engineering Task Force, Rquest for Comments: 3411 (STD: 62).*

ISO/IEC 9596 (1991). *Common Management Information Protocol Specification*.

Mccloghrie, K., Perkins, D., Schoenwaelder, J., Case, J., Rose, M., & Waldbusser, S. (1999). Structure of Management Information Version 2 (SMIv2). *Internet Engineering Task Force, Request for Comments: 2578 (STD: 58).*

Neisse, R., Granville, L. Z., Almeida, M. J., & Tarouco, L. (2003). A Dynamic SNMP to XML Proxy Solution. In *8th IFIP/IEEE International Symposium on Integrated Network Management* (pp 481-484).

Neisse, R., Vianna, R. L., Granville, L. Z., Almeida, M. J. B., & Tarouco, L. M. R. (2004). Implementation and Bandwidth Consumption Evaluation of SNMP to Web Sevices Gateways. In *IEEE/IFIP Network Operations and Management Symposium* (pp. 715-728).

Oh, Y. J., Ju, H. T., Choi, M. J., & Hong, J. W. (2002). Interaction Translation Methods for XML/SNMP Gateway. In *13h IFIP/IEEE International Workshop on Distributed Systems: Operations and Management* (pp. 54-65).

Orfali, R., & Harkey, D. (1998). *Client/Server Programming with Java and CORBA*. John Wiley and Sons, New York, USA, 2nd edition.

OSI (1987). Specification of Abstract Syntax Notation One (ASN.1). *International Standard* 8824.

Pavlou, G., Flegkas, P., Gouveris, S., & Liotta, A. (2004). On Management Technologies and the Potential of Web Services. *IEEE Communications, 42*(7).

Presuhn, R., Case, J., Mccloghrie, K., Rose, M., & Waldbusser, S. (2002). Version 2 of the Protocol Operations for the Simple Network Management Protocol (SNMP). *Internet Engineering Task Force. Request for Comments: 3416 (STD: 62).*

Saydam, T., & Sirsikar, R. (1998). Design of CMIP-SNMPv2 Proxy Gateway for Network Management Interoperability. *Journal of Network and Systems Management*, *6*(2). doi:10.1023/A:1018758725407

Schoenwaelder, J., Quittek, J., & Kappler, C. (2000). Building Distributed Management Applications with the IETF Script MIB. *IEEE Journal on Selected Areas in Communications*, *18*(5), 702–714. doi:10.1109/49.842986

Schönwälder, J., Pras, A., & Martin-Flatin, J. P. (2003). On the Future of Internet Management Technologies. *IEEE Communications Magazine*, *41*(10), 90–97. doi:10.1109/MCOM.2003.1235600

Strauss, F. (2003). *libsmi - A Library to Access SMI MIB Information*. From http://www.ibr.cs.tu-bs.de/projects/libsmi/

Strauss, F., & Klie, T. (2003). Towards XML Oriented Internet Management. In *8th IFIP/IEEE International Symposium on Integrated Network Management* (pp. 505-518).

KEY TERMS AND DEFINITIONS

Integrated Network Management: The study of integrating network management operations among each other or with other external disciplines.

Network Management: A computer science discipline that investigates activities, methods, procedures, and tools to operate, administrate, maintain, and provide networking infrastructures.

Protocol Evaluation: The observation of a network protocol in terms of its behavior and the impact associated to it over the communication infrastructure.

Protocol Gateway: A software entity able to translate protocol messages, operations, and/or services from one specific protocol to another specific protocol.

SNMP: Stands for *Simple Network Management Protocol* and is the *de facto* standard network management protocol for TCP/IP networks.

This work was previously published in Handbook of Research on Telecommunications Planning and Management for Business, edited by I. Lee, pp. 724-741, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 6.7
Web Services, Service-Oriented Computing, and Service-Oriented Architecture: Separating Hype from Reality

John Erickson
University of Nebraska - Omaha, USA

Keng Siau
University of Nebraska - Lincoln, USA

ABSTRACT

Service-oriented architecture (SOA), Web services, and service-oriented computing (SOC) have become the buzz words of the day for many in the business world. It seems that virtually every company has implemented, is in the midst of implementing, or is seriously considering SOA projects, Web services projects, or service-oriented computing. A problem many organizations face when entering the SOA world is that there are nearly as many definitions of SOA as there are organizations adopting it. Further complicating the issue is an unclear picture of the value added from adopting the SOA or Web services paradigm. This article attempts to shed some light on the definition of SOA and the difficulties of assessing the value of SOA or Web services via return on investment (ROI) or nontraditional approaches, examines the scant body of evidence empirical that exists on the topic of SOA, and highlights potential research directions in the area.

INTRODUCTION

Service-oriented architecture (SOA); Web services; mash-ups; Ajax; Web 2.0; some of their underlying middleware realization schemas such as SOAP (simple object access protocol), UDDI (universal description, discovery, and integration), XML (extensible markup language), and CORBA (common object request broker architecture); and many other ideas or approaches to cutting-edge information system architectures have become the buzzwords of the day for many in the business world and also in the IT and IS communities. It is quite difficult, perhaps nearly impossible, to pick up any relatively current practitioner publication without encountering an article focusing on at least one of the above topics. A recent library database search using keywords *service-oriented architecture, Web services*, and *SOA* resulted in 800-plus returns. Further investigation revealed that roughly 25 of those 800 articles were sourced in research journals while the other (still roughly 800) articles were all from more practitioner-oriented sources.

When it comes to adopting and implementing SOA, it appears that businesses are doing it at astounding rates. Of course, what they are actually doing, even though they may say that their efforts represent a move toward service-oriented architecture, may not match anyone else's definition of SOA but their own. Furthermore, how can SOA be defined, and how can we define the benefits of moving toward such architectures? It seems that there is little agreement among practitioners and researchers alike as to a standard definition of SOA.

Worse still, a growing number of practitioners are now beginning to question the business return of some of the approaches. For example, Dorman (2007), Havenstein (2006), Ricadela (2006), and Trembly (2007) indicate that there is doubt emerging as to the real value of SOA to adopting businesses and organizations. Perhaps the question of return on investment (ROI) should not be that surprising since it sometimes seems that each organization has its own definition of what SOA really is.

This article attempts to reach for a clearer understanding of what SOA really is, and proposes some possible areas of research into SOA that could help clear up some of the definitional confusion, which could in turn help lead to better understanding of ROI as it relates to SOA. First is the introduction. Second, the article provides existing definitions of SOA, Web services, and some of the related and underlying technologies and protocols. The next section combines the various definitions of SOA into a more coherent form, while the section after that proposes ideas about what SOA should be. The fifth section discusses research possibilities and provides recommendations for future research efforts. Next, we look at ways of measuring and justifying SOA and SOC (service-oriented computing) success. Finally, we conclude the article.

BACKGROUND AND HISTORY OF SERVICE-ORIENTED ARCHITECTURE

A minimum of nine formal definitions of SOA exist as of this writing, from sources such as the Organization for the Advancement of Structured Information Standards (OASIS), the Open Group, XML.com, Javaworld.com, Object Management Group (OMG), the World Wide Web Consortium (W3C), Webopedia, TechEncyclopedia, WhatIs.com, and Webopedia.org. In addition, many other definitions put forth by numerous industry experts, such as those from IBM, further cloud the issue, and worse yet, other formal definitions might also exist. In other words, the concept of service-oriented architecture appears in many ways to be a virtually content-free description of an IT-based architecture. It is not our intent here to add yet another definition to this already crowded arena of definitions, but to try to cull the common, base meanings from the various distinct definitions.

Prior to about 2003, the term service-oriented architecture was not in general use for the most part, according to Wikipedia ("SOA," 2007). However, since that time, SOA has exploded nearly everywhere in the business and technology world. SOA appears to derive or develop in many cases from more basic Web services. These services can include enabling technologies such as SOAP, CORBA, EJB (Enterprise Java Beans), DCOM (distributed component object model), and even

SIP (session-initiated protocol) among many others; services may also include other middleware created with XML (Lee, Siau, & Hong, 2003; Siau & Tian, 2004; Sulkin, 2007; Walker, 2007).

Service-Oriented Architecture Definitions

The Open Group (2007) defines SOA as “an architectural style that supports service orientation.” The definition goes on to also include descriptions of architectural style, service orientation, service, and salient features of SOA. OASIS defines SOA as “a paradigm for organizing and utilizing distributed capabilities that may be under the control of different ownership domains.” The OASIS definition includes what they call a “reference model” in which the details of the definition are expanded and formalized. The Object Management Group (2007) defines SOA as “an architectural style for a community of providers and consumers of services to achieve mutual value.” OMG adds that SOA allows technical independence among the community members, specifies the standards that the (community) members must agree to adhere to, provides business and process value to the (community) members, and “allows for a variety of technologies to facilitate (community) interactions” (OMG, 2007).

W3C (2007) defines SOA as “a form of distributed systems architecture that is typically characterized by…a logical view, a message orientation, a description orientation, granularity and platform neutrality.” W3C adds details describing what it means by logical view, message and description orientations, granularity, and platform neutrality. XML.com (2007) defines SOA as follows:

SOA is an architectural style whose goal is to achieve loose coupling among interacting software agents. A service is a unit of work done by a service provider to achieve desired end results for a service consumer. Both provider and consumer are roles played by software agents on behalf of their owners.

The Javaworld.com SOA definition, composed by Raghu Kodali (2005), is as follows: “Service-oriented architecture (SOA) is an evolution of distributed computing based on the request/reply design paradigm for synchronous and asynchronous applications.” Kodali also goes on to describe four characteristics of SOA. First, the interfaces composed in XML, using WSDL (Web services description language), are used for self-description. Second, XML schema called XSD should be used for messaging. Third, a UDDI-based registry maintains a list of the services provided. Finally, each service must maintain a level of quality defined for it via a QoS (quality of service) security requirement.

Finally, IBM proposes that SOA “describes a style of architecture that treats software components as a set of services” (UNL-IBM System in Global Innovation Hub, 2007). Furthermore, it insists that business needs should “drive definition” of the services, and that the value proposition be centered on the reusability and flexibility of the defined services.

SERVICE-ORIENTED ARCHITECTURE

We begin the SOA discussion with an overview of SOA provided by Krafzig, Banke, and Slama (2005). They proposed a three-level hierarchical perspective on SOA in which Level 1 includes the application front end, the service, the service repository, and the service bus (SB). Accordingly, only the service child has children, consisting of the contract, implementation, and interface. Finally, the last level of the proposed hierarchy is composed of business logic and data, children of implementation. The next subsections will discuss the general ideas of the elements included in the hierarchy proposed by Krafzig et al. described previously. This is not to recommend adoption of the hierarchy and description as the final description of SOA, but rather as a framework for discussing the meaning of SOA for the remainder of this article.

Application Front End

This part of SOA comprises a source-code interface, and in SOA terminology, it is referred to as the application programming interface (API). In accordance with most commonly accepted design principles, the underlying service requests, brokerage (negotiation), and provision should be transparent to the end user.

Service Repository

The service repository could be thought of as the library of services offered by a particular SOA. This would likely consist of an internal system that describes the services, and provides the means in the user interface to call a particular service. UDDI could be seen as a realization of the service repository idea. UDDI is a global registry that allows businesses to list themselves on the Internet. UDDI is platform independent and XML based. The point of UDDI is for businesses to list the Web or SOA-type services that they provide so that other companies searching for such services can more easily locate and arrange to use them.

Service Bus

The SB, more commonly referred to as the enterprise service bus (ESB), provides a transportation pathway between the data and the end-user application interface. Using an ESB does not necessarily mean SOA is being implemented, but ESB or some sort of SB use is almost always part of an SOA deployment. According to Hicks (n.d.), Oracle's idea of an ESB includes multiple protocols that "separate integration concerns from applications and logic." What this means is that ESBs have now become commercialized, and can be licensed for use much like other UDDI-based services. So, companies searching for ESB solutions as part of an SOA effort now have multiple choices and do not necessarily have to re-create the wheel by building their own ESB.

Common Services

It seems apparent from many of the SOA definitions that many of the technologies included in an SOA definition, and by default SOA implementations, are established and conventional protocols. To better understand the services provided in many SOA definitions, a brief explanation of some of the more commonly used underlying technologies is provided. A particular service may or may not be explicitly Web based, but in the end it matters little since the services provided by the architecture should be transparently designed, implemented, and provided. The general consensus from most involved in Web services is that the services are meant to be modular. This means that no single document encompasses all of them, and furthermore, that the specifications are multiple and (more or less) dynamic. This results in a small number of core specifications. Those core services can be enhanced or supported by other services as "the circumstances and choice of technology dictate" ("Web Service," 2007).

XML allows users to define and specify the tags used to capture and exchange data, typically between distinct and usually incompatible systems from different companies or organizations. This means that XML is a good example of middleware; it also means that XML enables Web services. XML was one of the initial drivers that provided the ability to conduct e-business for many businesses in the Internet era. XML cannot really be considered a service, but as the language used to write many of the Web services or service stack protocols.

SOAP, like all protocols, consists of a set list of instructions detailing the action(s) to be taken in a given circumstance. SOAP is designed to call, access, and execute objects. The original SOAP was typically for communications between computers, and usually involved XML-based messages.

SOAP and its underlying XML programming comprised one of the first Web service communication stacks. One of the original Web services that SOAP provided was called remote procedure call (RPC), which allowed a remote computer to call a procedure from another computer or network. More recently, SOAP has taken on a somewhat modified meaning so that the acronym now means service-oriented architecture protocol. In both cases, what SOAP does is to use existing communications protocols to provide its services. The more common early SOAP contracts included XML applications written for HTTP (hypertext transfer protocol), HTTPS (HTTP over secure socket layer), and SMTP (simple mail transfer protocol), among others. It should be apparent from these that many early SOAP implementations involved e-commerce or e-business applications, which means that the concern at the time when many applications were first developed was to move sales and other data collected in Web portals to back-end data stores.

CORBA is an OMG-developed standard that allows different software components that are usually written in different languages and installed on different computers to work together (Zhao & Siau, 2007). CORBA was developed in the early 1990s, and while not overtly an SOA at the time, it actually performs many of the functions in an SOA, using an IIOP- (Internet inter-orb protocol) based service stack.

EJB is a component typically situated on the server that "encapsulates the business logic of an application" ("EJB," 2007). EJB enables the creation of modular enterprise (and other) applications. The intent of EJB is to facilitate the creation of middleware that acts as a go-between tying front-end applications to back-end applications or data sources.

SIP is a signaling protocol designed for use in telecommunications at the application layer. It has generally become one of the primary protocols used in VoIP (voice over Internet protocol), H.323, and other communications standards. SIP can be seen as a primary provider of Web services for Internet-based voice communications such as VoIP (Sulkin, 2007).

Contract (Services)

Components of a service contract typically include primary and secondary elements. The primary elements consist of the header, functional requirements, and nonfunctional requirements. Subelements for the header consist of the name, version, owner, RACI, and type. Under functional requirements are functional requirement descriptions, service operations, and invocation. Nonfunctional requirements include security constraints, QoS, transactional requirements (the service part of a larger transaction), service-level agreement, and process ("SOA," 2007). The contract generally includes metadata about itself, who owns it, and how it is brokered, bound, and executed.

Interface

At this level of service provision, the interface referred to is a segment of code that connects the service with the data and/or business logic (process). The interface describes how data will be moved into and out of the data source by the service, and must be designed to comply with the physical (data, data structures, etc.) and process (business logic) requirements of the existing and/or legacy system.

Implementation

The implementation specifies the contract and interface to be used for each service requested, and contains the direct pathway into the data and business logic.

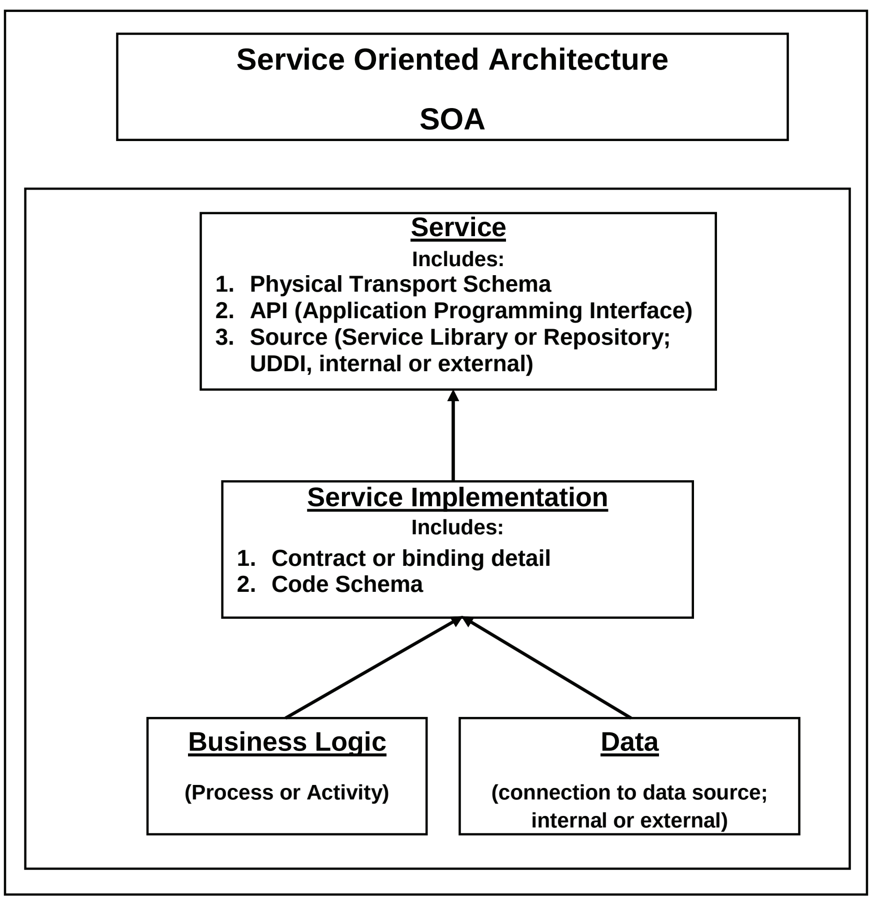
Architecture

The service component of SOA has been discussed, though admittedly at a high level. However, the architecture component has not yet been addressed and it will be helpful to speak briefly about the architecture segment of SOA. Architecture in general refers to the art (or science) behind the design and building of structures. Alternatively, an architecture may refer to a method or style of a building or a computer system. So, if SOA is taken literally as a description of its function, it could be taken to mean a structured way of organizing or arranging the services in a business or organization.

SOA FRAMEWORK

It is apparent from the existing definitions and models that service-oriented architecture is commonly seen as an architecture or way of assembling, building, or composing the information technology infrastructure of a business or organization. As such, SOA is not a technology in itself; rather, it is a way of structuring or arranging other technologies to accomplish a number of other tasks. This naturally leads to the problem of a multiplicity of definitions of SOA since many relatively similar structural arrangements of services are possible. Many of the definitions also indicate that the arrangement and relationships between modules should be loosely coupled rather than tightly coupled. This allows for customization of services based on need, and on-demand rather than some predetermined structure, but the downside is that it also leads toward a plethora of definitions and approaches to SOA implementation.

Figure 1. SOA framework



Some of the common features that seem sensible to include in a formal definition of SOA would relate to a common framework, such as that specified by Krafzig et al. (2005) or one of the other standards bodies. In other words, a framework would include metadata describing the various important features of SOA, how those features can be arranged, and the libraries or location of services that allow adopting organizations to arrange bindings or contracts between themselves and the service provider, independent of whether the service provider is internal or external. We propose the framework depicted in Figure 1 as a starting point for visualizing SOA.

Several of the standards bodies have taken a stance in creating or calling for a metamodel, at least in some form. Among them are the Open Group, OASIS, OMG, W3C, and to a lesser extent industry-related bodies such as Javaworld.com, XML.com, IBM, and Oracle.

UDDI has become a very well-known structured repository for services and service components, which speaks to the universality of the library or centralized database of services. However, more standardization efforts will be necessary to enhance the interoperability of UDDI.

It also appears, especially with the industry definitions of SOA, that the contracts, bindings, interfaces, service buses, and other implementation-related portions of SOA are important elements to be considered when attempting to give an overall definition of SOA. This unfortunately could easily represent a stumbling block in garnering consensus on a definition of SOA since each of these companies has invested significant time, human, and other likely resources toward development of their specific pieces of the SOA pie. Each company has invested heavily and thus will likely be less willing to risk that investment and any potential return and customer lock-in in order to simply agree on standards. We observed a similar occurrence of this type of behavior in the recently ended format war in the high-definition DVD market. Similarly, if the standards bodies have political or industry leanings, agreement on a common SOA definition and standards could be difficult to achieve.

Another more recent development comes from Shah and Kalin (2007). They proposed that organizations adopting SOA follow a specific path based on an analysis of business challenges, including SOA business drivers and IT barriers. This led them to speculate that a specific adoption model be used to guide the SOA implementation process. They indicated that an ad hoc SOA model is better where the benefits of new services are specific to each individual service, where the technologies may be inconsistently applied (different implementations for the same service in different projects), where services cannot be reused, and where the increases in technological complexity translate into decreased system response times. Shah and Kalin ended with a call for a strategy- or program-based SOA adoption model that is situational.

We propose that a common definition of SOA is possible and necessary, and call for negotiations among interested bodies with the aim of reaching a common definition of SOA. We realize that in practice it might prove difficult or even nearly impossible to expect such a consensus to be arrived at, but a common definition and structure of SOA would go a long way toward dealing with some of the confusion, misinformation, and hype regarding the entire subject. Difficult though it might be to expect this, a realization that SOAP, CORBA, RPC, and XML among many other technological tools have reached a point of relative agreement amongst users if not ubiquity, at least related to their underlying standards, should provide some evidence that agreements can be reached. Next, we will examine SOA from the research perspective.

POSSIBILITIES FOR RESEARCH

Research into SOA is extremely limited at this point in time. What studies exist can be classified into several distinct categories. The first includes exploratory or recommendation-type efforts that propose various means to approach SOA implementation. These investigations may or may not include proprietary industry software, but most of these research efforts propose the use of patterns or blueprints and a metamodel of SOA as a means to understanding the SOA perspective. Second, in this category are research proposals that examine company-specific technologies or tools (i.e., IBM proposing the use of Rational Software, including the Rational Unified Process) in relation to SOA design and implementation. Neither of the first two types of SOA research generally involve ideas on how to measure SOA in terms of success or failure, or even suggest metrics. Finally, the third type of research articles focus on empirical research.

SOA Development or Deployment Patterns and Blueprints, and the Meta-Approach

Stal (2006) took a roughly similar approach to what we are attempting to do in this article; he advocated using architectural patterns and blueprints (software engineering patterns) as a means to enable or foster efficient deployment of SOA. He supported loose coupling of services in a registry or library to the extent that he thought that removing the services' dependency on the registry's or provider's distinct location would benefit the deployment of SOA. Stal maintained that this would eliminate, or at least minimize, a layer in the SOA framework. He also proposed a more tightly defined and controlled integration of middleware using XML or similar tools. Basically, Stal suggested a metamodel and pattern approach to defining SOA, but did not suggest what the research might accomplish or how the research into SOA would be framed. Kim and Lim (2007) also proposed a distinct means to implementing SOA, using in this instance, business process management, in addition to a variant of the SOA framework specifically dealing with the telecommunications industry. Similar to Stal, Kim and Lim did not propose empirical research into SOA, but rather focused on implementation and standards in a specific industry.

Shan and Hua (2006) proposed an SOA approach for the Internet banking industry. They also compiled a list of patterns that have been proven successful for other online service industries. However, the models they used and ended up with are very detailed regarding how SOA should be implemented for first online companies in general, and then Internet banking specifically. This again does not propose or frame specific research but rather suggests an implementation approach and a structure for SOA.

The ESB is explained in detail, but from a general perspective rather than a company-specific approach in Schmidt, Hutchison, Lambros, and Phippen's (2005) expository. The article is informative regarding ESB implementation and design patterns, but it is not research oriented.

Crawford, Bate, Cherbakov, Holley, and Tsocanos (2005) proposed a different way to structure SOA, what they called on-demand SOA. They essentially proposed an even looser coupling of services and their connecting elements than in other perspectives of SOA. They argued that this would allow much more flexibility to the adopting organizations and the end users.

Company-Specific and Commercial Tool-Based SOA Deployment

Brown, Delbaere, Eeles, Johnston, and Weaver (2005) presented an industry-oriented perspective on the SOA puzzle. They suggested an approach to service orientation using the proprietary IBM Rational platform. Their recommendations follow similar paths as some previous research, but are also filtered through the IBM Rational lens. The article is primarily illustrative in nature, suggesting how to best implement SOA using IBM Rational tools. In a similar vein, Ferguson and Stockton (2005) also detailed IBM's programming model and product architecture.

De Pauw, Lei, Pring, and Villard (2005) described the benefits of Web Services Navigator, a proprietary tool created to provide a better visualization of SOA and Web services in a loosely coupled architecture. The tool can help with design-pattern, business-logic, and business-process analysis, and thus help with SOA architecture design and implementation.

Jones (2005) suggested that SOA, service, and Web service standards were "on the way" and provided a list of existing tools, such as UML (Unified Modeling Language) and/or the rational unified process that could aid the SOA (or service) design process. However, he also advocated the push toward formal definitions of such SOA basics as services, to the end of providing a more coherent and cohesive structure that he thought would enhance the ability of developers and adopters to understand and deploy SOA.

Research-Based Perspectives on SOA

Chen, Zhou, and Zhang (2006) proposed an ontologically based perspective on SOA, Web services, and knowledge management. They attempted, with some success, to integrate two separate research streams into one. They presented a solution to show that semantic- and syntactic-based knowledge representations could both be depicted with a comprehensive ontology that also described Web service composition. While their framework represents a step toward automated (Web) service composition, more research is still needed.

Borkar, Carey, Mangtani, McKinney, Pate, and Thatte (2006) suggested a way of handling XML-based data in an SOA or service environment. Their idea involved the use of data both able to be queried and unable to be queried, and would necessarily also involve XML-formatted data. This represents empirical research into a part of SOA, namely, the underlying services, and is at least a step in the right direction, although it does not enter the realm of research into the efficacy or ROI of SOA.

Duke, Davies, and Richardson (2005) recommended and provided details on using the Semantic Web to organize an organization's approach to SOA and Web service orientation. They suggested that combining the Semantic Web and SOA into what they called Semantic SOA would provide benefits to adopting organizations. Then they further proposed an ontological model of the Semantic SOA, attempting essentially to create a meta-metamodel of SOA using their experience with the telecommunications industry as a case example. This is one of the few high-level articles that can also be seen as empirical research.

Zhang (2004) explored the connection between Web services and business process management, and described the modular nature of the service (and Web service) perspective. He detailed the software industry's approach to Web services and provided evidence that standards development would quickly mature, beginning in 2005. He maintained that once standards were agreed upon, a connection to business process management would be easier to sell to businesses. Zhang also developed a prototype e-procurement system that composed external services to operate.

Malloy, Kraft, Hallstrom, and Voas (2006) developed an extension to WSDL. They insisted that Web services' specifications were "typically informal and not well-defined," and proposed what they called an intermediate step between requiring more formal and rigorous service specifications and the informal nature of the existing service specifications. They accomplished this balance by extending WSDL to include support for application arguments that would help automate and expand the ability of services to operate in multiple environments. They provided an example of how their WSDL extension could allow a single service to function successfully in different applications using multiple zip code formats (five vs. nine digits, and hyphens vs. no hyphens).

Verheecke, Vanderperren, and Jonckers (2006) proposed and developed a middleware level that they called the Web services management layer (WSML). They saw the primary advantage of their approach in that it provided a reusable framework. They further believed that the use of their framework would enable "dynamic integration, selection, composition, and client-side management of Web Services in client applications" (p. 49). They were aware that their approach could cause some problems in a distributed system since implementation of it resulted in a centralized architecture.

Hutchinson, Henzel, and Thwaits (2006) described a case in which an SOA-based system was deployed for a library extension collaboration project. Much of the case details the SOA approach itself, and explains the experiences of the project developers and implementers. They noted that while the SOA architecture could be expected to reduce the operational maintenance costs overall, the way the system was specified and delivered in this particular case might require more work from IT to keep some services, such as flash players, up to date. While the authors did not specifically mention it in the article, perhaps a more loosely coupled architecture might alleviate some of those operational maintenance costs.

Li, Huang, Yen, and Cheng (2007) proposed a methodology to migrate the functionality of legacy systems to a Web services or SOA architecture. They used a case study to investigate the efficacy of their proposed methodology, finding that while it was possible to make such a migration from legacy systems to SOA (or Web services), the changes that it required from the organization were considerable, and some process reengineering would likely be necessary.

MEASURING SOA AND SOC SUCCESS

Another tricky issue in SOA and SOC implementation is the measurement or evaluation of success. Traditionally, software (or system) successes and failures have been estimated by the usual suspects: traditional measures such as ROI, net present value (NPV), breakeven, internal rate of return (IRR), or other similar financially based approaches. Similarly, software itself has usually been measured in terms of errors or productivity via numeric methodologies such as lines of code, COCOMO (constructive cost model), and similar estimation techniques. These approaches are all based firmly on the idea that if we can assign some number to a system, then we can compare them across projects, systems, or organizations. The problem is analogous to the question often asked regarding enterprise resource planning (ERP) systems: If all of the Fortune 100 companies implement the same piece of software, such as SAP, then what allows one organization to differentiate itself from another if they have standardized on SAP's best processes and best practices? One way to answer that question is to examine other measures of success such as competitive advantages (Siau, 2003), competitive necessity, flexibility, agility (Erickson, Lyytinen, & Siau, 2005), nimbleness, responsiveness, and other relevant intangibles. We would even propose that the best way to evaluate SOA or SOC implementation is not ROI. Intangible but critical factors such as competitive necessity, agility, on-demand abilities, and responsiveness should be the decisive factors.

Nah, Islam, and Tan (2007) proposed a framework and critical success factors for estimating the success of ERP implementations. They empirically assessed a variety of implementation success factors including top-management support, project team competence, and interdepartmental cooperation, among many others. While the study answered a number of important questions regarding ERP implementations, the issue of assessing intangibles in terms of success factors remains a problem, not only for ERP-type implementations but also for other system types as well, especially for SOA since the SOA approach can be seen as an alternative in many ways to ERP.

Langdon (2007) noted that while many economic-based studies indicate that IT projects add value at the macrolevel, little has been done to assess how value is added at the more micro or individual project level. Specifically, Langdon proposed and evaluated a research model that included (IS) integration and flexibility as capabilities that could lead to IT business value. Of course, flexibility and integration are only two components of a larger IT capabilities structure, but the study indicates that the first steps have been taken to study intangibles in the context of an IT systems development project.

Two intangibles in the IT success-factor context are the oft-cited agility or nimbleness of a company or organization. An entire genre of systems development has emerged based on the principle of agility. However, there is little empirical evidence supporting the value added from such development approaches (Erickson et al., 2005). Since a growing number of SOA installations are constructed as ad hoc, which is in a basic sense agile, we propose that in environments where agility and nimbleness are important, so in turn are SOA and SOC important.

CONCLUSION

From the literature, it appears that only a few efforts can be said to be empirical research. A majority of the research efforts involved created tools or language extensions that would increase the interoperability of services, while other research proposed standards modifications. Many of the remaining articles published proposed new tools or the use of existing proprietary tools, described an approach to SOA from specific perspectives, or proposed model or metamodel changes. A limited number of case studies detailing SOA, Web services, or service deployments or implementation efforts provide experience reports on how best to implement such systems.

As far as we can determine, virtually no research has been formally done regarding the benefits and drawbacks of SOA or Web services. Two problems with this are likely to revolve around the nebulous nature of SOA and Web services in terms of the widely varying definition and the emerging standards issue. An effort to identify SOA and Web services metrics would help to get research into this area started.

Another area of interest involving SOA and Web services adoption is the cultural and structural impacts on the organization or business. A number of articles note the importance of those elements, but little has been accomplished in terms of research specifically connecting SOA or Web services with cultural and structural changes in organizations.

A variety of standards bodies are working separately toward formal definitions including metamodels, and a number of SOA vendors, among them some of the very large and established software industry players, have emerged. While the effort toward standardization is direly needed and commendable, a more collaborative approach would, in our opinion, benefit the industry and implementing companies and organizations as well. The seeming result of the rather haphazard approach to SOA appears to indicate that an increasing number of implementing organizations are finding it difficult to assess the cost benefit of the entire services approach. Research efforts at this point appear to be in a similar state of disarray. Until a more coherent picture of SOA emerges, its image is likely to remain slightly out of focus, and research in the area is likely to remain somewhat unfocused as a result.

REFERENCES

Borkar, V., Carey, M., Mangtani, N., McKinney, D., Patel, R., & Thatte, S. (2006). XML data services. *International Journal of Web Services Research*, *3*(1), 85-95.

Brown, A., Delbaere, M., Eeles, P., Johnston, S., & Weaver, R. (2005). Realizing service oriented solutions with the IBM Rational Software Development Platform. *IBM Systems Journal*, *44*(4), 727-752.

Chen, Y., Zhou, L., & Zhang, D. (2006). Ontology-supported Web service composition: An approach to service-oriented knowledge management in corporate financial services. *Journal of Database Management*, *17*(1), 67-84.

Crawford, C., Bate, G., Cherbakov, L., Holley, K., & Tsocanos, C. (2005). Toward an on demand service architecture. *IBM Systems Journal*, *44*(1), 81-107.

De Pauw, Lei, M., Pring, E., & Villard, L. (2005). Web services navigator: Visualizing the execution of Web services. *IBM Systems Journal*, *44*(4), 821-845.

Dorman, A. (2007). FrankenSOA. *Network Computing*, *18*(12), 41-51.

Duke, A., Davies, J., & Richardson, M. (2005). Enabling a scalable service oriented architecture with Semantic Web services. *BT Technology Journal*, *23*(3), 191-201.

EJB. (2007). *Wikipedia*. Retrieved October 12, 2007, from http://en.wikipedia.org/wiki/Ejb

Erickson, J., Lyytinen, K., & Siau, K. (2005). Agile modeling, agile software development, and extreme programming: The state of research. *Journal of Database Management*, *16*(4), 80-89.

Ferguson, D., & Stockton, M. (2005). Service oriented architecture: Programming model and product architecture. *IBM Systems Journal*, *44*(4), 753-780.

Havenstein, H. (2006). Measuring SOA performance is a complex art. *Computer World*, *40*(2), 6.

Hicks, B. (n.d.). *Oracle Enterprise Service Bus: The foundation for service oriented architecture*. Retrieved October 18, 2007, from http://www.oracle.com/global/ap/openworld/ppt_download/middleware_oracle%20enterprise%20service%20bus%20foundation_250.pdf

Hutchinson, B., Henzel, J., & Thwaits, A. (2006). Using Web services to promote library-extension collaboration. *Library Hi Tech*, *24*(1), 126-141.

Jones, S. (2005). Toward an acceptable definition of service. *IEEE Software*, *22*(3), 87-93.

Kim, J., & Lim, K. (2007). An approach to service oriented architecture using Web service and BPM in the Telcom OSS domain. *Internet Research*, *17*(1), 99-107.

Krafzig, D., Banke, K., & Slama, D. (2005). *SOA elements*. Prentice Hall. Retrieved October 2, 2007, from http://en.wikipedia.org/wiki/Image:SOA_Elements.png

Langdon, C. (2007). Designing information systems to create business value: A theoretical conceptualization of the role of flexibility and integration. *Journal of Database Management*, *17*(3), 1-18.

Lee, J., Siau, K., & Hong, S. (2003). Enterprise integration with ERP and EAI. *Communications of the ACM*, *46*(2), 54-60.

Li, S., Huang, S., Yen, D., & Chang, C. (2007). Migrating legacy information systems to Web services architecture. *Journal of Database Management*, *18*(4), 1-25.

Malloy, B., Kraft, N., Hallstrom, J., & Voas, J. (2006). Improving the predictable assembly of service oriented architectures. *IEEE Software*, *23*(2), 12-15.

Nah, F., Islam, Z., & Tan, M. (2007). Empirical assessment of factors influencing success of enterprise resource planning implementations. *Journal of Database Management*, *18*(4), 26-50.

Object Management Group (OMG). (2007). Retrieved September 25, 2007, from http://colab.cim3.net/cgi-bin/wiki.pl?OMGSoaGlossary#nid34QI

Open Group. (2007). Retrieved September 25, 2007, from http://opengroup.org/projects/soa/doc.tpl?gdid=10632

Organization for the Advancement of Structured Information Standards (OASIS). (2006). Retrieved September 25, 2007, from http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=soa-rm

Ricadela, A. (2006, September 4). The dark side of SOA. *Information Week*, pp. 54-58.

Schmidt, M., Hutchison, B., Lambros, P., & Phippen, R. (2005). Enterprise service bus: Making service oriented architecture real. *IBM Systems Journal*, *44*(4), 781-797.

Shah, A., & Kalin, P. (2007, July 6). SOA adoption models: Ad-hoc versus program-based. *SOA Magazine*.

Shan, T., & Hua, W. (2006). Service oriented solution framework for Internet banking. *Internet Journal of Web Services Research*, *3*(1), 29-48.

Siau, K. (2003). Interorganizational systems and competitive advantages: Lessons from history. *Journal of Computer Information Systems*, *44*(1), 33-39.

Siau, K., & Tian, Y. (2004). Supply chains integration: Architecture and enabling technologies. *Journal of Computer Information Systems*, *44*(3), 67-72.

SOA. (2007). *Wikipedia*. Retrieved September 25, 2007, from http://en.wikipedia.org/wiki/Service-oriented_architecture#SOA_definitions

Stal, M. (2006). Using architectural patterns and blueprints for service oriented architecture. *IEEE Software*, *23*(2), 54-61.

Sulkin, A. (2007). SOA and enterprise voice communications. *Business Communications Review*, *37*(8), 32-34.

Trembly, A. (2007). SOA: Savior or snake oil? *National Underwriter Life & Health*, *111*(27), 50.

UNL-IBM System in Global Innovation Hub. (2007). *Making SOA relevant for business*. Retrieved October 9, 2007, from http://cba.unl.edu/outreach/unl-ibm/documents/SOA_Relevant_Business.pdf

Verheecke, B., Vanderperren, W., & Jonckers, V. (2006). Unraveling crosscutting concerns in Web services middleware. *IEEE Software*, *23*(1), 42-50.

Walker, L. (2007). IBM business transformation enabled by service-oriented architecture. *IBM Systems Journal*, *46*(4), 651-667.

Web service. (2007). *Wikipedia*. Retrieved October 18, 2007, from http://en.wikipedia.org/wiki/Web_service

World Wide Web Consortium (W3C). (2007). Retrieved September 25, 2007, from http://colab.cim3.net/cgi-bin/wiki.pl?WwwCSoaGlossary#nid34R0

XML.com. (2007). Retrieved September 25, 2007, from http://www.xml.com/pub/a/ws/2003/09/30/soa.html

Zhang, D. (2004). Web services composition for process management in e-business. *Journal of Computer Information Systems*, *45*(2), 83-91.

Zhao, L., & Siau, K. (2007). Information mediation using metamodels: An approach using XML and common warehouse metamodel. *Journal of Database Management*, *18*(3), 69-82.

This work was previously published in Journal of Database Management, Vol. 19, Issue 3, edited by K. Siau, pp. 42-54, copyright 2008 by IGI Publishing (an imprint of IGI Global).

Chapter 6.8
Diffusion and Oscillation of Telecommunications Services: The Case of Web 2.0 Platforms

Tobias Kollmann
University of Duisburg-Essen, Campus Essen, Germany

Christoph Stöckmann
University of Duisburg-Essen, Campus Essen, Germany

Carsten Schröer
University of Duisburg-Essen, Campus Essen, Germany

ABSTRACT

The diffusion of a Web 2.0 product or services is, unlike to traditional consumer or industrial goods, not only based on purchase. Full acceptance of Web 2.0 platforms occurs by recurring utilization. The chapter focuses on diffusion characteristics of this innovative category of ICT products and provides management concepts for competition. The concept of critical mass is applied to different growth scenarios. Additional success factors are discussed. Particularly the permanent supervision of a platform regarding its compliance with qualitative, as well as ethical and legal standards is of great importance. Adjustments to external market conditions, proactive management, and a bilateral marketing approach are a key to lasting success within the Net Economy. Markets are never settled, due to the ever changing and oscillating conditions. The chapter shows that there is always a chance to capture a market or at least to grow against competition in a Web 2.0 setting.

DOI: 10.4018/978-1-60566-194-0.ch036

CRITICAL MASS AS A SUCCESS FACTOR

The extension of electronic networks and the use of information and telecommunication technologies for the digitalization of value creation lead to a new economic dimension (Lumpkin & Dess, 2004). This newly established level of value creation, the so-called **Net Economy**, provides room for innovative business models and successful start-up firms (Kollmann, 2006). An increasing number of companies participate in the economic potential of the internet which leads to a rising level of competition. Competing players either win a market and participate in a stable and sustainable business development or fail with their idea within a short period of time (Shapiro & Varian, 1999). The roots of this phenomenon are derived from an economies of scale effect that keeps aggravating itself instead of declining. Every new user of an offered **platform** (community or marketplace) helps to raise the value of a network and makes it even more attractive for further participants. A higher number of communication and transaction activities are the possible outcome. A rising quantity of community members also increases the perceived attractiveness (site stickiness) of a platform. This can be illustrated by the following two examples: A rising number of members subscribing to an ***E-Community*** (Kollmann, 2006) raises the chance to meet likeminded individuals or to receive answers to posted questions. Also a rising number of users to an ***E-Marketplace*** (Kollmann, 2006) rises the probability to find interested customers for offered products of a supplier.

According to the presented scenarios, a special focus has to be put on the **critical mass** phenomenon, because the subjectively perceived attractiveness of a system (e.g. community) is highly correlated with the already registered number of users. A certain number of users within a network are necessary to create value among the participants at a sophisticated level. Reaching this level is essential for a network, because the enrolled participants will be reinforced to use the system on an ongoing basis, and it will become easier to convince new users to join in (Kollmann, 1998). The minimum number of participants to maintain a sufficient utility on a long-term basis is referred to as the **critical mass** (Weiber, 1992).

Especially in a **Net Economy** setting young companies experience a very competitive environment to reach the **critical mass** (Kollmann, 1998). Oftentimes, the winners of this race drive smaller competitors or copycats off the market. This conception reinforces itself in a **Web 2.0** setting (O'Reilly, 2005), where customers or members leave the status of pure information consumers. Their status changes to an active information provider and editor role (O'Reilly, 2005). Therefore, growth at a fast pace in regards to the number of users becomes the critical success factor to leave the zone of competition as a winner. Actually, the winner of this battle is able to establish a close too monopolistic market position (Shapiro & Varian, 1999). The attractiveness for new users to join a network is even higher, if *everyone else* already joined in.

Following the stated assumptions Web 2.0 **critical mass** winners are destined for lasting company performance and profits. But the real life teaches another lesson. Apparently successful market leaders are frequently challenged by various inconveniences with the potential to jeopardize their market position. In accordance with the theoretical model ***eBay™*** for example, market leader for internet auctions, announced a growing number of membership accounts alongside with rising revenues and profits (eBay™, 2006). The unmentioned downside of this success story was a flood of insolvencies among professional ***eBay™*** dealers. *The International E-Business Association* (IEBA), an association for *power sellers*, sees the roots for many discontinued businesses closely connected with an increasing number of sellers and a resulting higher level of competition. Both factors lower profits and force sellers to predatory pricing strategies. In this context insolvency reasons for most of the dealers are not on an individual, entrepreneurial level. They are based on the market characteristics of electronic marketplaces, and a substantial number of insolvencies by professional dealers of a platform will sooner or later hit the marketplace vendors.

Other **critical mass** winners within the **Web 2.0** environment like the online community ***MySpace™*** or the video **platform** ***YouTube™*** are not only the centre of interest because of their enormous growth rates and success stories. Critical notes about security issues, copyright violations, or identity theft and fraud affairs are also on the spot of public interest. Web 2.0 companies might face severe challenges, if the offered content on their platforms violates ethical or legal standards. Insufficient qualities of the offered content, as well as a mismatch between information supply and demand in the case of ***eBay™***, gain the potential for adverse effects for a market position.

Therefore, the proposed chapter aims to show how quantitative, qualitative, as well as ethical and legal matters correspond with the market success of Web 2.0 platforms. In addition implications for the competition of platforms and the concept of **critical mass** as a foundation for success in the **Net Economy** will be discussed.

WEB 2.0 PLATFORMS IN THE NET ECONOMY

In the past the internet used to be recognized as a technology to publish and distribute data, information and media content. This view was based on split-up roles: Private and commercial publishers of web content with an active role on the one hand, and passive consumers on the other hand. This *golden rule* changed in 2005, when **Web 2.0** saw the light of day. A new category of website concepts was born. The established differentiation between active content providers and passive consumers diluted. Now users were able to generate and affect contents. User generated content became the slogan for the new internet. The active role of users built the basis for innovative business ideas, which were unthinkable some months ago. Many **Web 2.0** business models, like the online community ***MySpace™*** or the video **platform** ***YouTube™***, are centered around community structures. According to Kollmann (2006) an **E-Community** facilitates contact and interaction between individuals or institutions via a digital network. Therefore, the integration of innovative ICT supports data- and knowledge transfer. These two features characterize the core activity for most of the **Web 2.0** business models. Besides arranging and exchanging contacts and information, bringing together supply and demand for economic transactions, e.g. on an **E-Marketplace** (Kollmann, 2006), is an integral feature of numerous platforms.

The common goal of **Web 2.0** platforms is to win suppliers and consumers of information for their business model in order to match them (Kollmann, 2006). Users of a **Web 2.0** platform act alternating and parallel as information suppliers and consumers. Both activities have to be considered separately, because supplying and requesting information differ in regards to motivation and acceptance. This leads to a tripolar structure. The platform operator provides a matching service to perform an exchange of information or a business transaction at lower transaction costs (Lee & Clark, 1996). The value of a **platform** not only depends on the operators' service capability and willingness to perform, but also on the contributions of the suppliers and consumers of information (derivative capability aspect). Platforms depend on queries. A higher number of queries provide a broader scope for matching activities (Kollmann, 2001). Consequently a **Web 2.0** platform is solely dependent on the participation willingness (acceptance) of its users. Therefore business development efforts concentrate on the so-called matching as a target parameter (Kollmann, 2000, 2005).

When starting a **platform**, operators have to get awareness for their services. They are challenged by the question, which points are of interest to turn internet surfers into members or subscribers for a community, or respectively customers for a marketplace. This goal can only be reached by offering a matching **platform** that delivers an acceptable service. Therefore the scientific construct of acceptance obtained growing relevance in the marketing of ICT products and services over time (Kollmann, 2001). The reason for this is that ICT technologies and applications (e.g. interactive TV, internet, cellular phones) need a specific pattern of utilization. The pure purchase is not an indicator for further activities of a user and therefore not a sufficient indicator for the economic success of a **platform** operator. Augmenting this idea, a full acceptance of a **Web 2.0 platform** is closely linked to three conditions.

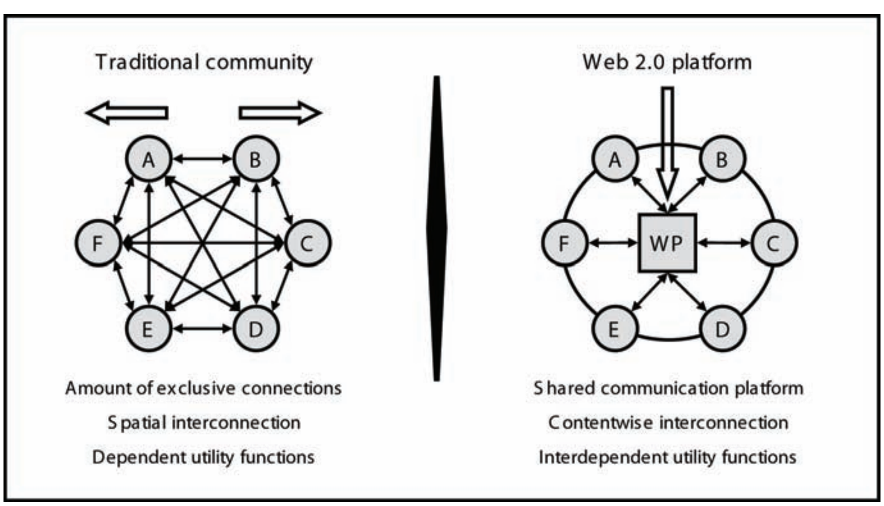
1. Connecting (Access to the platform): The customers have to get access to a platform via a security code (e.g. log-in) or provided access software. A first time registration process reflects the purchase.
2. Acting (Demand and contribution of information): The customers have to use the services of the electronic platform. Information has to be requested and provided. This opens up the potential for matching.
3. Interacting (Clearing and matching): The customers have to interact on the platform; otherwise a matching of requested and provided information cannot be executed.

Coordination of the participants is proceeded over the electronic **platform** (n information suppliers, m information consumers, and the platform operator without time and local limitations, see figure 1). Information and data between two or more counterparts is solely shared on the **platform**. The active placement and intervention into the matching process leads to a new responsibility for the platform operator, because the result of each matching affects all participants.

The attractiveness of traditional communities and marketplaces is primarily determined by the numbers of participants as a quantitative measure for their availability (see figure 1). The center of those system architectures is the exclusive connection between two counterparts (e.g. trade show or farmers market). The interaction between A and B usually has no direct qualitative effect for the utility of C. Technological external effects are conceivable under certain conditions, e.g. if all participants benefit from a network effect associated with the extension of a telephone network. However, this interpretation of a quantitative point of view is insufficient to explain attractiveness of progressive internet platforms.

In a **Web 2.0** system the interconnection of participants does not consist of exclusive one-on-one data links. An **E-Community** or **E-Marketplace** as a commonly shared platform represents the center of the system architecture (see figure 1). Information of a database is available for every user on the network. Options to alter, comment, or expand the provided content are inherent to the

Figure 1. Illustration of traditional communities and Web 2.0 platforms



system. A transparent provision of information leads to direct effects on the qualitative utility function of C (economies of scope), if A and B exchange information. An agreeing or derogative comment of B to a contribution from A, may lead to a valuation of C regarding the statement. Further, as Pavlou and Gefen (2005) state, a psychological contract violation with an individual seller is proposed to prompt a generalized perception of contract violation with the entire community of sellers in a marketplace. Internet auctions for a specific item are open to several participants. The bid of A has a direct impact on the utility function of all other bidders. The following paragraph will deal with the resulting implications for the diffusion of Web 2.0 platforms.

THE DIFFUSION OF WEB 2.0 PLATFORMS

Research on **diffusion** provides answers to the question how an innovation will spread on a market (Rogers, 2003, Pavlou & Fygenson, 2006). Services of a **Web 2.0** platform, which are offered as a commercial product, could be part of a study on diffusion, too. Research in this field is based on the presumption of a recurring use (acceptance) of a product or service, not on a one-time purchase (adoption). In regards to the diffusion of **Web 2.0** platforms the following three questions are of interest.

1. Which factors have an impact on the diffusion of a Web 2.0 platform on a market?
2. How fast does a Web 2.0 platform spread on a market?
3. What are the growth characteristics of a network?

The successful **diffusion** of a **Web 2.0 platform** is completed, if all interactions of a defined market are handled by this platform. For the evaluation of **diffusion** the aforementioned quantitative alignment of the network effect (higher number of participants = higher probability to find appropriate counterparts for interaction) appears with an economies of scope effect (nature, size and trend of an executed transaction, including its impact on the overall system). The following paragraphs will analyze the main problem areas of diffusion and oscillating degrees of utilization in the light of both effects.

Problem Areas of Diffusion

The **diffusion** of a **Web 2.0** platform is associated with quantitative, qualitative, ethical and legal challenges. Those issues will be discussed in the following.

Quantitative Problem Areas

Attractiveness of a web **platform** is significantly linked to the number of participants. A higher participation level raises the chance to reach other individuals. Every information supply (e.g. a provided video) as well as every information demand (e.g. on a personal level) need at least one counterpart to enable a platform provider to match requests. The service of the platform provider creates an indirect utility that is derived from the usage of an interactive relationship within the communication system, the so-called derivative capability aspect (Katz & Shapiro, 1985; Farrell & Saloner, 1985). The derivative utility following the usage of such a good increases with the number of participants, and the intensity of use by the other participants (Weiber, 1992). The result is a network effect. Common examples for goods with direct network effects are all types of ICT-systems. The utility of each participant is advanced with every new customer, who helps to grow the network. In connection with the bilateral customer orientation (information supply and demand) of the platform provider, specific characteristics within the **diffusion** of a **platform** can be derived for different development stages (Kollmann, 2001).

1. Chicken-and-Egg-Problem: One reason for matching problems on a **Web 2.0** platform is derived from the so-called Chicken-and-Egg-Problem (Durand, 1983; Earston, 1980). The following two examples aim to illustrate this circumstance. An insufficient number of suppliers or offers lead to an absence of customers on the platform. An insufficient number of customers or requests lead to a lack of suppliers. The dilemma situation, which counterpart (supplier or consumer) at first has to get involved with the platform, is deemed as an obstacle for the development of a business.
2. Collateral-Critical-Mass-Problem: The installed basis, i.e. the number of users already present in a platform, determines the utility of the platform for new users since a greater number of users also increases the number of potential interactions (Farrell & Saloner, 1986). The larger the installed basis, the larger is the derivative utility for the participants (Kollmann, 2001). **Web 2.0**-platform providers are confronted with a collateral **critical mass**, because of the bilateral orientation (Kollmann, 1998). Suppliers need a certain level of counterparts or requests, in order to commit to or use a marketplace. Simultaneously, a certain level of suppliers or offers has to be provided, in order to persuade a customer to facilitate a marketplace. This problem supersedes itself, if the customer base on both sides grows to a sufficient point, where the derivative utility exceeds a certain level.
3. Equilibrium problem: Bilateral matching results in a mutual state of dependence regarding the number of suppliers and consumers, and respectively their offers and requests. Consequently the platform provider has to take into consideration that offers and requests almost equate themselves. Bilateral marketing activities support this endeavor (Kollmann, 1998) and help maintaining a high matching level (one offer = one request).

Qualitative Problem Areas

Contrary to the established belief in an exclusive utilization act between supplier and consumer interaction, the critical phase of a matching includes an additional economies of scope effect associated with the quality of interaction. The decision to subscribe to and use a **platform** has to be expanded. Besides complying with quantitative issues, meeting the qualitative requirements of the suppliers and consumers with information is of equal importance. If they realize that the web platform complies with their demand and interaction needs, they are willing to utilize the platform's services. The following issues with regard to qualitative problems have to be solved (Kollmann, 2001).

1. Matching performance problem: An exclusive focus on the number of suppliers and consumers is insufficient to measure the quality of the interacting counterparts, as well as their level of satisfaction with regard to the exchange of information. The demanded level of interaction has to correspond with the expectations of the participants. The degree of satisfaction is closely linked to three core areas of need, which are information, relationships, and business (Hagel & Amstrong, 1997), as well as the related concept of the heterogeneity-dependent level of commitment. Participants look for like-minded individuals on a platform and relevant content to fulfill their information needs. New discussion threads have to be established to acquire additional members, which cover further fields of interest. Platform operators are exposed to a dilemma situation, because of the diametrical impact of the heterogeneity of discussion threads with regard to a growing member base and the persistent

commitment of current participants.

2. Reality check problem: The structural conditions of virtual platforms disallow to validate provided information with reality. Anonymous publishing options among some platforms aggravate this issue. Information and reality fall apart frequently. Some of those discrepancies occur inadvertently, e.g. if a change of address is not entered into a database or an information is provided on a non-current standard of knowledge. More frequently intentional misrepresenting takes place by sugar-coating one's profile on the web, or even worse with criminal intention. Because the roots of a reality gap are unknown to the participants, misrepresentations might reduce the commitment or ongoing patronage of a user.

Ethical and Legal Issues

The addressed willful misrepresentation is an example for the multitude of ethical and legal issues associated with user generated content, among other legal problems of internet **platform**s and their foundation. Current lawsuits on those topics will give answers to important questions and provide future guidance on duties of **Web 2.0** platform operators. Liabilities for provided content on a company's website and linked content from external sources, as well as infringements of users (e.g. announcement of a criminal offense) and related duties of care have to be clarified. The results of those decisions will inevitably have a major impact on the further diffusion and development of **Web 2.0** platforms. In the following section two problems will be discussed.

1. Freedom of expression problem: Despite the fact that every human being possesses the right to express an opinion, legal and ethical standards have to be obeyed. Sometimes those standards are violated on anonymous web-based communication platforms. Comments with an extremist, offending, or sexually harassing content cannot be tolerated by any platform operator. Other categories, like advertising and promotion activities, can be classified as unwanted too. Guidance for communication on the internet is given by the so-called *netiquette*, derived from internet and etiquette. The recommended behavior of the netiquette is not legally binding, but helps to maintain and develop a positive net culture. The voluntary agreed upon rules are frequently incorporated in codes of conduct of platform operators. A breach of the rules leads to a closing of discussion threads, cancelation of comments, or dismissal of accounts, because a negative communication culture bears the lasting potential to lower the acceptance of a platform.
2. Problems associated with the adoption of external content:

Particularly on video platforms users provide, intentionally or unintentionally, copyrighted material from other websites or real sources. By now, copyright holders mandate agencies (e.g. *copyrightcontrol*^{TM}) to retrieve their protected material. The platform operator is responsible for inflicting penalties of participants and to remove copyrighted materials. Preventing an upload of protected material is virtually impossible, because the violation of a copyright just becomes apparent after a user has posted illegal content. Even reactive behavior leaves a legal stain on the platform operator's vest, as despite the fact of a fast content removal a breach of law already happened. Since there is no appropriate method available on the market to avoid the upload of copyrighted material, all prominent platform operators strive towards general licensing agreements with bailees. Those arrangements would allow them to leave copyrighted material on their websites and protect them from costly lawsuits and negative impacts on their market penetration.

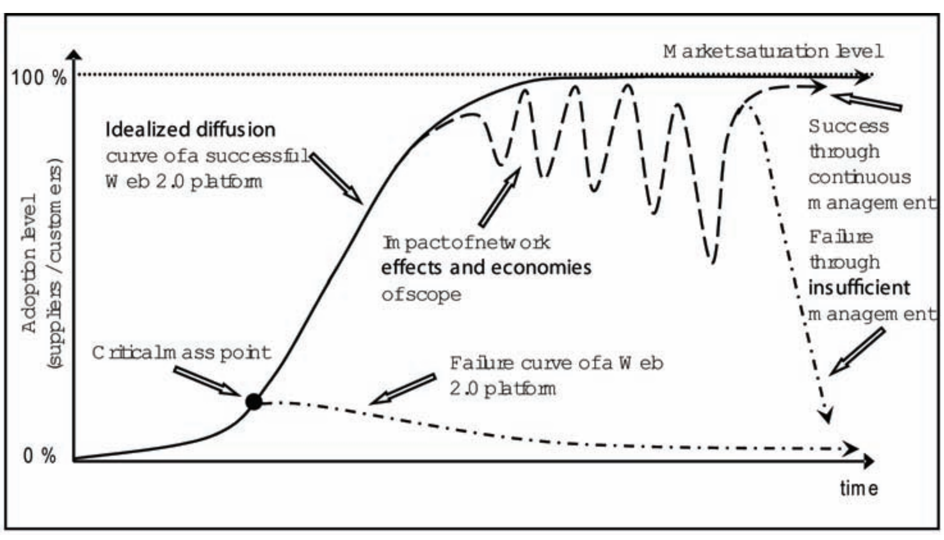
Oscillating Effects of Web 2.0 Platforms

The classical theory of **diffusion** was developed for so-called singular goods. Their diffusion proceeds with the act of buying. In a **critical mass** system the scope has to be extended by the variables of *connecting* and *acting* (see chapter *Web 2.0 Platforms in the Net Economy*) as constitutive determinates for diffusion (Weiber, 1992). With traditional consumer and producer goods the act of buying creates a positive and irreversible impact for diffusion. However, ICT service contracts bear the risk of being cancelled (e.g. mobile phone contract, website account), which limits the chance to realize demand synergies. As an extreme example reversible utility could cause a declining diffusion (see figure 2; Weiber, 1992). The characteristics of a **diffusion** curve in a **critical mass** system generally do not reflect a monotonically increasing function; in fact a considerable drop is also possible.

Diffusion Characteristics

The traditional model needs an expansion within a **Web 2.0** setting, because the connecting act is an insufficient parameter to evaluate **diffusion**. It is just a necessary requirement for adoption. The market success of a **Web 2.0 platform** depends directly on the participants' constant utilization and interaction as a reliable measure for adoption, and therefore acceptance (Kollmann, 2001). An adequate utilization and interaction discipline supports a premium quality of information and knowledge transfer among the participants of a platform with positive effects for the whole market system. Also the recurring utilization and interaction is a prerequisite to realize constant cash flows for the platform operators. Accordingly the concept of **diffusion** has to be extended beyond the purchase dependent quantitative measure of participants to the utilization and interaction dependent quantitative interaction measure. The reversibility of utilization and interaction has to be considered in this context. Due to the planning interval the sequence of the three adoption and acceptance acts is interpreted as a discontinued multiple event. This process is characterized by permanent oscillations

Figure 2. Diffusion of Web 2.0 platforms



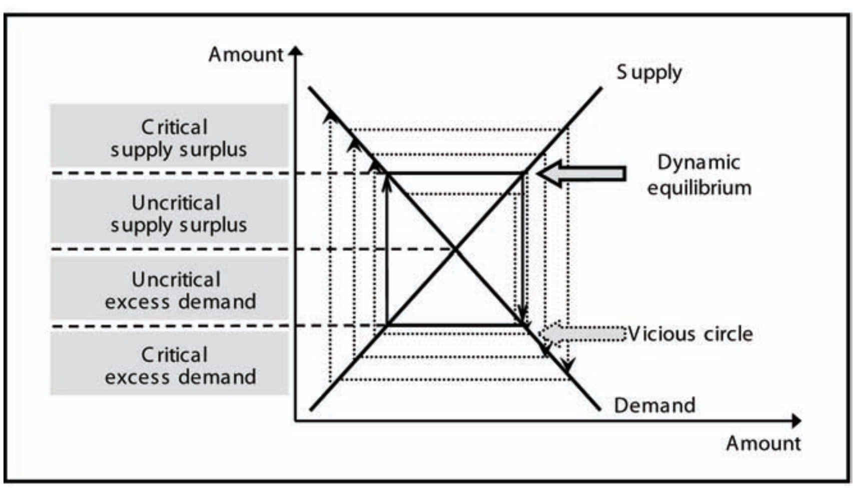
making **diffusion** a permanent companion. The reach of market saturation is not only linked with negative adoption and acceptance ratios, but also with alternating positive and negative ratios. The direct consequence is an oscillating development at market saturation level.

This effect is caused by the circumstance of negative utilization and interaction levels, which lead to a renouncing or deregistration by the participants. The original decision of adoption is withdrawn. Potential reasons to resign are caused by effects from quantitative, qualitative, ethical and legal issues. Suppliers and consumers of information feel uncomfortable about these aspects. The interconnection of participants has a negative impact on their individual utility function and the whole market system. An oscillating **diffusion** is the result (see Figure 2). In this context special attention has to be paid to the proportion of information supply and demand. As postulated, both factors should create an equilibrium to meet all matching requests and create a high level of satisfaction among the community members. The ever-changing level of participants leads to an alternating level of supply and demand. **Web 2.0 platform**s with a transaction oriented business model (e.g. ***eBay***^{TM}) use the price of goods as an instrument to regulate the proportions of supply and demand. Even though the control mechanism does not converge, due to continuously changing prices and numbers as well as suppliers and consumers, an indifferent dynamic equilibrium is the possible outcome (see figure 3).

Communication oriented platforms are faced with different challenges. A price based self-regulating mechanism does not exist in their business environment. Monitoring each market situation is important to perform intervening actions. From a platform operator's point of view the dynamic participation of suppliers and consumers is not necessarily associated with challenges. The dynamics reveal a harmful potential, if a critical supply surplus or excess demand is detected. Both scenarios cause the same effect (see figure 3).

The reduction of supply side participants will most likely lead to a decreasing information supply. Hence there is not enough potential to match all information requests. This causes a permanent loss of attractiveness on the demand side, as the offered information does not meet expectations. A lack of quantity and quality causes consumers to leave the **platform**. Likewise the lower number of requests makes it even more unattractive for suppliers to provide information. In this case the

Figure 3. Adoption process of information supply and demand



development of a platform does not induce an oscillation, but moreover a creeping and declining growth. Neither suppliers nor consumers of information are stimulated to return to the platform and utilize the services again (see figure 3).

In conclusion, the control mechanism of supply and demand generates a self-aggravating effect with a positive (virtuous circle) and negative (vicious circle) development potential (Kollmann, 1999). This control mechanism makes the management of a **Web 2.0 platform** complex, because the operator as an independent facilitator faces a bipolar user group simultaneously. Matching efforts should consider individual as well as general interaction requests to fulfil the expectation of opposed groups of interest.

The positive scenario (virtuous circle) in accordance with the **critical mass** effect leads to a continuous growth of power, achievement potential, and attractiveness of the platform. A significant gain of information suppliers typically shows an increase of requests with positive effects on the choice of the selected set of matching opportunities. In turn, the positive impact raises consumer satisfaction and the number of information requests. The flourishing demand for information lifts up the platform attractiveness for the suppliers and so forth (positive loop).

As stated before in this paragraph, changed market conditions may possibly restart a negative control mechanism. Within this vicious circle a significant loss of information suppliers leads to a clear reduction of requested information. Negative impacts on the number of choices for a matching lead to a high number of unmatched interaction demands. This has a negative impact on consumer satisfaction and consequently on the amount of information request. A declining demand reduces the attractiveness of the platform for suppliers, which leads to an ongoing downturn of supply and so forth (negative loop). The **Web 2.0** platform could suffer from those effects by a continuous loss of power and achievement potential. At the bottom line the existence of the platform might be at stake.

The oscillating characteristics of the **diffusion** curve lead to serious implications regarding the management of a **Web 2.0 platform** and the competition between web-based communication platforms in general. Those aspects along with the illustration of the **critical mass** effect as a success guarantee will be explained in the following paragraph.

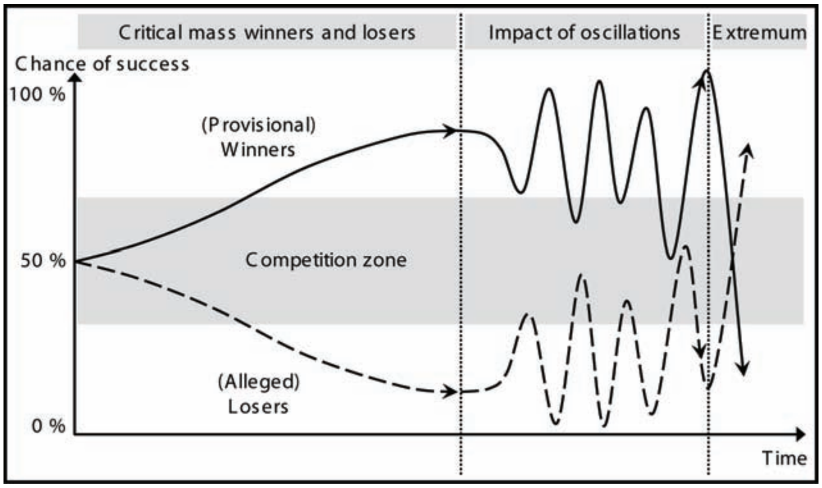
Competition on Diffusion

Competition in the **Net Economy** is characterized by an oscillating utilization of **platform**s. Therefore reaching a **critical mass** does not automatically assure a promising and stable development (see figure 4). Even on saturated or apparently settled markets threats for **critical mass** winners occur on a frequent basis. Alleged losers or innovative start-up companies occasionally create surprisingly good chances to grow against reputable competition.

As demonstrated, established **Web 2.0** platforms are endangered to lose participants. In a worst case scenario they start a vicious circle with the potential to bring their company down. The goal of successful **platform**s is therefore to maintain equilibrium between the bipolar groups as well as safeguarding compliance with qualitative, legal, and ethical standards. Lasting survival is closely connected with a high level of commitment among the participants and protection against competition. The continued management of a web **platform** becomes the critical factor of success. Provisional **critical mass** winners have to be on permanent alert, instead of relaxing in their accomplished position. Ongoing market evaluations and proactive influence on current developments on the respective network through bilateral marketing are inevitable.

Weak phases of critical mass winners provide opportunities for start-ups and established competitors to attack the supremacy of market

Figure 4. Competition in oscillating areas of conflict[1]



leaders. The existence of financial strength and survivability supports the gain of market shares (see figure 4). It is unrealistic to turn around a market completely or drive a leader in a certain field off the market, but addressing special target groups with innovative niche products receives growing popularity. Current examples are university and high school student communities, which recently started their services. Starting a positive control mechanism (positive loop) is the beginning to overcome the **critical mass** sustainably. Especially young start-ups are confronted with the challenge to reach **critical mass**. Their brand name is usually unknown to the broad public and the network attractiveness of their platform is limited. But the value of a network product is not solely based on the number of participants. Future development expectations also play an influential role (Hagel & Armstrong, 1997). Well timed and promising announcements to the market in advance (vapor marketing) and the management of expectations combined with additional online and offline marketing activities bear the potential to occupy a niche by the massive acquisition of new customers.

CONCLUSION

The **critical mass** plays a vital role for the implementation of internet business models. As described, the **critical mass** concept is of crucial importance in the age of **Web 2.0**. Current developments show further challenges besides this success factor. Those challenges have to be taken into account with the management of a **Web 2.0 platform**, because they could have a severe impact on company development. Particularly the permanent supervision of a platform regarding the compliance with qualitative, as well as ethical and legal standards is of great importance. Adjustments to external market conditions, proactive management of the platform, and a bilateral marketing approach are key for a lasting success within the **Net Economy**. Finally competitors and founders of (new) ventures should keep in mind that a market is never settled, because of the ever changing and oscillating market conditions. There is always a chance to capture a market or at least to grow against competition.

AUTHOR NOTE

Tobias Kollmann, Christoph Stöckmann, and Carsten Schröer, E-Business and E-Entrepreneurship Research Group, University of Duisburg-Essen, Campus Essen. Correspondence concerning this article should be addressed to Tobias Kollmann, E-Business and E-Entrepreneurship Research Group, University of Duisburg-Essen, Campus Essen, Universitässtr. 9, 45141 Essen, Germany. E-mail: tobias.kollmann@uni-due.de

REFERENCES

Durand, P. (1983). The public service potential of videotext and teletext. *Telecommunications Policy*, *7*(6), 149–162. doi:10.1016/0308-5961(83)90070-8

Earston, A. (1980). Viewpoint. *Telecommunications Policy*, *4*(9), 220–225.

eBay (2006). *eBay Inc. Announces fourth quarter and full year 2005 financial results*. Retrieved March 31, 2007, from http://investor.ebay.com/news/Q405/EBAY0118-123321.pdf.

Farrell, J., & Saloner, G. (1985). Standardisation, compatibility and innovation. *The Rand Journal of Economics*, *16*(1), 70–83. doi:10.2307/2555589

Farrell, J., & Saloner, G. (1986). Installed base and compatibility: Innovation, product, preannouncement and predation. *The American Economic Review*, *76*(5), 940–955.

Hagel, J., & Armstrong, A. G. (1997). *Net gain*. Boston, MA: Harvard Business School Press.

Katz, M. L., & Shapiro, C. (1985). Network externalities, competition, and compatibility. *The American Economic Review*, *75*(3), 424–440.

Kollmann, T. (1998). Marketing for electronic market places – the relevance of two critical points of success“. *Electronic Markets*, *8*(4), 36–39. doi:10.1080/10196789800000039

Kollmann, T. (1999). Virtual marketplaces: Building management information systems for internet brokerage. *Virtual Reality (Waltham Cross)*, *4*(4), 275–290. doi:10.1007/BF01421811

Kollmann, T. (2000). Competitive strategies for electronic marketplaces. *Electronic Markets*, *10*(2), 102–109. doi:10.1080/10196780050138155

Kollmann, T. (2001). Measuring the acceptance of electronic marketplaces. *Journal of Computer-Mediated Communication*, *2*(6).

Kollmann, T. (2005). The matching function for electronic market places – determining the probability of coordinating of supply and demand. *International Journal of Electronic Business*, *5*(3), 461–472. doi:10.1504/IJEB.2005.008520

Kollmann, T. (2006). What is e-entrepreneurship? – Fundamentals of company founding in the net economy. *International Journal of Technology Management*, *33*(4), 322–340. doi:10.1504/IJTM.2006.009247

Lee, H. G., & Clark, T. H. (1996). Impacts of the electronic marketplace on transaction cost and market Structure. *International Journal of Electronic Commerce*, *1*(1), 127–149.

Lumpkin, G., & Dess, G. (2004). E-business strategies and internet business models: how the internet adds value. *Organizational Dynamics*, *33*, 161–173. doi:10.1016/j.orgdyn.2004.01.004

O'Reilly, T. (2005). *What Is Web 2.0. Design patterns and business models for the next generation of software*. Retrieved March 31, 2007, from http://www.oreillynet.com/ pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

Pavlou, P. A., & Fygenson, M. (2006). Understanding and predicting electronic commerce adoption: An extension of the theory of planned behavior. *MIS Quarterly*, *30*(1), 115–143.

Pavlou, P. A., & Gefen, D. (2005). Psychological contract violation in online marketplaces: Antecedents, consequences, and moderating role. *Information Systems Research*, *16*(4), 372–399. doi:10.1287/isre.1050.0065

Rogers, E. M. (2003). *Diffusion of Innovations* (5th ed.). New York: Free Press.

Shapiro, C., & Varian, H. R. (1999). *Information rules. A strategic guide to the network economy* (Reprint). Boston: Harvard Business School.

Weiber, R. (1992). *Diffusion von Telekommunikation*. Wiesbaden: Gabler.

KEY TERMS AND DEFINITIONS

Critical Mass: Of users is an important success measure for Web 2.0 platforms, because the subjectively perceived attractiveness of a system (e.g. community) is highly correlated with the already registered number of users. A certain number of users within a network are necessary to create value among the participants at a sophisticated level. Reaching this level is essential for a network, because the enrolled participants will be reinforced to use the system on an ongoing basis, and it will become easier to convince new users to join in. The minimum number of participants to maintain a sufficient utility on a long-term basis is referred to as the critical mass.

E-Community: Stands for a virtual community. E-Communities are one sort of communication platform on the internet, and support or initiate business processes. They are used to build constant, self dynamic communication and interaction processes.

E-Marketplaces: Are virtual-based organizations to exchange services. An E-Marketplace has an institutional framework for transaction processes. They can be interpreted as virtual space where supply and demand is coordinated.

Net Economy: Refers to the commercial/business dimension of electronic data networks and is therefore a digital network economy. Different electronic platforms perform the execution of information, communication and transaction processes. The starting point for its expansion is the development of the information society. The basis of the Net Economy is formed by four technological innovations: telecommunication, information technology, media technology and entertainment (the so-called TIME markets). These innovations have, and continue to, significantly impact the possible ways in which information, communication and transactions are managed. The increased support of business processes using electronic systems takes centre stage here. There are a number of terms for this that can be identified (e.g. e-business, e-commerce, information economics, network economics), which can, to some degree, be used synonymously.

Netiquette: Is derived from the terms internet and etiquette. Despite the fact that every human being possesses the right to express an opinion on the internet, legal and ethical standards have to be obeyed. Sometimes those standards are violated, especially on anonymous web-based communication platforms. Comments with an extremist, offending, or sexually harassing content should not be tolerated by any individual or platform operator. Guidance for communication on the internet is given by the so-called netiquette. The recommended behavior of the netiquette is not legally binding, but helps to maintain and develop a positive net culture. The voluntary agreed upon rules are frequently incorporated in codes of conduct of platform operators.

Vapor Marketing: Is characterized by promising announcements to the market on products or services in advance and the management of expectations combined with additional online and offline marketing activities.

Web 2.0: Is the next evolutionary step of the internet. In the past the internet used to be recognized as a technology to publish and distribute data, information and media content. This view was based on split-up roles: Private and commercial publishers of web contents with an active role on the one hand, and passive consumers on the other hand. This golden rule changed in 2005, when Web 2.0 concepts as a new category of websites were established. The traditional differentiation between active content providers and passive consumers diluted. On Web 2.0 platforms users are able to generate and affect contents. User generated content became the slogan of Web 2.0. The active role of users built the basis for innovative business ideas, which were unthinkable before. Many Web 2.0 business models like online communities or video platforms are centered around community structures.

ENDNOTE

[1] Based on Shapiro & Varian (1999, p. 177)

This work was previously published in Handbook of Research on Telecommunications Planning and Management for Business, edited by I. Lee, pp. 557-570, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Section VII
Critical Issues

This section addresses conceptual and theoretical issues related to the field of Web technologies, which include issues related to usage, as well as failures and successes in Web implementation. Within these chapters, the reader is presented with analysis of the most current and relevant conceptual inquires within this growing field of study. Particular chapters address privacy concerns in Web logging, Web information extraction, and Web rules. Overall, contributions within this section ask unique, often theoretical questions related to the study of Web technologies and, more often than not, conclude that solutions are both numerous and contradictory.

Chapter 7.1
Applying an Organizational Uncertainty Principle:
Semantic Web–Based Metrics

Joseph Wood, LTC
US Army, USA

Hui-Lien Tung
Paine College, USA

Tina Marshall-Bradley
Paine College, USA

Donald A. Sofge
Naval Research Laboratory, USA

James Grayson
Augusta State University, USA

Margo Bergman
Northwest Health Services Research & Development (HSR&D), USA

W.F. Lawless
Paine College, USA

ABSTRACT

The theory of bistable perceptions in the interaction indicates the existence of an uncertainty principle with effects amplified at the organizational level. Traditional theory of the interaction, organizational theory, and the justification for an organizational uncertainty principle are reviewed. The organizational uncertainty principle predicts counterintuitive effects that can be exploited with the Semantic Web to formulate a set of metrics for organizational performance. As a preliminary test of the principle, metrics derived from it are applied to two case studies, both works in progress, with the first as an ongoing large system-wide application of web-based metrics for organizational performance and the second as a case study of a small college where web-based metrics are being considered and constructed. In preparation for the possibility of machine-based real-time metrics afforded by the Semantic Web, the results demonstrate a successful theory and application in the field of an uncertainty principle for organizations.

INTRODUCTION

Overview. No theory of organizations is widely accepted today (Pfeffer & Fing, 2005). In this chapter, we provide a brief discussion of the problems with traditional organizational theory and, focusing on fundamentals, a classical (quantum) alternative model that accounts for predictions from traditional theory and at the same time its supposedly spurious but ultimately disconfirming findings. With its focus on the individual, traditional theory, also known as methodological individualism, encompasses social learning theory (SLT) and game theory. SLT includes classical or Pavlovian conditioning, operant or Skinnerian reinforcement, and modeling (for a revised version and summary, see Bandura, 1989). In contrast to SLT, game theory focuses on the interaction between two or more individuals (Von Neuman & Morgenstern, 1953), but like SLT, it is static; an attempt to make game theory dynamic employs repeated presentations of static game matrices. But the need for the classical (quantum) alternative is inherently based on the fundamental questions raised by the traditional focus on the individual.

In addition to theory, we review field data and the application of the organizational uncertainty principle in the form of performance metrics to two case studies, one of an ongoing, long-term nature and the other incipient. Both case studies are web-based. We include a review of the future semantic web and its implications for the two case studies. Finally, we discuss future prospects with the semantic web for theory, tests and computational models of the organizational uncertainty principle, and a path forward for the two case studies.

From the perspective of the Semantic web, our objectives are to review traditional social learning and game theory for organizations and the alternative organizational uncertainty principle. Our objective for the organizational uncertainty principle is to justify its formulation based on evidence and to review two case studies that use metrics to exploit the organizational uncertainty principle. Our final objective is to provide a path forward with automatic machine-based data generating real-time online metrics for future research with the semantic web.

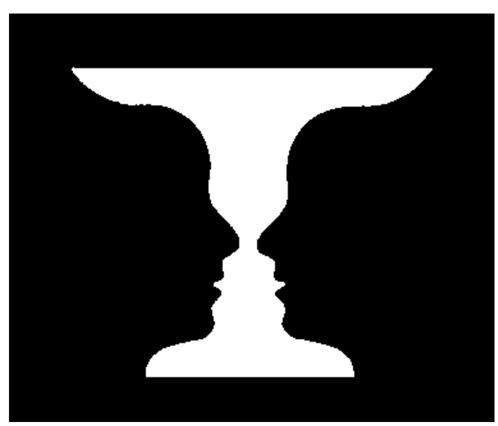
BACKGROUND

Definitions

In the Background, we define bistability, the organizational uncertainty principle, SLT, game theory, organizational theory, and Semantic web. In addition, after each term is defined, we provide a brief literature review for each term. At the end of the references, we summarize these key terms.

Bistability. Bistability is best explained with an example of an illusion (see Figure 1). It occurs when one data set can produce two mutually exclusive interpretations. While the entire data set is processed cognitively by an individual, both interpretations of a bistable illusion cannot be held in awareness simultaneously (Cacioppo et al., 1996). According to Bohr (1955), multiple interpretations support the existence of different cultures. Further, given the importance of feed-

Figure 1. An example of bistability



In this very simple two-faces vase illusion, an observer is incapable of observing both aspects of this single picture at the same time.

back to social dynamics (Lawless et al., 2007), the possibility of rapid shifts ($j\omega$, where j is the imaginary number $\sqrt{-1}$ and ω is the discrete frequency in radians per second) between each bistable interpretation increases uncertainty in the non-observed interpretation which not only can create social problems between different groups, but also support the existence of an uncertainty principle, characterized as tradeoffs between incommensurable views (e.g., the rapid shifts in the interpretations presented to a neutral jury during the cross-fires of courtroom debates). We propose that measurements of bistable phenomena collapse interdependence, decreasing uncertainty in observed aspects of a bistable phenomenon while increasing uncertainty in non-observed aspects; e.g., questionnaires of interdependent cognitive-behavior states.

Organizational uncertainty principle. The organizational uncertainty principle acts as a tradeoff in attention directed at reducing the uncertainty in one factor, such as a worldview or business model, with the result that the uncertainty in a second, interdependent factor is inversely increased, such as the execution of a business model. It is based on Bohr's (1955) famous notion that the quantum uncertainty principle applied to social situations is captured by the bistability of action and observation. That is, the more focused a collective of individuals are on acting out a series of steps, the less observant they become of their actions. Applied to societies, clusters of action-observation uncertainty couples form a causal path for different cultures based on multiple interpretations of the same business model or worldview (e.g., religion, liberalism, conservatism). The organizational uncertainty principle we have proposed links uncertainty between planning and execution as well as between resource availability and duration for plan execution (Lawless et al., 2006).

Social learning theory. SLT is a term coined by Bandura (1977) to include the three different schools of thoughts that accounted for learning by organisms, but particularly apt for humans. These three schools were classical conditioning (associations), operant conditioning (rewards and punishments), and modeling, Bandura's own school of thought. According to Bandura, modeling subsumed the other schools of learning. SLT works well in changing behavior at the individual level, irrespective of an individual's cognitive contributions or willingness to change. However, it offers little in the way of insight into organizational dynamics or organizational change, such as mergers and acquisitions (M&A), restructuring, or solving ill-posed cultural problems. More relevant to our thesis, SLT has been adapted to cognitive behaviors, which are dependent on surveys of interdependent physical behaviors. The critical assumption usually made for SLT is that self-reports of the cognitive perceptions of behavior match the behaviors that actually occur, an assumption not supported empirically.

Game theory. Game theory was invented in the 1940's by Von Neuman & Morgenstern (1953). It is a one-shot game, or a series of repeated games, played by two or more agents. In its most basic form, the game configuration presents two static, arbitrary choices for payoffs to each player. Payoffs are interdependent. For example, in the well-known and well-studied Prisoner's Dilemma Game, two players (prisoners) who cannot communicate to each other must decide whether to cooperate with each other, thereby gaining the most points for their dyad, or to compete against each other, producing the least total points for their dyad (Nash equilibrium). However, the best individual payoff occurs if one player competes under the condition that the other player cooperates. Although no communication occurs between the two "prisoners" undergoing simulated interrogation "by the police" in two isolated rooms, the feedback afforded by repeated games from observing prior partner choices affects future choices, leading to Axelrod's (1984) "evolution" of cooperation.

One of the problems with game theory is that it is normative (Körding, 2007). The results from playing these games follow the prevailing social norms independently of fundamental human behavior. That is, the value of the choices available to be made by the participants are not based on empirical support on improving social welfare, but on society's worldview of the ideal social behavior that promotes social welfare; viz., cooperation is superior to competition (this arbitrary choice by scientists is analogous to choosing market winners by authoritarian or command decision-making governments). Second, game theory inverts the assumption made by SLT about behavior. Game theory assumes that a static configuration imposed on cognition generates the desired (normative) behavior exactly. As a consequence, although providing very clear predictions, the results from games do not predict human behavior (Sanfey, 2007).

In sum, when considering the two prevailing models in the social sciences, game theory and social learning theory, these two models of methodological individualism do not focus on improving social welfare, generating creativity, or solving social problems; however, they expect to derive these benefits secondarily. Further, both theories assume the existence of a 1:1 mapping between interdependent states of cognition and behavior, leaving no room for the collapse of interdependence; e.g., by measurements or decision-making. This situation places social scientists in a box—they recognize that restrictions for self-reports are necessary in at least the case where extreme claims made about behavior cannot equal actual behavior (for hypochondria and alcoholic denial, one over states self-reported behavior and the other under states self-reported behavior).

Organizations. Organizations are social groups that perform a function which often cannot be done by an individual alone. Organizations do this by assigning interdependent roles to a set of independent individuals, which requires information coordination and blocking to form them into a working collective, but which consequently amplifies the capabilities of an individual (Ambrose, 2001). An organization arises to serve a function when its operational costs are less than the benefits it accrues and provides to its members (Coase, 1937). It is likely constructed around a geographical centroid (x) about which its business attributes are centered, planned and modeled (i.e., BM_x; derived from Sukthankar, 2008). But multiple theories of organizations exist (Weick & Quinn, 1999). Pfeffer and Fong (2005) concluded that one of the problems is the lack of a foundational theory for organizations; consequently, they proposed the need to incorporate illusions into basic organizational theory. Although "illusions" could be metaphorical, imaginary factors are instrumental in engineering to model oscillations ($j\omega$). We propose that active illusions bruited about during an open discussion interact with real world feedback to generate discussion oscillations ($j\omega$) until interdependence collapses.

Semantic Web. The Semantic web is an ongoing project to extend the World Wide Web (WWW) to permit humans and machines to collaborate efficiently. As envisioned by Berners-Lee (2007), inventor of WWW (and web languages URI, HTTP, and HTML), the future Web should evolve into a universal exchange for data, information and knowledge. Without a universal standard for machine access, HTML data is difficult to use on a large scale. The Semantic web solves this problem with an efficient global mesh for information access by both humans and machines.

The Semantic web includes a Health Care and Life Sciences Interest Group (HCLSIG, 2008) to establish interoperable data standards for "connected knowledge" to improve collaboration across the health care and life sciences, in our case, military medical research training services. The goal for HCLSIG is to reduce medical errors, increase physician efficiency and advance patient care and satisfaction. It includes document annotation and rule processing (with XML formats, OWL ontologies and SWRL rule processors).

A future option for the Semantic web, but one we pursue now in Case Study 1, may be electronic dashboards to link scientific publications and electronic medical records to associate disease, drug compounds, biology and pathway knowledge between R&D groups. As a final concern for HCL-SIG, there is today no widely recognized machine accessible semantic differentiation between a manuscript and publication; illustration and experimental image data; or between an experiment, data, interpretations, and the hypothesis that an experiment was designed to validate. Initially, our first web-based study in Case Study 1 addresses only parts of these problems with an adaptive electronic Institutional Review Board (eIRB) for research protocols rather than medical records; but associated with the eIRB, we are considering business intelligence for individual organization and system-wide performance metrics, and linking scientific publications from multiple military R&D groups to improve patient care.

Brief Literature Review

In addition to the literature reviewed in the background, an additional but brief review is provided here to place our work in a historical context. On its face, Durkheim's (1893) "social facts" stand against Weber's (1958) methodological individualism, today ingrained in game theory, where the choices available to those playing games are influenced by the social and religious norms existing within a culture (Körding, 2007). As an example, the choice to cooperate with a partner in the Prisoner's Dilemma Game is configured with a higher value than the choice to defect from a partner, even though from an information theory perspective, society often gains significantly more social benefits from competition than cooperation (Lawless & Grayson, 2004). While social norms should not be disparaged but studied, neither should scientists favor the norm of cooperation by configuring it with a higher social welfare value, similar to an industrial policy that chooses the winners for its society. But there are limitations to Durkehim's view, too. If reality is bi-stable, social facts are open to multiple interpretations.

Parsons and Luhman contributed to cybernetics and control theory as an information approach to controlling and modeling society. Parsons (1966) developed a systems approach as a grand theory of society. He used systems as a tool to analyze the function of society, concluding the systems that adapt to their environment had evolved into more efficient systems; however, in that the environment is ever changing, adaptation is not an optimal control strategy (Conant & Ashby, 1970). Parsons influenced Luhmann's (1984) theory of autopoietic, or self-organizing, systems. Luhman believed that autopoietic systems filtered information from the environment, independently of other systems, making them autonomous, but also apart from society. Elias (1969) contributed to cybernetics with his ideas on figurational, networked or interconnected structures as the source of power over other systems. Crozier and Friedberg (1981) used game structures to explicitly analyze power and strategy between organizations and their members as interdependent actors. But the limitations remain for game theory from the influence of social norms and the lack of a theory of uncertainty.

Finally, and contrary to Weber's view of different beliefs producing structural differences in a society, Montesquieu (1949) suggested that checks and balances contribute to society by limiting power. Madison applied Montesquieu's idea by building a constitutional government based on checks and balances (Hamilton, Madison, & Jay, 1787-1788), concluding that social structure controls and stabilizes government independently of the social norms in existence. Further, not only do checks and balances recognize the limits of situational awareness, motivated by the search for meaning at the individual level (Carley, 2002); but also, consensus rules and compromise dilute the added information provided to society by checks and balances, their strength. However, compromise leads to an "action consensus" based on a concrete plan of action, compared to the unified worldview of consensus seeking, which reduces the likelihood of action (Lawless et al., 2008b). This is not to conclude that Weber's ideas missed the mark. Just the opposite. Weber understood that the tradeoffs between the incommensurable beliefs of Confucianism and Puritanism produced profound differences in the control of and social welfare benefits for two social systems, which agrees with the uncertainty relations presented below.

MAIN FOCUS OF THE CHAPTER

In general, most of social science is predicated on the assumption that observations of behavior, especially the self-observations made in response to questionnaires, provide perfect or near perfect information about a target behavior, thereby leaving no room for an uncertainty principle. However, striking problems exist with asking agents about the causes of their behavior (self-reports, surveys, structured interviews, case studies). Baumeister et al. (2005) found that a 30-year meta-analysis of survey data on self-esteem correlated poorly with academic and work performance, casting doubt on one of the most studied phenomena in psychology and also on the ability of self-reports to capture targeted phenomena. Similarly, in an attempt to prove the value of air combat maneuvering for Air Force educators, Lawless and his colleagues (2000) found no association between air combat outcomes (wins-losses) and examination scores on air-combat knowledge. And at the end of his distinguished career in testing game matrices, Kelley (1992) found no association between the preferences as measured by surveys before games were played and the choices actually made during games. Along the same line, Axsom and Lawless (1992) found that scientists easily misinterpreted the causes of behavior measured in effort justification experiments designed to reduce public speaking anxiety even when the scientists observed the changes directly.

In their review of decision theory, Shafir and LeBoef (2002) concluded that justifications of actions already taken were not consistent with the actions taken, including for expert judges. In addition, they found that the widely held belief by theoreticians that expectations of well-being lead to well-being was systematically violated, even for experts. But even though the evidence in support of widespread claims based on self-reports does not exist, many social models continue to endorse the belief that cooperation enhances the value of social welfare more than competition. In agreement with Pfeffer and Fong, the lack of fundamentals has produced a subjective social science. In response, we take a more theoretical approach based on the impact that cooperation and dissonance have on the diminution or generation of information (Lawless & Grayson, 2004).

To summarize, metrics must not interfere with the process of measurement; doing so collapses interdependence and invokes the organizational uncertainty principle (e.g, surveys of self-esteem at the individual level by Baumeister et al., 2004; decision-making at the organizational level; Lawless & Grayson, 2004). Perceptions are integral to behavior, as the Coca-Cola Company discovered when it decided to close out its traditional Coca-Cola brand due to its inability to best Pepsi-Cola in internal taste tests (en.wikipedia.org/wiki/New_Coke). But following considerable public criticism, the firm brought back its traditional cola and re-branded it "Classic Coke". As Baumeister has re-discovered, the measurement of perceptions in interdependent states with behavior collapses the interdependence, producing static information.

We plan to study organizations with computational models. However, Bankes (2002) and Conzelmann and his colleagues (2004) have both concluded that current computational models of organizations are not predictive, principally with Agent-Based Models (ABMs). We plan two correctives: first, to test models using social configurations addressed by our organizational uncertainty model to reproduce the results of collapsed interdependent states that we have predicted and found in the field and laboratory; and second, to build bistable organizations constituted with bistable artificial agents.

Organizational Theory and Uncertainty Principle

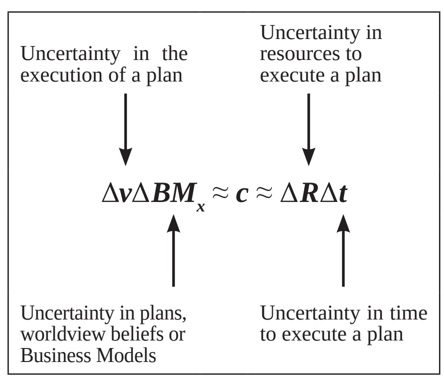
In contrast to traditional social science, we have attempted to combine individuals with organizations and systems, statics with dynamics, and empirical approaches with theory. We incorporate dynamics in our model with the effects of feedback on oscillations within an organization, but as a metric for its performance. We incorporate organizations in our model by introducing control as organizations seek to improve in performing or revising their mission (Mattick & Gagen, 2005; also, May, 1973). Finally, in our approach, an empirical approach alone precludes formal approaches and optimal solutions; our immediate goal, then, is to build and be guided by theory and empirical findings.

To implement control theory (Csete & Doyle, 2002), we need to quantify an organizational or system's level model. In line with earlier arguments, an organization controls at least four aspects of the decision-making process. First, by helping to set or choose its reference or threshold set-points (e.g., culture, decision processes, expectations, planning; and in Case Study 1, mission and vision). Second, by damping unexpected disturbances. Third, by filtering and transforming incoming information about system internal states, inputs, and responses to form patterns and decisions. Finally, by taking actions then collecting feedback to revise decisions. However, Conant and Ashby (1970) concluded that feedback to minimize errors is not an optimal solution for control, that the optimum solution avoided errors (e.g, with a plan that produces the most efficient operation possible).

As metrics for our control theory, we have proposed inverting the organizational uncertainty principle in Figure 2 to link uncertainty between planning and execution as well as between resource availability and the duration of plan execution.

In Figure 2, uncertainty in the social interaction is represented by an interdependence between business models, strategy, plans, or knowledge uncertainty (ΔBM_x, where the knowledge or business model is a function of the social location where it was learned; from Latané, 1981 and Sukthankar, 2008) and uncertainty in the rate of change in knowledge or its execution as $\Delta v = \Delta\ (\Delta BM/\Delta t)$. This relationship agrees with Levine and Moreland (2004) that as consensus for a concrete plan increases (ΔBM_x reduces), the ability to execute the plan increases (Δv increases). By extension,

Figure 2. Measurement problem



The measurement problem occurs as the result of the organizational uncertainty principle. The measurement problem arises from the interdependence between the two factors on each side of the equation. It states that both factors on either side of the equation cannot simultaneously be known exactly. For example, a decrease in the uncertainty in the strategy for an organization results in an increase in uncertainty for the execution of that strategy. In practice, decreasing strategic uncertainty increases action; increasing strategic uncertainty slows action (Busemeyer, 2008). At the same time, the uncertainty principle informs us that only one of the factors on either side of the equation can be known with certainty (Lawless et al., 2007).

interdependence also exists in the uncertainty in the resources expended to gain knowledge, ΔR, and by uncertainty in the time it takes to enact knowledge, Δt. That these two sets of bistable factors are interdependent means that a simultaneous exact knowledge of the two factors in either set is precluded, due to a collapse of interdependence. However, a partial or proportional collapse is not ruled out (i.e., tradeoffs).

We have used the model in Figure 2 to study human organizations making decisions under uncertainty by addressing complex situations like the environmental cleanup of its nuclear facilities undertaken by the Department of Energy, or mergers and acquisitions. The primary characteristic of this interdependence is reflected in tradeoffs between coordinating social objects communicating to solve problems while in states of uncertainty (Lawless & Grayson, 2004). In Case Study 1, we apply Organizational Uncertainty theory to a system of seven MDRCs (Medical Department Research Training Center). Our goal is to help those MDRCs become more productive in meeting their assigned mission. This means that the MDRC system would shift from a fragmented to a more ordered group of organizations, thereby increasing productivity. In the future, to exploit the power of the semantic web, we propose to use a rate equation to measure in real-time with machines the system performance, thus offering management insight as to the factors to change in a tradeoff that enhances organizational performance.

In addition, we have proposed that alignment of humans and thinking machines (agents) in an organization ranges from disordered in the lowest energy or individual state to one focused on the mission (Lawless et al., 2007). But, by focusing on the mission exclusively as in the latter case, organizations become vulnerable to change. Therefore, it is important to use feedback not only to fine tune an organization's effectiveness over the short term, but to restructure by revising its mission over the long term (Smith & Tushman, 2005). We propose that the tension can be best constructed, maintained and controlled over time by using semantic web-based metrics.

Evidence: Field

Department of Energy Citizen Advisory Boards

In our search for a classical organizational uncertainty principle, we have found in the field and confirmed in the laboratory a planning cognitive-execution tradeoff between consensus-seeking and majority rule decision-making as citizen groups made decisions over complex issues like nuclear waste management (Lawless et al., 2005). Two forms of consensus were found to exist: Worldview consensus and action consensus. The former is more likely to be derived from cooperative processes and the latter from competitive processes (Wood et al., 2008). In the first field study, we looked at the decisions of all nine of the Department of Energy's Citizen Advisory Boards as they responded to DOE's formal request to support DOE's plans to speed the shipments of transuranic wastes to its repository in New Mexico (i.e., the WIPP facility; see www.wipp.energy.gov) as part of its mission to accelerate the cleanup of DOE facilities across the U.S. These nine DOE Boards were geographically separated and located at the DOE sites where the transuranic wastes were being removed and shipped to WIPP. DOE's plans were entailed in 13 concrete recommendations and explained to the various Boards by DOE engineers (e.g., recommendation #8: "DOE in consultation with stakeholders and regulators initiate action to assure that WIPP has the capacity to accommodate all of the above listed TRU waste"). As predicted, four-fifths of DOE's majority-rule boards endorsed these recommendations, while three-fourths of its consensus-ruled boards rejected them. In addition, the time spent in deciding for majority-ruled boards was about one-fourth of the amount of time taken by the consensus-ruled boards.

In a follow-on field study of consensus decisions by the Hanford Board in Washington State and majority rule decisions at the Savannah River Site Board in South Carolina, Boards located at the two DOE sites with the largest cleanup budgets, we found that consensus rule decisions produced a cognitive congestion that resulted in behavioral "gridlock" when the single worldview of the Board conflicted with DOE's vision, increasing social volatility (Lawless & Whitton, 2007). We have found that cognitive congestion is more likely under cooperative decision making because of the inability to accept challenges to illusions (Lawless et al., 2008b). In contrast, we have found that the cognitive disambiguation from competition improves social welfare with practical decisions that feedback amplifies or accelerates.

Relative to the SRS-CAB, Bradbury and her colleagues (2003) analyzed interviews and other self-reported measures to conclude that Hanford CAB members felt very positive about their consensus-seeking process, that they very much wanted a cleaned-up environment, and they felt that DOE at its Hanford site was very responsive to their demands. However, the results from field metrics at DOE Hanford and DOE Savannah River Site (SRS) across three measures of cleanup (high-level radioactive wastes, transuranic wastes, and the environmental remediation of contaminated sites) indicated the opposite (e.g., Lawless et al., 2005). Compared to the SRS CAB and the SRS site, this difference between perceptions at the Hanford CAB and the results in the field represented an increase in risk perceptions (i.e, an unchecked increase in the number of illusions) among the Hanford CAB members that had kept them from making concrete recommendations to accelerate the environmental cleanup at Hanford.

Evidence: Laboratory

Preliminary data from a laboratory experiment nearing completion with college students making recommendations to improve their college experiences appears to have fully replicated the DOE CAB study. In this study, we asked college students in 3-person groups (N = 53 groups) at a Historically Black College and a nearby University to proposed open-ended recommendations to improve operations affecting them at their schools (e.g., with cafeteria food, library, student government, etc.). Students were randomly assigned to three-person groups who made recommendations either under consensus (CR) or majority rules (MR). Time for both types of groups was held constant. Tentatively, we predicted and found that while CR produces significantly more discussions (oscillations or $j\omega$), indicating less time available to craft recommendations, MR produces significantly more total recommendations (our analyses are ongoing).

Evidence: Case Study 1: Military Medical Department Research Training Centers (MDRCs)

Guided by our theoretical and field results in applying the organizational uncertainty principle, we have been assisting a system of seven military MDRCs (Wood et al., 2008) to become more productive; e.g., produce more research with greater scientific impact; improve patient care; and reduce the costs of care. Specifically, when we began this case study, we found little knowledge existed at the organizational level that directly linked each research product (publications, presentations, workshops) with MDRCs assigned mission. Instead, MDRC collected basic citations for each publication; not all publications were captured in its data-base; nor were all conferences attended captured in their data base.

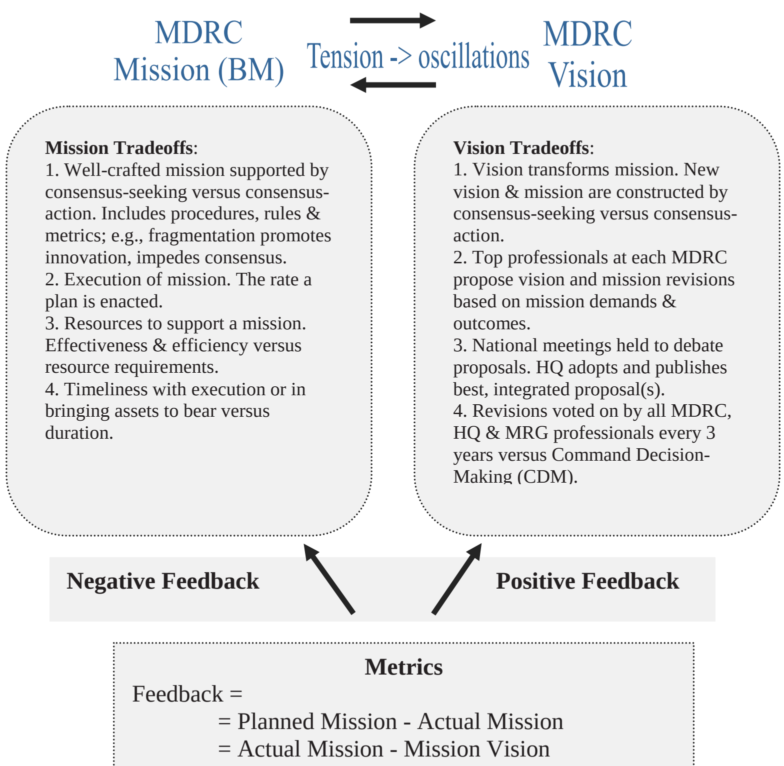
We began with a preliminary set of metrics that indicated the efficiency in meeting MDRCs mission per research protocol across the factors of scholarly activity, personnel availability, space, and funding. But at the same time, these Centers wanted to be able to transform their mission as necessary. These two goals are contradictory. But Smith and Tushman (2005) concluded that satisfying contradictory goals like these could make an organization more productive now, and more transformative in the future (see Figure 3).

Based on feedback from metrics of organizational performance linked to eIRB's, administrators have the ability to execute their mission effectively and efficiently; e.g., with Lean Six Sigma processes. But efficiency alone reduces adaptability to uncertain future missions (Smith & Tushman, 2005). Thus, concomitantly, a group internal to each MDRC and a national group of elite professionals from all MDRC units could gather annually to transform its mission, goals, and rules guided by the same feedback and metrics. As these two systems compete in a bistable relationship to control the Mission, the two systems operate in tension (interdependence), producing a natural evolution of the system.

Evidence: Case Study 1: Application of the Theoretical Model

The military (Wood, 2007) has funded a secure web-based system for one MDRC for the submission of IRB research protocols by its investigators (viz., human and animal research Institutional Review Boards). The other MDRCs are included in the product evaluation selection process in the hope that the benefits of the funded eIRB will

Figure 3. Future proposal for a semantic-web based system of seven MDRCs



The initial guidance based on theory were: Mission success makes a lean organization more efficient but also more vulnerable to change; change in a business model or its execution in reaction to environmental change was not optimum (Conant & Ashby, 1970); and a sweet spot exists where mission performance is optimum, errors are at a minimum, and at the same time the mission and the organizations it guides are modernized.

secure funding for the other sites. The first eIRB includes routing of submissions to IRB members; receipt of comments from IRB reviewers; transmission of modification requests to investigators; development of IRB meeting minutes; tracking of protocol status; automatic notification of investigators of continuing review deadlines; and tracking metrics. The technology provides a platform for collaboration across the organization between Principal Investigators and team members when drafting protocol proposals. It provides feedback among IRB reviewers, the PI and study team, and Administrators. It tracks Adverse Events (medical and drugs); provides guided electronic input and assistance and error checking and reporting to PI's and Administrators; but more importantly, it is a platform for integrated management and reporting.

The vision for this eIRB project is to achieve an end state to:

allow all research proposals, supporting documents, and scholarly products to be submitted and managed by a secured web based electronic system that allows for the real time calculation of research metrics of workload, productivity and quality. Additionally, this kind of system will allow for better management of the necessary documentation for human research protection and ensure a better environment of operational security oversight for potentially actionable medical information. This will be developed with joint execution in mind and have input from our DoD counterparts. A system that effectively captures all aspects of the research process, from protocol submission and processing to publication of scholarly products or novel therapeutics will generate the highest quality data for productivity analysis and metric development. We believe this can best be achieved by development of an electronic protocol submission and management system with the capacity to generate real time metrics of productivity and quality. (Wood, 2007, pp. 4-5)

In installing the eIRB, MDRC will be better positioned to leverage business intelligence (BI) tools that automatically pull together data for metrics with machines from this new electronic system and from other disparate database systems already in place (e.g., electronic medical records). However, only until MDRC has database systems across all aspects of biomedical research and medical care delivery and the BI tools to link these often incongruent systems together will it be able to generate real time data for semantic-web machines to study, define and improve their processes. Once in place, MDRC can make decisions in real-time rather than with data many months old thereby closing the gap between the mission and the vision and pushing the organization faster towards innovation. The natural tension and gap between the mission and vision, as it closes, will decrease the cycle time between these two perspectives propelling MDRC along the pathway of necessary transformation. We believe the ability to quickly and effectively manage knowledge is the key to organizational change.

Knowledge management is one of the fastest growing sectors in the business community. In parallel with the rapid growth of knowledge generated by automation systems, organizations having the capability and diversity of BI tools to analyze their performance against their own chosen metrics should help to accelerate system-wide transformation. These tools can afford a seamless reach across different platforms to easily allow for the automatic generation of dashboards that can visually depict metrics of organizational importance in a manner not previously available. As the present web evolves into the Semantic web, so will the capability of knowledge management with BI tools.

Current Status. A case in point to demonstrate the power of web-based technology and knowledge management has been the virtual collaboration systems used by the MDRC working group planning for an eIRB. Leaders geographically separated were able to meet approximately thirty times over almost two years and work together to solve common problems in a manner that would have been cost-prohibitive in the past. MDRC leaders from Hawaii, Washington State, Texas, Washington DC, and Georgia worked as a networked virtual organization for approximately 60 hours using web-based collaboration technology with visual and audio communication that lead ultimately to the successful funding of the eIRB system (for a review of Networked and Virtual Organizations, see Lawless et al., 2008a). Members simply logged onto the web from the convenience of their own office to participate in problem solving and closing the gap of tension between their mission and vision. Using this virtual collaboration in conjunction with a mind-mapping program (similar to a semantic network) for more effective brainstorming allowed the saving of thousands of dollars in travel and personnel time.

Assessment of Case Study One. We began Case Study 1 by contrasting the organizational performance of MDRC against the specifics listed in its assigned mission: improving patient care in the field; reducing the costs of care; and increasing the impact of research products. We found no clear link between research products and the mission; no measure of publication impacts; and no direct way to measure organizational productivity against its peers (reduced or negligible states of interdependence). In general, the organizations in the MDRC network appeared to be fragmented, with each pursuing its own path to mission success. No overarching measure of system performance existed for the MDRCs that the separate organizations could follow to guide their collective behavior. As a consequence, long-term work practices and cultural differences predominated. Subsequently, the move to adopt a web-based eIRB has set the stage to turn around the lack of organizational and system-wide knowledge. MDRC is prepared for real-time organizational and system-wide based metrics, improvements and future transformations (based on maintaining interdependent states). We believe that the semantic web can enhance these metrics by operating in real-time with data collected by machines to distinguish between classes of data sources (using OWL's vocabulary to label separately a site's physician students, physician scientists and medical scientists across the different sites, etc.). At the same time, we will be diligent in preventing web machines from either the inadvertent disclosure of patient records or the premature release, identity and location of researcher data.

Evidence: Case Study 2: Application of Theoretical Model to a College

After developing and applying metrics for a government organization whose primary mission is training military physicians in medical research practices, it was helpful to apply similar web-based metrics to an organization with a very different purpose. The subject of Case Study 2 is an organization whose primary function is higher education. Although all institutions of higher education are tasked with the production of new knowledge within fields where it offers degrees, this organization's primary purpose is to train the next generation of citizens through the use of a liberal arts curriculum. In its Vision statement, technology is highlighted and indicates that the institution "provides information technologies that link its students' total academic and social experiences to the global world." (Bradley, 2008)

Today's institutions of higher education are faced with an interesting dilemma with faculty members who have come of age during a period of tremendous technological upheaval. During the last twenty years, institutions of higher education have started making significant investments in administrative information systems. Higher education institutions are being asked by policy makers, accrediting bodies, the business community and the public for evidence that college graduates have a demonstrated knowledge base predicated on their degree. With the mounting cost of higher education, consumers are asking for accountability from colleges and universities (Bradley, 2008). Institutions of higher education as well as most organizations must focus on systems that must be in place to ensure that the decisions made in the future take advantage of the best data possible.

Institutions are engaged in a delicate dance of remaining true to their purpose in society while responding to calls for accountability for their actions. Laws such as the Family Education Rights and Privacy Act (FERPA) caused some campus officials to develop extremely strict policies regarding information about student records. These policies were strictly enforced even when it was known that the aggregation and analysis of data from student records would provide the institution with invaluable information for making informed decisions about ways to improve academic programs, increase retention, and address issues being raised by outside entities. Institutional research projects were strangled by the fear of litigation regarding the privacy of student information (Green, 2007).

According to Green (2007) "institutions of higher education have seen an emergence of a wide, rich, and mission-critical array of student and institutional services that are directly linked to core campus information services (or Enterprise Resource Planning (ERP) functions). Yet these new functions and services—alumni services, course/learning management systems, digital content, e-portfolios, e-services (online registration, fee payment), and portals—are all firmly dependent not only on the Web but also on real-time interaction with the core elements of the "old" management information system (MIS), particularly students records and institutional finances." Many of these functions at institutions, particularly small institutions are informal and units within the organization form their own fiefdoms many times as a way of managing the complexity of a system that is governed by external policies and procedures as well as the end users of the services. In an earlier age when students walked from one office to another to engage personnel in the business of enrolling in courses, acquiring financial aid, paying their bills, and obtaining housing, these systems worked. However, in an age where information drives decisions for the organization as well as the consumer, the earlier model is no longer feasible.

The organization employs approximately 200 individuals with the majority of individuals serving as instructional personnel providing instruction for a student body of less than 1,000 individuals studying at the undergraduate level. Besides instructional staff (faculty), there are administrative staff members, staff who provide support services to students, a unit that manages the fiscal enterprise of the organization, and a unit responsible for external partnerships and fund raising. All units of the institution rely on the efficient function of the other areas but are limited in the operational knowledge generated by these other units from the lack of technological (web-based) interconnectivity.

Current Status

Computing and technology support in an academic environment provide the technology infrastructure for academic and administrative activities that have become essential for the operational effectiveness of institutions of higher education. There is a need to analyze the current technology infrastructure due to the present isolation between subsystems and organizational operations. Multiple systems exist but each organizational unit works with its own "preferred" one, producing fragmentation. The different systems are not integrated causing record sharing and management problems. Currently, information technology (IT) support is being done by two staff members, one deals with hardware issues and the other with support software plus the network as part of the college's infrastructure. There is no system request form or work-list. Priority is given to network issues and calls from very important persons (VIPs) within the organization, likely impeding performance.

With a new administration, this organization has realized the need to evaluate the current IT infrastructure and the need for changes to fulfill its vision and mission. After the preliminary investigation, the first need identified was to overhaul and redesign the website. The previous version did not represent the academic organization due to its commercial feel. Then an IT inventory survey was conducted to find out what systems are available, which system is being utilized by which unit (or not at all), the merit of these choices, and the costs associated with each system. To find an enterprise-wide solution, the institution is considering having an IT-consultant company to evaluate the current infrastructure (conceptual model), and suggest the best solution. The institution also needs a chief information officer (CIO) (or MIS director) who is capable of implementing the plan.

All institutional areas that are impacted by or use technology should be evaluated. Either after purchasing an enterprise information system (EIS) or after choosing from currently available systems for a single "main" system that supports most unit functions plus a Transaction Process System (TPS) for business/financial unit online transactions, performance measurement should be enacted. Focus, however, would not be placed on the network *per se*, but on the organization's performance as measured with its EIS. Critical Success Factors for an EIS in a higher education institute like this one which should be measured are:

- Instructional support, as measured by the number of courses offered or supported via the Internet or other electronic methods, number of instructional classrooms supported, number of student computer labs, student accounts, technology in residence halls and shared spaces (i.e. campus center) or other means
- Research support, as measured by access to research databases, high speed network connectivity, other data collection and analysis mechanisms, and specialized research functions
- Cost of services, either measured in the aggregate, or on a per-student full-time equivalent (FTE) or per-faculty FTE basis, including comparisons with peer institutions
- Number of graduates compared with admission
- Student learning outcomes: assessments support

Assessment of Case Study Two

While it is too early in the process to assess this college, and while a measurable semantic-web based baseline is being built, certain areas to measure performance are already obvious. For example, after implementation of the EIS, do faculty publication numbers and the impact of research (quantity and quality) improve? Does the new IT web system improve or assist the College in its assessment processes? After the EIS system is operational, we plan to review its performance as well as the College's.

FUTURE TRENDS

The most important future trend is the use of agent-based models (ABM's) to model social and organizational effects to measure their effectiveness with the semantic web. Agent based systems have been endowed with auction based mechanisms for distributing their resources (Gibson and Troxell, 2007). In this scenario, the various entities would "bid" for the information they require, ensuring that the entity that valued the information the most would receive it in the timeliest manner for their decision making. Double auctions have been used for similar analyses with genetic algorithms (Choi, Ahn and Han, 2008).

Working with mathematics, ABM's, and artificial intelligence, the organizational uncertainty principle can be generalized to interdependent probability distribution functions with the standard deviation of Fourier transform pairs (i.e., the standard deviations from a Gaussian distribution and its Fourier transform form a "Fourier pair"; in Cohen, 1995; Rieffel, 2007). Next, we construct circuits to provide a basic model of social decision making (Yu & Efstathiou, 2002). Circuits can be modeled using virtual natural selection processes (e.g., machine learning, natural computation). Rate equations would then provide a detailed prediction of outcomes that we plan to estimate with Monte Carlo simulations. Completing the process, sensitivity analyses with the rate equation parameters provides a direct link back to the organizational uncertainty principle.

Circuits

Based on entropy measures, Yu and Efstathiou (2002) found that series network circuits underperformed parallel circuits. We expect to find that group decision-making, especially around a table, is similar to a series circuit, with subgroups or subcommittees acting like parallel circuits. However, we also expect that consensus rules (CR) will be serial and sequential, producing the time lags observed in the field and laboratory, but that majority rules (MR) with discussion drivers will act like a parallelization of neutrals, producing the time speedup also observed.

Natural Computation

Natural computation models will permit us to test field data and model the organizations that produce this data, especially the MDRC system in Case Study 1 and later the college in Case Study 2. We propose to test the data and organizational models with artificial agents evolved using biologically inspired natural selection (De Jong, 2008) and social methods of decision-making (e.g. "voting" mechanisms, ensembles). Based on our field research, we predict longer decision times and more oscillations under consensus rule (CR) than majority rule (MR). That is, we expect CR to model serial sequential individual decision processes. Surowiecki (2004) presented evidence and case studies of why agent ensembles often outperform individual experts. Earlier, Opitz and Maclin (1999) empirically showed that ensembles often outperform individuals, with theoretical support provided by Brown (2005) and Tang (2006).

Monte Carlo Simulations

Monte Carlo simulation is a technique that allows the simultaneous iteration of many uncertain variables to understand the impact of input uncertainties on one or more outcome variables. Developed during the 1940s as part of the Manhattan Project, and named after the famous casino in Monaco, Monte Carlo techniques are used today in fields ranging from manufacturing to finance, engineering and life sciences.

The basic concept in Monte Carlo simulation is that each uncertain variable, which we call a random variable, is simulated by a probability distribution. For each trial of a simulation, each random variable is sampled from its corresponding probability distribution and the sampled value is used to compute the output variable(s) for the model. Many such trials are conducted and a value is collected for each outcome variable for each simulation trial. At the conclusion of all trials a distribution of outcomes can be constructed to better understand the distribution of uncertainties for an outcome given the uncertainties in the input variables.

Rate Equation

Lawless and his colleagues (2007) devised a mathematical model of social interaction rates (this approach will allow future tests of this model constructed with machine learning using recombination operators; De Jong, 2008). We propose to adapt this model to guide our future research on organizations, e.g., training MDRC physicians with the experimental method or educating students unprepared for college courses with enhancement classes. In the latter case, the model becomes,

$$\Gamma = N_1 N_2 v_{12} \sigma_{12} \, exp \, (-\Delta A/<A>), \tag{1}$$

where Γ is the college graduation rate; N_1 the population in the group of those who have learned; N_2 those in the population who have not yet learned; v_{12} represents the velocity of knowledge passed between them, with the minimum effect occurring under censorship; σ_{12} represents how well the two groups match their beliefs, with the maximum effect occurring under social agreement (resonance); and $exp\ (-\Delta A/<A>)$ represents the probability of graduation or knowledge exchanges, where ΔA represents the energy or effort required for the knowledge to be acquired, and $<A>$ represents the average amount of effort being expended by the targeted HBCU, its professors and support staff, and its fellow students. Before we address the implications of equation (1), let's rearrange it. If χ represents the knowledge required before a student can be declared to become a graduate, then $\Gamma = \partial\chi/\partial t \approx \Delta\chi/\Delta t$, and

$$\Delta\chi = \Delta t \, N_1 N_2 v_{12} \sigma_{12} \, exp \, (-\Delta A/<A>). \tag{2}$$

From equation (2), given an average time to matriculate from the target HBCU, various opportunities exist as tradeoffs for it as an organization to improve the probability that one of its students will graduate ($\Delta\chi$) from this college. Increasing the numbers of those who actively support the student increases the occurrence of teacher-support group (N_1) to student (N_2) speech acts. Increasing the velocity (v_{12}) of knowledge passed between the two groups improves the acquisition of knowledge. Increasing the match (σ_{12}) between faculty-support groups and student groups can dramatically increase the knowledge gained (e.g., study groups; student focus groups; faculty-student focus groups; *enhancement* groups). But also the probability of graduation can be increased by reducing barriers for students ($-\Delta A$; e.g., either lowering requirements, choosing better qualified entrants, or *enhancing* the skills of the weaker entrants). Finally, by increasing the overall average effort or excitement by the HBCU directed toward learning and graduation ($<A>$), a college can strongly improve the odds that its students will graduate. Inversely, changing these factors can also decrease or adversely increase the time required for graduation.

Using the equations that we have laid out, with machines automatically collecting the data over the semantic web, we believe that real-time metrics will become possible. This information will not only be able to inform colleges or MDRCs whether they are on-target to achieve their mission as they themselves have defined it, but whether they are making progress evolving into the vision that they themselves have also proposed. With machine readable data feeding real-time metrics, organizations like MDRC will also be able to tune their performance. For the first time, we will know the actual cost of controlling their organizations to realize their benefits.

CONCLUSION

A preliminary web-based metric modeled after the plans for the new semantic web Health Care and Life Sciences Interest Group (HCL SIG) using electronic spreadsheets indicates that researcher protocol effectiveness can be established and measured as part of an organization's mission. In the metric, for theoretical reasons we have chosen the interdependent factors of planning-execution and resources-timing. As a result, the organizational uncertainty principle has proven to be a fertile source for theory and a tool to guide a system of military units in the field as they move into a new web-based collaboration system, and for a college as it begins to establish a web-based EIS system with real-time metrics. Future trends and our next steps along the path forward with natural computation, Monte Carlo simulation and Agent-Based Models (ABM's) were also reviewed. Finally, we will assure that semantic web machines do not inadvertently disclose patient records nor prematurely release data from researchers.

REFERENCES

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic.

Axsom, D., & Lawless, W. F. (1992). Subsequent behavior can erase evidence of dissonance-induced attitude change. *Journal of Experimental Social Psychology*, *28*, 387-400.

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist*, *44*(9), 1175-1184.

Bandura, A. (1977). *Social learning theory*. New York: General Learning Press.

Bankes, S. C. (2002). Perspective. Agent-based modeling. In *Proceedings of the National Academy of Sciences*, *99*(3), 7199-7200.

Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. D. (2005, January). Exploding the self-esteem myth. *Scientific American*.

Berners-Lee, T. (2007, March 27). *Video: Tim Berners-Lee on the Semantic Web*. Retrieved February 20, 2008, from http://www.technologyreview.com/Infotech/18451/

Bradley, G. (2008). *Confidential source: Presentation to the faculty by the new president of the subject college while articulating that building technology infrastructure must be one of the goals for the college*.

Bohr, N. (1955). Science and the unity of knowledge. In L. Leary (Ed.), *The unity of knowledge* (pp. 44-62). New York: Doubleday.

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: A survey and categorization. *Journal of Information Fusion, 6*, 5-20.

Busemeyer, J. (2008). *Second quantum interaction symposium (2008).* In P. Bruza, W. F. Lawless, K. von Rijsbergen, D. Sofge, B. Coecke, & S. Clark (Eds.), Oxford, UK: Oxford University.

Cacioppo, J. T., Berntson, G. G., & Crites, S. L., Jr., (Eds.). (1996). Social neuroscience: Principles, psychophysiology, arousal and response. In *Social psychology handbook of basic principles.* New York: Guilford.

Choi, J., Ahn, H., & Han, I.(2008) Utility-based double auction mechanism using genetic algorithms. *Expert Systems with Applications: An International Journal*, *34(*1), 150-158.

Coase, R. (1937). The nature of the firm. *Economica*, *4*, 386.

Cohen, L. (1995). *Time-frequency analysis: Theory and applications.* Upper Saddle River, NJ: Prentice Hall Signal Processing Series.

Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science***,** *1*(2), 899-97.

Conzelmann, G., Boyd, R., Cirillo, V., Koritarov, C., Macal, M., North, P., Thimmapuram, R. & Veselka, T. (2004). Analyzing the potential for market power using an agent-based modeling approach: Results of a detailed U.S. power market simulation. In *Proceedings of the International Conference on Computing, Communication and Control Technologies*, Austin, TX.

Crozier, M., & Friedberg, E. (1981). *Actors and systems (l'acteur et le système).* Chicago: Chicago University Press.

Csete, M. E., & Doyle, J. C. (2002). Reverse engineering of biological complexity. *Science*, *295*, 1664-69.

De Jong, K. A. (2008, February). Evolving intelligent agents: A 50 year quest. *Computational Intelligence Magazine*, *3*(1), 12-17.

Durkheim, E. (1893/1997). *The division of labor in society*. The Free Press.

Elias, N. (1969/2000). *The civilizing process (Üder den prozess der zivilisation).* Oxford, UK: Blackwell.

Green, K. C. (2007). Prodding the ERP turtle. *EDUCAUSE Review*, 148-149.

Kelley, H. H. (1992). Lewin, situations, and interdependence. *J Social Issues*, *47*, 211-233.

Körding, K. (2007). Decision theory: What "should" the nervous system do? *Science, 318*, 606-10

Latane, B. (1981). The psychology of social impact. *American Psychologist*, *36*, 343-356.

Lawless, W. F., Castelao, T., & Ballas, J. A. (2000). Virtual knowledge: Bistable reality and solution of ill-defined problems. *IEEE Systems, Man, & Cybernetics*, *30*(1), 119-124.

Lawless, W. F., & Grayson, J. M. (2004). A quantum perturbation model (QPM) of knowledge and organizational mergers. In L. van Elst & V. Dignum (Eds.), *Agent mediated knowledge management* (pp. 143-161). Berlin, Germany: Springer.

Lawless, W. F., Bergman, M., & Feltovich, N. (2005). Consensus-seeking versus truth-seeking. *ASCE Practice Periodical of Hazardous, Toxic, and Radioactive Waste Management, 9*(1), 59-70.

Lawless, W. F., Bergman, M., & Feltovich, N. (2006). The physics of organizational uncertainty: Perturbations, measurement and computational agents. In S. H. Chen, L. Jain, & C. C. Tai (Eds.), *Computational economics: A perspective from computational intelligence* (pp. 286-298). Hershey, PA: Idea Group Publishing.

Lawless, W. F., Bergman, M., Louçã, J., Kriegel, N. N., & Feltovich, N. (2007). A quantum metric of organizational performance: Terrorism and counterterrorism. *Computational & Mathematical Organizational Theory, 13*, 241-281.

Lawless, W. F., Howard, C. R., & Kriegel, N. N. (2008a). A quantum real-time metric for NVO's. In G. D. Putnik & M. M. Cunha (Eds.), *Encyclopedia of networked and virtual organizations.* Hershey, PA: Information Science Reference.

Lawless, W. F., Whitton, J., & Poppeliers, C. (2008b). Case studies from the UK and US of stakeholder decision-making on radioactive waste management. *ASCE Practice Periodical of Hazardous, Toxic, and Radioactive Waste Management*, *12*(2), 70-78.

Levine, J. M., & Moreland, R. L. (1998). Small groups. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 415-469). Boston, MA: McGraw-Hill. Luhmank, N. (1984), *Soziale systeme: Grundriß einer allgemeinen theorie*, Frankfurt: Suhrkamp.

Hamilton, A., Madison, J., & Jay, J. (1787-1788). *The federalist papers.* New York newspapers.

Mattick, J. S., & Gagen, M. J. (2005). Accelerating networks. *Science, 307,* 856-8.

May, R. M. (1973/2001). *Stability and complexity in model ecosystems.* Princeton, NJ: Princeton University Press.

Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos Science, Special Issue*, 125-130.

Montesquieu, C.-L., II. (1949). *The spirit of the laws.* (T. Nugent trans.). New York: MacMillan.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *11*, 169-198.

Parsons, T. (1966). *Societies: Evolutionary and comparative perspectives*. Englewood-Cliffs, NJ: Prentice Hall.

Pfeffer, J., & Fong, C. T. (2005). Building organization theory from first principles: The self-enhancement motive and understanding power and influence. *Organizatonal Science, 16*(4), 372-388.

Rieffel, E. G. (2007). Certainty and uncertainty in quantum information processing. In *Proceedings of the Quantum Interaction: AAAI Spring Symposium*, Stanford University. AAAI Press.

Sanfey, A. G. (2007). Social decision-making: Insights from game theory and neuroscience. *Science, 318*, 598-602.

Smith, W. K., & Tushman, M. L. (2005). Managing strategic contradictions: A top management model for managing innovation streams. *Organizational Science, 16*(5), 522-536.

Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology, 53*, 491-517.

Sukthankar, G. (2008, June 10). *Robust and eficient plan recognition for dynamic multi-agent teams*. Presentation to Information Technology Division, Naval Research Laboratory, Washington, DC.

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. Little, Brown, & Co.

Tang, E. K., Suganthan, P. N., & Yao, X. (2006). An analysis of diversity measures. *Machine Learning, 65*, 247-271.

Troxell, W., & Gerald, G. (2007). *U.S. Patent No. 20,070,124,026*. Washington, DC: U. S. Patent and Trademark Office.

Von Neumann, J., & Morgenstern, O. (1953). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

W3C-SWHCLS. (2008). *W3C Semantic Web health care and life sciences interest group*. Retrieved February 20, 2008, from http://www.w3.org/2001/sw/hcls/

Weber, M. (1904-5/1930). *The protestant ethic and the spirit of capitalism (Die protestantische ethik und der 'geist' des kapitalismus)*. (T. Parsons trans.). Scribner.

Winston, W. L., & Albright, S. C. (2007). *Practical management science* (3rd ed.). Thomson Press.

Wood, J. (2007). *PROPOSAL TITLE*. Southeast Regional Medical Command (SERMC) Institutional Review Board (IRB) Automation Solution. (PI-LTC Joseph Wood).

Wood, J., Tung, H.-L., Grayson, J., Poppeliers, C., & Lawless, W. F. (2008). A classical uncertainty principle for organizations. In M. Khosrow-Pour (Ed.), *Encyclopedia information science & technology reference* (2nd ed.). Hershey, PA: Information Science Reference.

Yu, S. B., & Efstathiou, J. (2002, April 9-10). An introduction of network complexity. In *Proceedings of Tackling Industrial Complexity: The ideas that make a difference*, Downing College, Cambridge, UK.

KEY TERMS AND DEFINITIONS

Bistability: Bistability occurs when one data set produces two mutually exclusive interpretations that cannot be held in awareness simultaneously (Cacioppo et al., 1996). Bohr (1955) concluded that multiple interpretations support the existence of different cultures. Further, given the importance of feedback to social dynamics (Lawless et al., 2007), rapid shifts between bistable perceptions increase uncertainty in the non-observed perception which not only underwrites social problems between different groups, but also supports the existence of an uncertainty principle.

Game Theory: Game theory was invented in the 1940's by Von Neuman & Morgenstern (1953). It is a one-shot game, or repeated games, played by 2 or more agents. In its most basic form, the game configuration presents two choices for payoffs to each player. Payoffs are interdependent. The values in the configuration of choices offered to participants are arbitrary and normative.

Health Care and Life Sciences Interest Group: The Semantic Web includes a Health Care and Life Sciences Interest Group (HCLSIG, 2008) to establish interoperable data standards for "connected knowledge" to improve collaboration across the health care and life sciences. The goal for HCLSIG is to reduce medical errors, increase physician efficiency and advance patient care and satisfaction. It includes document annotation and rule processing (with XML formats, OWL ontologies and SWRL rule processors).

Organizations: Organizations are social collectives performing a function that often cannot be done by an individual alone. Organizations do this by assigning interdependent roles to individuals, which requires coordinating the output of individuals, but also amplifies the capabilities of the individual working alone (Ambrose, 2001).

Organizational Uncertainty Principle: The organizational uncertainty principle acts as a tradeoff in attention directed at reducing the uncertainty in one factor, such as a worldview, with the result that the uncertainty in a second interdependent factor is increased inversely. It is based on Bohr's (1955) famous notion that the uncertainty principle at the atomic level applied to social situations is captured by human action and observation. That is, the more focused individuals are on acting out a series of steps, the less observant they become of their action. Applied to societies, action-observation uncertainty couples that open the path to multiple interpretations of the same social behavior lie at the root of different cultures.

Semantic Web: The Semantic Web is an on-going project to extend the World Wide Web (WWW) to permit humans and machines to collaborate efficiently. As envisioned by Berners-Lee (2007), inventor of WWW (and web languages URI, HTTP, and HTML), the future Web should evolve into a universal exchange for data, information and knowledge. Without a universal standard for machine access, HTML data is difficult to use on a large scale. The Semantic Web solves this problem with an efficient global mesh for information access by humans and machines.

Social Learning Theory: SLT is a term coined by Bandura (1977) that includes the three different schools of ideas that accounted for learning by organisms, but with a primary focus on humans. These three schools were classical conditioning (Pavlovian associations), operant conditioning (Skinnerian rewards and punishments), and modeling, Bandura's own school of thought.

This work was previously published in Handbook of Research on Social Dimensions of Semantic Technologies and Web Services, edited by M. M. Cruz-Cunha; E. F. Oliveira; A. J. Tavares; L. G. Ferreira, pp. 469-488, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 7.2
Bridging the Gap between Mobile Application Contexts and Web Resources

Stefan Dietze
Open University, UK

Alessio Gugliotta
Open University, UK

John Domingue
Open University, UK

ABSTRACT

Context-awareness is highly desired, particularly in highly dynamic mobile environments. Semantic Web Services (SWS) address context-adaptation by enabling the automatic discovery of distributed Web services based on comprehensive semantic capability descriptions. Even though the appropriateness of resources in mobile settings is strongly dependent on the current situation, SWS technology does not explicitly encourage the representation of situational contexts. Therefore, whereas SWS technology supports the allocation of resources, it does not entail the discovery of appropriate SWS representations for a given situational context. Moreover, describing the complex notion of a specific situation by utilizing symbolic SWS representation facilities is costly, prone to ambiguity issues and may never reach semantic completeness. In fact, since not any real-world situation completely equals another, a potentially infinite set of situation parameters has to be matched to a finite set of semantically defined SWS resource descriptions to enable context-adaptability. To overcome these issues, the authors propose Mobile Situation Spaces (MSS) which enable the description of situation characteristics as members in geometrical vector spaces following the idea of Conceptual Spaces (CS). Semantic similarity between situational contexts is calculated in terms of their Euclidean distance within a MSS. Extending merely symbolic SWS descriptions with context information on a conceptual level through MSS enables similarity-based matchmaking between real-world situation characteristics and predefined resource representations as part of SWS descriptions. To prove the feasibility, the authors provide a proof-of-concept prototype which applies MSS to support context-adaptation across distinct mobile situations.

DOI: 10.4018/978-1-60566-290-9.ch009

INTRODUCTION

Current and next generation wireless communication technologies will encourage a widespread use of available resources – data and services - via a broad range of mobile devices resulting in the demand for a rather context-adaptive resource retrieval. Context-adaptation is a highly important feature across a wide variety of application domains and subject to intensive research throughout the last decade (Dietze, Gugliotta & Domingue, 2007; Schmidt & Winterhalter, 2004; Gellersen, Schmidt & Beigl, 2002). Whereas the context is defined as the entire set of surrounding situation characteristics, each individual situation represents a specific state of the world, and more precisely, a particular state of the actual context (Weißenberg, Gartmann & Voisard, 2006). Particularly, a situation description defines the context of a specific situation, and it is described by a combination of situation parameters, each representing a particular situation characteristic. Following this definition, context-adaptation can be defined as the ability of Information Systems (IS) to adapt to distinct possible situations.

To achieve this, we base on a promising technology for distributed and highly dynamic service oriented applications: Semantic Web Services (SWS). SWS technology (Fensel et al., 2006) addresses context-adaptation by means of automatic discovery of distributed Web services as well as underlying data for a given task based on comprehensive semantic descriptions. First results of SWS research are available in terms of reference ontologies – e.g. OWL-S (Joint US/EU ad hoc Agent Markup Language Committee, 2004) and WSMO (WSMO Working Group, 2004) – as well as comprehensive frameworks (e.g. DIP project[1] results). However, whereas SWS technology supports the allocation of appropriate services for a given goal based on semantic representations, it does not entail the discovery of appropriate SWS goal representations for a given situation. Particularly in mobile settings, the current situation of a user heavily determines the intentional scope behind a user goal and consequently, the appropriateness of particular resources. For instance, when attempting to retrieve localized geographical information, the achievement of a respective goal has to consider the location and device of the user.

Despite the strong impact of a (mobile) context on the semantic meaning and intention behind a user goal, current SWS technology does not explicitly encourage the representation of domain situations. Furthermore, the symbolic approach - describing symbols by using other symbols without a grounding in the real world - of established SWS and Semantic Web (SW) representation standards in general, such as RDF (World Wide Web Consortium, W3C, 2004a), OWL (World Wide Web Consortium, W3C, 2004b), OWL-S (Joint US/EU ad hoc Agent Markup Language Committee, 2004), or WSMO (WSMO Working Group, 2004), leads to ambiguity issues and does not entail semantic meaningfulness, since meaning requires both the definition of a terminology in terms of a logical structure (using symbols) and grounding of symbols to a conceptual level (Cregan, 2007; Nosofsky, 1992).Moreover, while not any situation or situation parameter completely equals another, the description of the complex notion of a specific situation in all its facets is a costly task and may never reach semantic completeness. Apart from that, to enable context-adaptability, a potential infinite set of (real-world) situation characteristics has to be matched to a finite set of semantically defined parameter representations. Therefore, we claim, that fuzzy classification and matchmaking techniques are required to extend and exploit the current functionalities provided by SWS and match the specific requirements of context-aware mobile applications.

Conceptual Spaces (CS), introduced by Gärdenfors (Gärdenfors, 2000; Gärdenfors, 2004) follow a theory of describing entities at the conceptual level in terms of their natural characteristics similar to natural human cognition in order to avoid the symbol grounding issue. CS enable representation of objects as vector spaces within a geometrical space which is defined through a set of quality dimensions. For instance, a particular color may be defined as point described by vectors measuring the quality dimensions hue, saturation, and brightness. Describing instances as vector spaces where each vector follows a specific metric enables the automatic calculation of their semantic similarity, in terms of their Euclidean distance, in contrast to the costly representation of such knowledge through symbolic SW representations. Even though several criticisms have to be taken into account when utilizing CS (Section 0) they are considered to be a viable option for knowledge representation.

In this chapter, we propose Mobile Situation Spaces (MSS) as a specific derivation of Conceptual Situation Spaces (CSS). MSS utilize CS to represent situations and are mapped to standardized SWS representations to enable first, the situation-aware discovery of appropriate SWS descriptions and finally, the automatic discovery and invocation of appropriate Web services to achieve a given task within a particular situation. Extending merely symbolic SWS descriptions with context information on a conceptual level through MSS enables a fuzzy, similarity-based matchmaking methodology between real-world situation characteristics and predefined SWS representations within mobile environments. Since semantic similarity between situation parameters within a MSS is indicated by the Euclidean distance between them, real-world situation parameters are classified in terms of their distance to predefined prototypical parameters, which are implicit elements of a SWS description. Whereas current SWS technology addresses the issue of allocating services for a given task, our approach supports the discovery of SWS task representations within a given mobile situation. Consequently, the expressiveness of current SWS standards is extended and fuzzy matchmaking mechanisms are supported.

To prove the feasibility of our approach, a proof-of-concept prototype is provided which uses MSS to support context-adaptation by taking into account context parameters such as the current location and desired knowledge subject.

The paper is organized as follows. The following Section 2 provides background information on SWS, whereas Section 3 introduces our approach of Conceptual Situation Spaces which are aligned to current SWS representations. Section 4 illustrates the application of CSS to mobile settings by introducing MSS. Utilizing MSS, we introduce a context-adaptive prototype in Section 5. Finally, we conclude our work in Section 6 and provide an outlook to future research.

SEMANTIC WEB SERVICES AND WSMO

SWS technology aims at the automatic discovery, orchestration and invocation of distributed services for a given user goal on the basis of comprehensive semantic descriptions. SWS are supported through representation standards such as *WSMO* and *OWL-S*. We refer to the *Web Service Modelling Ontology (WSMO)*, a well established SWS reference ontology and framework. The conceptual model of WSMO defines the following four main entities:

- *Domain Ontologies* provide the foundation for describing domains semantically. They are used by the three other WSMO elements. WSMO domain ontologies not only support Web service related knowledge representation but semantic knowledge representation in general.
- *Goals* define the tasks that a service requester expects a Web service to fulfill. In this sense they express the requester's intent.
- *Web service* descriptions represent the functional behavior of an existing deployed Web service. The description also

outlines how Web services communicate (*choreography*) and how they are composed (*orchestration*).
- *Mediators* handle data and process interoperability issues that arise when handling heterogeneous systems.

WSMO is currently supported through several software tools and runtime environments, such as the *Internet Reasoning Service IRS-III* (Cabral et al., 2006) and WSMX (WSMX Working Group, 2007). IRS-III is a *Semantic Execution Environment (SEE)* that also provides a development and broker environment for SWS following WSMO. IRS-III mediates between a service requester and one or more service providers. Based on a client request capturing a desired outcome, the goal, IRS-III proceeds through the following steps utilizing the set of SWS capability descriptions:

1. Discovery of potentially relevant Web services.
2. Selection of set of Web services which best fit the incoming request.
3. Invocation of selected Web services whilst adhering to any data, control flow and Web service invocation constraints defined in the SWS capabilities.
4. Mediation of mismatches at the data or process level.

In particular, IRS-III incorporates and extends WSMO as core epistemological framework of the IRS-III service ontology which provides semantic links between the knowledge level components describing the capabilities of a service and the restrictions applied to its use.

However, even though SWS technologies enable the dynamic allocation of Web services for a given goal, it does not consider the adaptation to different user contexts. In order to fully enable context-aware discovery of resources as required by mobile settings (Section 1), the following shortcomings have to be considered:

I1. *Lack of explicit notion of context*: current SWS technology does not entirely specify how to represent domain contexts. For example, WSMO addresses the idea of context: Goal and web service represent the user and provider local views, respectively; the domain ontologies define the terminologies used in each view; and the mediators are the semantic bridges among such distinct views. However, WSMO does not specify what a context description should define and how the context elements should be used.

I2. *Symbolic Semantic Web representations lack grounding to conceptual level*: the symbolic approach, i.e. describing symbols by using other symbols, without a grounding in the real world, of established SWS, and Semantic Web representation standards in general, leads to ambiguity issues and does not entail semantic meaningfulness, since meaning requires both the definition of a terminology in terms of a logical structure (using symbols) and grounding of symbols to a conceptual level (Cregan, 2007; Nosofsky, 1992).

I3. *Lack of fuzzy matchmaking methodologies*: Describing the complex notion of a specific situation in all its facets is a costly task and may never reach semantic completeness. Whereas not any situation and situation parameter completely equals another, the number of (predefined) semantic representations of situations and situation parameters is finite. Therefore, a possibly infinite set of given (real-world) situation characteristics has to be matched to a finite set of predefined parameter instance representations which are described within an IS. Consequently, fuzzy classification and matchmaking techniques are required to classify a real-world situation based on a limited set of predefined parameter descriptions.

CONCEPTUAL SITUATION SPACES

To address the issues I1 - I3 introduced in Section 0, we propose *Mobile Situation Spaces (MSS)* as a setting-specific realisation of our metamodel for *Conceptual Situation Spaces (CSS)* (Dietze, Gugliotta & Domingue, 2008).

CSS Formalisation

CSS enable the description of a particular situation as a member of a dedicated CS. As defined in (Weißenberg et al., 2006) a situation is defined as:

$$S^n = \left\{ \left(t_1, t_2, cp_1, cp_2, .., cp_n \right) \middle| cp_i \in CP \right\}$$

Where t_1 is the starting time of a situation, t_2 represents the end time of a situation and cp_i being situation parameters which are invariant throughout the time interval defined through t_1 and t_2. Referring to (Gärdenfors, 2004; Raubal, 2004), we define a CSS (*css:Conceptual Situation Space* in Figure 1) as a vector space:

$$C^n = \left\{ \left(c_1, c_2, ..., c_n \right) \middle| c_i \in C \right\}$$

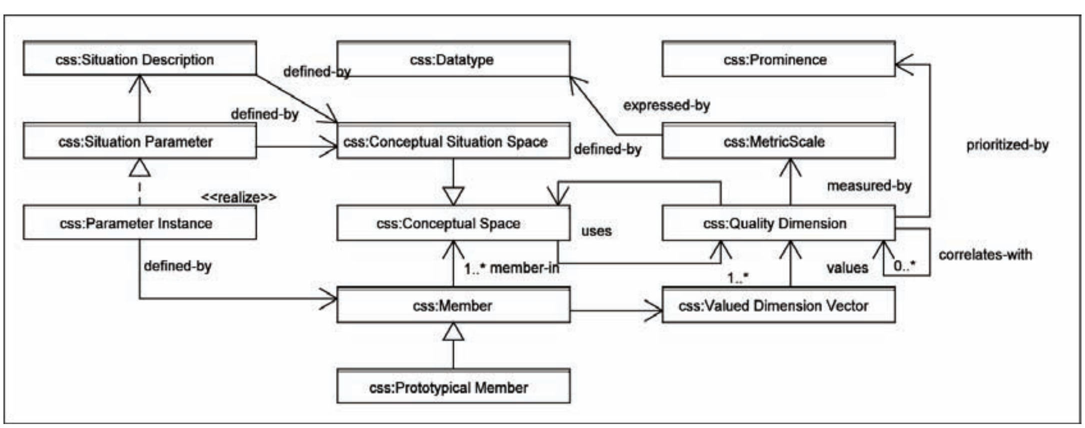
with c_i being the quality dimensions *(css:Quality Dimension)* of *C*. In that, a CSS *C* represents a particular situation *S* whereas its situation parameters cp_i are represented through certain quality dimensions c_i. Please note, that we do not distinguish between dimensions and domains - beings sets of integral dimensions (Gärdenfors, 2004) - but enable dimensions to be detailed further in terms of subspaces. Hence, a dimension within one space may be defined through another conceptual space by using further dimensions (Raubal, 2004). In such a case, the particular quality dimension c_j is described by a set of further quality dimensions with

$$c_j = D^n = \left\{ \left(d_1, d_2, ..., d_n \right) \middle| d_k \in D \right\}.$$

In this way, a CSS may be composed of several subspaces and consequently, the description granularity of a specific situation can be refined gradually. To reflect the impact of a specific quality dimension on the entire CSS, we consider a prominence value *p (css:Prominence)* for each dimension. Therefore, a CSS is defined by

$$C^n = \left\{ \left(p_1 c_1, p_2 c_2, ..., p_n c_n \right) \middle| c_i \in C, p_i \in P \right\}$$

Figure 1. The CSS metamodel.



where *P* is the set of real numbers. However, the usage context, purpose and domain of a particular CSS strongly influence the ranking of its quality dimensions. This clearly supports our position of describing distinct CSS explicitly for specific domains only.

Particular members (*css:Member*) in the CSS are described through a set of valued dimension vectors (*css:Valued Dimension Vectors*). Symbolic representations of domain situations and parameters, such as *css:Situation Description* and *css:Situation Parameter*, refer to particular CSS *(css:Conceptual Situation Space)* whereas parameter instances are represented as members *(css:Member)*.

Moreover, referring to Gärdenfors (2004) we consider prototypical members (*css:Prototypical Member*) within a particular space. Prototypical members enable the classification of any arbitrary member *m* within the a specific CSS, by simply calculating the Euclidean distances between *m* and all prototypical members in the same space to identify the closest neighbours of *m*. For instance, given a CS to describe apples based on their shape, taste and colour, a green apple with a strong and fruity taste may be close to a prototypical member representing the typical characteristics of the Granny Smith species. Figure 1 depicts the CSS metamodel.

The metamodel introduced above has been formalized into a *Conceptual Situation Space Ontology (CSSO)*, utilizing OCML (Motta, 1998). In particular, each of the depicted entities is represented as a concept within CSSO whereas associations are reflected as their properties in most cases. The correlation relationship indicates whether two dimensions are correlated or not. For instance, when describing an apple the quality dimension describing its sugar content may be correlated with the taste dimension. Information about correlation is expressed within the CSSO through axioms related to a specific quality dimension instance. CSSO is aligned to a well-known foundational ontology: the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) (Gangemi, Guarino, Masolo, Oltramari, Schneider, 2002) and, in particular, its module Descriptions and Situations (D&S) (Gangemi, Mika, 2003). The aspect of gradually refining a CSS through subspaces corresponds to the approach of DOLCE D&S to gradually refine a particular description by using parameters where each parameter can be described by an additional description.

With respect to (Raubal, 2004), we define the semantic similarity between two members of a space as a function of the Euclidean distance between the points representing each of the members. However, we would like to point out, that distinct distance metrics, such as the Taxicab or Manhattan distance (Krause, 1987), could be considered, even though the nature of the space and its possible metrics suggests the Euclidean distance as a useful metric to calculate similarities. Applying a formalization of CS proposed in Raubal (2004) to our definition of a CSS, we formalize the Euclidean distance between two members in a CSS as follows. Given a CSS definition *C* and two members represented by two vector sets *V* and *U*, defined by vectors v_0, v_1, …,v_n and u_1, u_2,…,u_n within *C*, the distance between *V* and *U* can be calculated as:

$$\left|d(u,v)\right|^2 = \sum_{i=1}^{n} (z(u_i) - z(v_i))^2$$

where $z(u_i)$ is the so-called Z-transformation or standardization (Devore, Peck, 2001) from u_i. Z-transformation facilitates the standardization of distinct measurement scales which are utilized by different quality dimensions in order to enable the calculation of distances in a multi-dimensional and multi-metric space. The z-score of a particular observation u_i in a dataset is calculated as follows:

$$z(u_i) = \frac{u_i - \bar{u}}{s_u}$$

where $\bar{u}$ is the mean of a dataset *U* and s_u is the standard deviation from *U*. Considering prominence values p_i for each quality dimension *i*, the Euclidean distance *d(u,v)* indicating the semantic similarity between two members described by vector sets *V* and *U* can be calculated as follows:

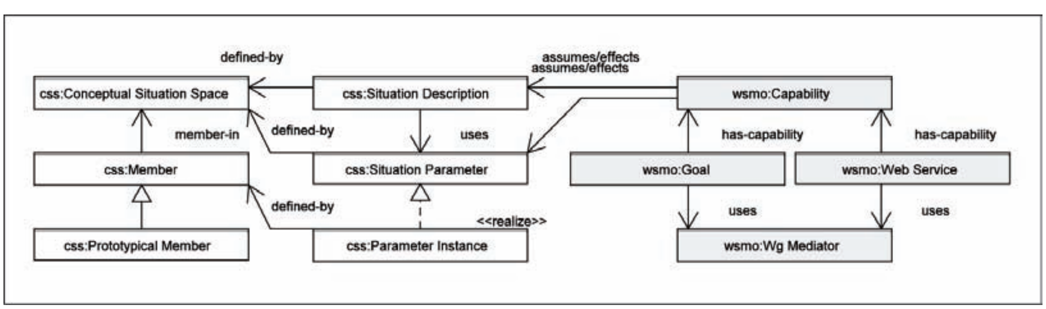
$$d(u,v) = \sqrt{\sum_{i=1}^{n} p_i((\frac{u_i - \bar{u}}{s_u}) - (\frac{v_i - \bar{v}}{s_v}))^2}$$

Utilizing CSS for SWS Selection

Whereas the discovery of distributed Web services for a given user goal is addressed by current SWS technology, such as WSMO, and corresponding reasoners, the context-aware selection of a specific SWS goal representation for a given situation is a challenging task to be tackled when developing SWS-driven applications. By providing an alignment of CSS and SWS, we address this issue by enabling the classification of an individual situation along predefined situation descriptions - used within SWS descriptions - based on semantic similarity calculation. Therefore, CSS are aligned to WSMO to support the automatic discovery of the most appropriate goal representation for a specific situation. Since both metamodels, WSMO as well as CSS, are represented based on the OCML representation language (Motta, 1998), the alignment was accomplished by defining relations between concepts of both ontologies as depicted in Figure 2.

Grey colored concepts in Figure 2 represent concepts of WSMO. A goal description (*wsmo:Goal*) utilizes particular situation parameters (*css:Situation Parameters*) to semantically describe its capabilities, i.e. its assumptions, effects, preconditions and postconditions in terms of semantic situation descriptions *(css:Situation Description)*. A WSMO runtime reasoning engine utilizes capability descriptions to identify SWS (*wsmo:Web Service*) which suit a given Goal. In contrast, the preliminary selection of the most appropriate goal description for a given situation is addressed by classification of situation parameters through CSS. For instance, given a set of real-world situation parameters, described as members in a CSS, their semantic similarity with predefined prototypical parameters *(css:Prototypical Member)* is calculated. Given such a classification of a particular real-world situation, a goal representation which assumes matching prototypical parameter instances is selected and achieved through the reasoning engine.

Figure 2. Alignment of CSS and WSMO.



Deriving CSS for Certain Application Contexts

As stated in Gärdenfors (2000), the definition and prioritization of quality dimensions within a CS is highly dependent on the purpose and context of the space. For instance, when describing an apple, dimensions may be differently weighted, dependent on whether the apple is subject to visual cognition exclusively or to full sensory perception, what would be the case if the apple is supposed to be eaten. Whereas in the first case, dimensions such as color and shape are highly ranked, taste and texture may additionally be important in the latter case.

Consequently, the derivation of an appropriate space for a certain purpose is considered an important task which usually should be carried out by a qualified individual such as an application designer. We particularly foresee a procedure consisting of the following steps:

S1. Identification of situation parameters eligible for representation as quality dimension c_i.
S2. Assignment of prominence values p_i to each quality dimension c_i
S3. Assignment of metrics to each quality dimension c_i.

With respect to *S1*, one has to take into account which aspects of a situation are relevant from an application perspective, i.e. which characteristics have an impact on the applied context adaptation strategy or rules. In the case of our intended usage of CSS for SWS selection, only parameters are important, which are considered within SWS capability representations (Section 0).

Since several dimensions might have a different impact factor on the entire space, *S2* is aimed at assigning a prominence value p_i to each dimension c_i. Prominence values should usually be chosen from a predefined value range, such as 0...1. However, since the assignment of prominences to quality dimensions is of major importance for the semantic meaning of calculated distances within a space, this step is not straightforward and most probably requires ex post re-adjustment.

During the final step *S3*, a quantitative metric has to be assigned to each previously defined dimension. Whereas certain dimensions naturally are described using qualitative measurements, such as a size or a weight, other dimensions are usually described using rather qualitative values. The latter applies for instance to the notion of a color. In case no quantitative metric can be assigned to a certain quality dimension c_i, a subspace has to be defined which refines the particular dimension through further dimensions. For instance, in the case of the color dimension, a subspace could be defined using the quantitative dimensions hue, saturation and brightness. Hence, the proposed procedure has to be repeated iteratively until a sufficient description depth has been achieved leading to the definition of a CSS *C* of the form (Section 0):

$$C^n = \left\{ \left(p_1 c_1, p_2 c_2, \ldots, p_n c_n \right) \middle| c_i \in C, p_i \in P \right\}$$

A MOBILE SITUATION SPACE

Following the steps introduced in Section 0, we derive a CSS aimed at representing situations in mobile settings. A mobile situation is defined by parameters such as the technical environment used by a user, his/her current objectives and particularly the current location. Since each of these parameters apparently is a complex theoretical construct, most of the situation parameters cannot be represented as a single quality dimension within the CSS, but have to be represented as dedicated subspaces which are defined by their very own dimensions (Section 0). Moreover, applying CSS to represent a particular concept is only reasonable in cases where similarity calculation is possible and semantically meaningful, i.e. a particular measurement can be applied to each quality dimension. For instance, the native language of a user is a crucial important situation parameter, but in this case, only a direct match is reasonable in order to provide appropriate information resources in the correct language to the user.

Therefore, this section focuses exemplarily on the representation of two parameters through a CSS subspace, which are of particular interest: the *location* and the *subject* a user is interested in. Due to the complex and diverse nature of a particular subject or spatial location, traditional symbolic representation approaches of the Semantic Web are supposed to fail since it is nearly impossible to define either a subject or a location in a non-ambiguous and comprehensive way by just following a symbolic approach.

Moreover, a one-to-one matchmaking between different locations and subjects is hard to achieve, since fairly not any instance of these parameters completely equals another one. Therefore, fuzzy similarity detections, as enabled through MSS, have to be utilized.

To represent spatial locations, we define a CSS subspace L with 2 quality dimensions l_i representing the latitude and longitude of a particular location

$$L^2 = \left\{ \left(l_1, l_2 \right) \middle| l_i \in L \right\}$$

In order to represent a particular subject, we currently consider 4 dimensions (history, geography, culture, languages) which are used to describe the semantic meaning of a particular subject within subspace S:

$$S^4 = \left\{ \left(s_1, s_2, s_3, s_4 \right) \middle| s_i \in L \right\}$$

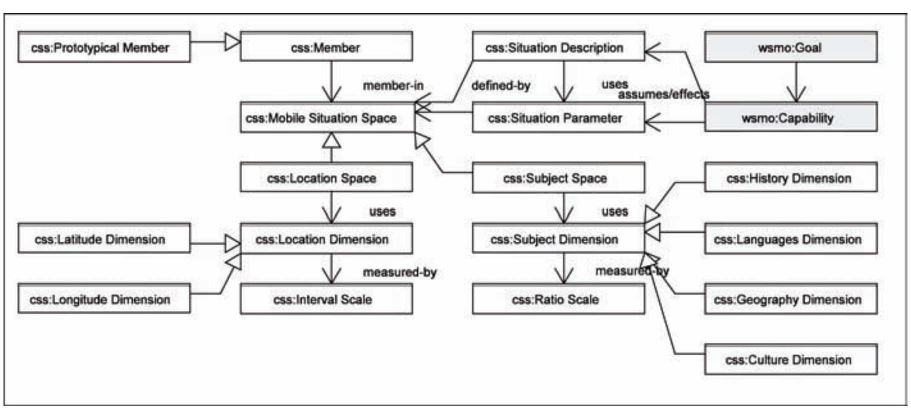
Figure 3 depicts the key concepts of the ontology describing L and S as subspaces (*css:Location Space, css:Subject Space*) within the mobile space *(css:Mobile Situation Space)*.

Moreover, Figure 3 depicts the relation of the subspace L (*css:Location Space*) and subspace S (*css:Subject Space*) with WSMO-based SWS descriptions, represented via grey-colored concepts (Section 0).

Instances of a situation parameter representing a subject are defined by particular members within the space S (*css:Subject Space*), which itself uses 4 quality dimension c_i, whereas instances of a parameter representing a spatial location are defined by members within the space L (*css:Location Space*), which itself uses 4 quality dimension $l_{i.}$ The metric scale, datatype and value range for each dimension s_i and l_i are presented in Table 1:

As depicted in Table 1, each quality dimension l_i is ranked on an interval scale with value ranges being float numbers between -90 and +90 in case of the latitude and between -180 and +180

Figure 3. Key concepts representing mobile situation subspaces.



in case of the longitude. Furthermore, each quality dimension c_i is ranked on a ratio scale with value ranges being float numbers between 0 and 100. The authors would like to highlight, that no prominence values have been assigned since each dimension has an equal impact to define a particular member. It is obvious, that the assignment of prominence values is a highly subjective process, strongly dependent on the purpose, context and individual preferences. Therefore, future work is aimed at enabling users to assign rankings of quality dimensions themselves in order to represent their individual priorities regarding the service retrieval process.

To classify an individual mobile situation, we define prototypical members (*css:Prototypical Member*) in the Mobile Situation Space. For instance, to describe particular cities as members within *L*, we utilized geodata, retrieved from GoogleMaps[2], to describe a prototypical member for each location which is targeted by a particular SWS. A few examples of prototypical location members used in the current prototype application are represented in Table 2:

An example of how such parameters are represented in a formal knowledge modeling language is given in Section 0. Moreover, we predefined several prototypical subjects in *S*, each representing the maximum value of a particular quality dimension s_i what resulted in the following 4 prototypical subjects.

Apart from the depicted subjects, each subject which is described as part of a symbolic SWS capability representation had been referred to an individual member in *S*.

Table 1. Metric scale, range, and data type of quality dimensions l_i and s_i.

	Quality Dimension	Metric Scale	Data-type	Range
l_1	Latitude	Interval	Float	-90..+90
l_2	Longitude	Interval	Float	-180..+180
s_1	History	Ratio	Float	0..100
s_2	Culture	Ratio	Float	0..100
s_3	Geography	Ratio	Float	0..100
s_4	Language	Ratio	Float	0..100

Table 2. Prototypical members within L.

Prototype	l_1 (Latitude)	l_2 (Longitude)
L1: Milton Keynes (UK)	52.044041	-0.699569
L2: London (UK)	51.500152	-0.126236
L3: Brighton (UK)	50.820931	-0.139846
L4: Paris (FR)	48.85667	2.350987
L5: Toulouse (FR)	43.604363	1.442951

Table 3. Prototypical members within S.

Prototype	s_1	s_2	s_3	s_4
S1: History	**100**	0	0	0
S2: Culture	0	**100**	0	0
S3: Geography	0	0	**100**	0
S4: Languages	0	0	0	**100**

SIMILARITY-BASED SWS SELECTION AND ACHIEVEMENT IN A MOBILE SETTING

To prove the feasibility of our approach, we provide a proof-of-concept prototype application, which utilizes MSS (Section 4) - based on the CSS metamodel introduced in Sections 0 - and supports context-adaptation in a mobile environment based on SWS and CSS.

Runtime Support for CSS and SWS

The following Figure 4 depicts the general architecture adopted to support reasoning on MSS and SWS in distinct domain settings through a Semantic Execution Environment (SEE), which in our case is IRS-III (Section 0).

Multiple mobile devices - such as PDAs, mobiles or any other portable device hosting a Web browser - can serve as user interface of the

SEE, enabling the user (and the device itself) to provide information about his/her goal and the current real-world situation.

The SEE makes use of semantic representations of the CSS formalisation (CSS ontology, CSSO), specifically derived for mobile settings, and of SWS annotations based on WSMO in order to discover and allocate the most appropriate resource for a given user goal within a current situation. Ontologies had been represented using the OCML knowledge modeling language (Motta, 1998).

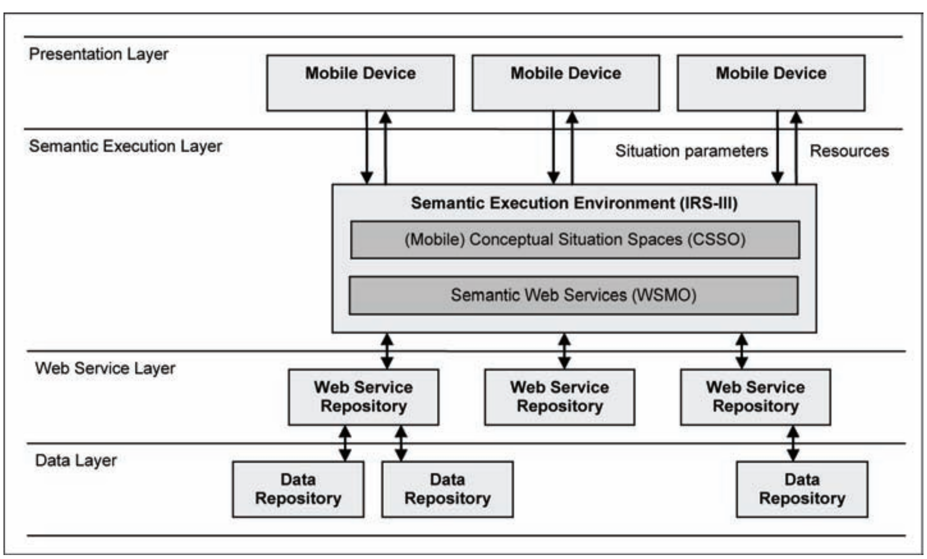
WSMO capabilities are represented by defining the assumptions and effects of available SWS and goals in terms of certain situation description or situation parameter instances (Section 0). Such situation descriptions are refined as particular prototypical members of an associated CSS, such as prototypical members of the MSS S and L introduced in Section 4.

As mentioned in Section 3, CSSO allows us to describe a specific mobile situation description instance in terms of a collection of situation parameter instances. Mobile situation description instances are automatically and gradually defined at runtime by the SEE as the result of the user interaction with the mobile device. On the basis of the detected context parameters, the SEE performs the following steps:

1. Computation of similarities between the detected real-world context parameters—obtained from the user and its device—and symbolic representation of prototypical situation parameters;
2. Progressive update of the current mobile situation description with the closest prototypical situations parameters;
3. Determination of (WSMO) goal matching the refined situation description;
4. Achievement of selected goals by means of discovery and orchestration of available web services.

Consequently, we enable the classification of real-world context parameters along available predefined parameters in order to enable a similarity-based selection and orchestration WSMO goals.

Figure 4. Architecture to support runtime reasoning on CSS and SWS.



Context Classification and Adaptation

As outlined in the previous section, the SEE automatically detects the semantic similarity of specific situation parameters with a set of predefined prototypical parameters to enable the allocation of context-appropriate resources. In this section, we further detail these aspects, since they are central in the contribution of this chapter. In particular, we specify the concepts of classification and adaptation.

Referring to CSS subspaces L and S described in Section 0, given a particular member U in L or S, its semantic similarity with each of the prototypical members is indicated by their Euclidean distance. Since we utilize spaces described by dimensions which each use the same metric scale and no prominence value, the distance between two members U and V can be calculated disregarding a Z-transformation (Section 0) for each vector:

$$d(u,v) = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2}$$

Please note, that it would be possible to calculate distances either between entire situations (members within *css:Mobile Situation Space*) or between particular parameter (members in subspaces such as L and S). Since individual semantic similarities between instances of parameters such as the current location or the desired subject are usually important knowledge when deciding about the appropriateness of resources for a given context, the current application calculates distances between each parameter, i.e. between members within each individual subspace.

The calculation of Euclidean distances using the formula shown above is performed by a standard Web service, which is annotated as SWS and invoked through IRS-III at runtime. Given a particular CSS description, a member (representing a specific parameter instance) as well as a set of prototypical member descriptions (representing prototypical parameter instances), similarities are calculated by the Web service at runtime in order to classify a given situation parameter.

For instance, a user is currently located in Eastbourne (UK) and is interested in historical information about the surrounding area. Consequently, the particular situation description (*css:MobileSituation Desccription*) includes a location parameter which is defined by a member E in the specific location space (*css:Location Space*) with the following vectors describing latitude and longitude of Eastbourne:

$$E = \left\{ \left(e_1 = 50.766868, e_2 = 0.284804 \right) \middle| e_i \in L \right\}$$

To represent the current aim of the user, a user selects one of the subject prototypes (Section 0), in this case *S1* (Table 3), which is added to the situation description.

Figure 5 depicts a screenshot of a mobile device showing the application web-interface while supporting a user to semi-automatically locate him-/herself utilizing geodata dynamically retrieved from GoogleMaps. By providing incomplete knowledge about the current location, for instance the current city, full geospatial datasets, including the latitude and longitude of a location, are retrieved dynamically to enable similarity-based location matchmaking.

Figure 5. Mobile device showing semi-automatic location detection.

Based on the current situation description, SWS are selected which are able to address the situation. Whereas parameters which are not defined by members in a specific CSS require a direct match with a corresponding SWS description, a similarity-based match is computed for parameters which are described in a CSS, e.g. the location or the subject. Hence, distance calculation was utilized to identify similarities between current context parameters – such as *E* and *S1* – and prototypical parameters which had been defined as part of SWS capability descriptions in order to represent the parameters targeted by available SWS. In order to illustrate the representation of prototypical CSS members, the following OCML code defines a location parameter instance representing the geospatial location Brighton, as well as the respective prototypical member (*L3*) in the MSS *L*.

Calculating distances between *E* and targeted locations – represented as prototypical MSS members - led to the identification of the following distances to the three closest matches: (Table 4)

Since not any SWS targets historical interests (*S1*) exclusively – as desired by the user - no direct match between the situation and subjects targeted by available SWS was achieved. However, similarity calculation identified related subject areas, which partially target historical information. Table 5 indicates their vectors and distances to the required subject *S1*.

The subjects *S5*, *S6* and *S7* as well as the locations *L1*, *L2*, and *L3* shown in Table 4 and Table 5 had been described as prototypical members in the MSS (Section 0) during the development of SWS representations targeting certain subjects and locations. By following our alignment from Section 0, this task could be performed by either the Web service provider or any SWS expert who is providing and publishing a semantic representation of available Web services.

As indicated by the Euclidean distances depicted in Tables 4 and 5, the closest matching SWS provides historical and cultural (*S7*) resources for the Brighton (*L3*) area, as these show the lowest distances. Provided these similarities, a user is able to select predefined parameters that best suit his/her specific preferences within the current situation. In that, the use of similarity-based classification enables the gradual refinement of

Listing 1. Partial OCML code defining location parameter instance and respective MSS member.

```
(def-instance brighton-location location
    ((has-instance-title “Brighton”)
    (defined-by p2-location-brighton)))

(def-instance p2-location-brighton location-prototypical-member
    ((has-title “Location-Brighton “)
    (has-description “Prototype describing Brighton”)
    (member-in location-space)
(has-valued-dimension (brighton-valued-lat-vector brighton-valued-long-vector))))

(def-instance brighton-valued-lat-vector location-valued-dimension-vector
    ((values latitude-dimension)
     (has-value 50.820931)))

(def-instance brighton-valued-long-vector location-valued-dimension-vector
    ((values longitude-dimension)
     (has-value -0.139846)))
```

Table 4. Distances between E and targeted locations.

Prototype	Euclidean Distance
L1: Milton Keynes (UK	1.6125014961413195
L2: London	0.8406303029608179
L3: Brighton	0.42807759865356176

Table 5. Distances between S1 and targeted subjects.

Subject	Euclidean Distance
S5 (50,0,50,0)	70.71067811865476
S6 (65,0,0,35)	49.49747468305833
S7 (70, 30,0, 0)	35.35533905932738

a situation description and fuzzy matchmaking between real-world situations, and prototypical parameters predefined within a SWS description. For example, the following OCML code defines the partial capability description of a Web service that provides historic and cultural information for the area of Brighton:

In fact, the assumption expression presented above describes that situation description representing the current situation (*has-situation*) consider the location *Brighton* and the subject *S7*.

As a result, in our approach, the actual mobile situation description (i.e. the actual context) is the result of an iterative process that involves several distance calculations to map symbolic representations and real world characteristics. Notice that this process actively involves the end users in providing observables and validating the distance calculations. According to the obtained situation parameters and the selected user goal, the SEE discovers and orchestrates annotated Web services, which show the capabilities to suit the given situation representation. Whereas discovery and orchestration are addressed by existing SWS technology, the context-aware selection of a specific SWS goal representation is addressed through CSS by enabling similarity-based classifications of individual situations as described in the previous sections.

RELATED WORK

Since our work relates to several different but related research areas, we report here related work on (i) Semantic Web Services, (ii) Context-adaptive systems, and (iii) Context-adaptation in mobile environments. Moreover, by comparing our approach with related work in (iii) we describe our contribution to the current state of the art in context-adaptive mobile and ubiquitous computing.

SWS: OWL-S (OWL-S Coalition. 2004) is a comparatively narrow framework and ontology for adding semantics to Web service descriptions. In order to identify problematic aspects of OWL-S and suggest possible enhancements, a contextualized core ontology of services has been described in Mika et al. (2004). Such an ontology is based on DOLCE (Gangemi et al., 2002) and its specific module D&S (Gangemi, Mika, 2003). Even though we followed a similar approach, we adopt WSMO (WSMO Working Group, 2004) instead of OWL-

Listing 2. Partial OCML code representing SWS capability in terms of assumed MSS members.

```
(def-class lpmo-get-brighton-his-and-cult-LOs-ws-capability (capability) ?capability
((used-mediator:value lpmo-get-brighton-his-and-cult-LOs-mediator)
    (has-assumption:value
         (KAPPA (?web-service)
(and (= (get-location (wsmo-role-value ?web-service ‘has-situation)) “ Brighton”))
(= (get-subject (wsmo-role-value ?web-service ‘has-situation)) “S7”)))))
```

S as reference ontology for SWS. Moreover, the aim of our resulting ontology is not proposing changes to WSMO, but creating domain-specific models which incorporate WSMO-based SWS representations.

Context-adaptive systems: in Bouquet et al. (2003) the authors define contexts as the local models that encode a party's view of a domain. They distinguish contexts from ontologies, since the latter are shared models of some domain that encode a view which is common to a set of different parties. Contexts are best used in those applications where the core problem is the use and management of local and autonomous representations with a need for a lack of centralized control. For example, the notion of contexts is used in some applications of distributed knowledge management Bonifacio et al. (2003), pervasive computing environments (Chen, Finin & Joshi, 2003) and peer-to-peer applications (Serafini et al., 2003). According to the definition introduced in Bouquet et al. (2003), we propose a novel use of contexts. The local models encode party's view of SWS-based process descriptions.

Context-adaptation in mobile environments: Weissenberg et al. (2006) adopt an approach to context-adaptation in mobile settings which shows some similarities to ours: given a set of context parameters – based on sensor data – first a context is identified and then a matching situation. However, they rely on manually predefined axioms which enable such a reasoning compared to the automatic detection as proposed in this paper. Korpipaa et al. (2003) propose a related framework but firstly, require client-side applications to be installed and, secondly, relies on Bayesian reasoning for matching between measured lower-level contexts and higher-level context abstractions represented within an ontology. Hence, as a major lack, it is required to provide information about contexts and their relations within a Bayesian Network in order to perform the proposed reasoning. Gu, Wang, Pung & Zang (2004) propose a context-aware middleware which also distinguishes between lower-level and higher-level contexts. However, there is no mechanism to automatically identify relationships between certain contexts or context parameters. The same criticism applies to the approaches to a semantic representation of user contexts described in Toivinen, Kolari & Laako (2003) and Sathish, Pavel & Trossen (2006).

Finally, it can be highlighted, that current approaches to context-adaptation in mobile settings usually rely on the manual representation of mappings between a given set of real-world context data and predefined context representations. Since this approach is costly and time-consuming, our approach could contribute there significantly by providing a similarity-based and rather fuzzy method for automatically identifying appropriate symbolic context representations given a set of detected context parameters.

CONCLUSION

In this paper, we proposed an approach to support fuzzy, similarity-based matchmaking between real-world situation parameters in mobile settings and predefined semantic situation descriptions by incorporating semantic context information on a conceptual level into symbolic SWS descriptions based on Conceptual Situation Spaces. Given a particular mobile situation, defined by parameters such as the location and device of the user, the most appropriate resources, whether data or services, are discovered based on the semantic similarity, calculated in terms of the Euclidean distance, between the real-world situation and predefined resource descriptions as part of SWS representations. Even though we refer to the SWS framework WSMO in this paper, we would like to highlight, that our approach could be applied to other SWS reference ontologies such as OWL-S (OWL-S Coalition. 2004). Consequently, by aligning CSS to established SWS technologies, the expressiveness of symbolic SWS standards is extended with context information on a conceptual level described in terms of natural quality dimensions to enable fuzzy context-aware delivery of information resources at runtime. Whereas current SWS frameworks address the allocation of distributed services for a given (semantically) well-described task, Mobile Situation Spaces particularly address the similarity-based discovery of the most appropriate SWS task representation for a given situation. To prove the feasibility of our approach, a proof-of-concept prototype application was presented, which applies the MSS to enable context-adaptive resource discovery in a mobile setting.

However, although our approach applies CS to solve SWS-related issues such as the symbol grounding problem, several criticisms still have to be taken into account. Whereas defining situational contexts, respectively members within a given MSS, appears to be a straightforward process of assigning specific values to each quality dimension, the definition of the MSS itself is not trivial at all and strongly dependent on individual perspectives and subjective appraisals. Whereas the semantics of an object are grounded to metrics in geometrical vector spaces within a MSS, the quality dimensions itself are subject to ones perspective and interpretation what may lead to ambiguity issues. With regard to this, MSS do not appear to solve the symbol grounding issue but to shift it from the process of describing instances to the definition of a MSS. Moreover, distinct semantic interpretations and conceptual groundings of each dimension may be applied by different individuals. Apart from that, whereas the size and resolution of a MSS is indefinite, defining a reasonable space for a specific domain and purpose may become a challenging task. Nevertheless, distance calculation as major contribution of the MSS approach, not only makes sense for quantifiable parameters but also relies on the fact, that parameters are described in the same geometrical space.

Consequently, CS-based approaches, such as MSS, may be perceived as step forward but do not fully solve the issues related to symbolic Semantic Web (Services)-based knowledge representations. Hence, future work has to deal with the aforementioned issues. For instance, we foresee to enable adjustment of prominence values to quality dimensions of a specific space to be accomplished by a user him/herself, in order to most appropriately suit his/her specific priorities and preferences regarding the resource allocation process, since the prioritization of dimensions is a highly individual and subjective process. Nevertheless, further research will be concerned with the application of our approach to further domain-specific situation settings.

REFERENCES

Bonifacio, M., Bouquet, P., Mameli, G., & Nori, M. (2003). Peer-mediated knowledge management. In *AAAI-03 Spring Symposium on Agent-Mediated Knowledge Management (AMKM-03).*

Bouquet, P., Giunchiglia, F., van Harmelen, F., Serafini, L., & Stuckenschmidt, H. (2003). C-OWL: Contextualizing Ontologies. *ISWC-2003* [Springer Verlag.]. *LNCS, 2870*, 164–179.

Cabral, L., Domingue, J., Galizia, S., Gugliotta, A., Norton, B., Tanasescu, V., & Pedrinaci, C. (2006). IRS-III: A Broker for Semantic Web Services based Applications. *Proceedings of the 5th International Semantic Web Conference (ISWC)*, Athens, USA.

Chen, H., Finin, T., & Joshi, A. (2003). An Ontology for Context Aware Pervasive Computing Environments. *The Knowledge Engineering Review, 18*, 197–207. doi:10.1017/S0269888904000025

Coalition, O. W. L.-S. OWL-S 1.1 release. (2004). http://www.daml.org/services/owl-s/1.1/

Cregan, A. (2007). Symbol Grounding for the Semantic Web. *4th European Semantic Web Conference 2007*, Innsbruck, Austria.

Devore, J., & Peck, R. (2001). *Statistics—The Exploration and Analysis of Data*. 4th ed. Pacific Grove, CA: Duxbury.

Dietze, S., Gugliotta, A., & Domingue, J. (2007). A Semantic Web Services-based Infrastructure for Context-Adaptive Process Support. *Proceedings of IEEE 2007 International Conference on Web Services (ICWS)*, Salt Lake City, Utah, USA.

Dietze, S., Gugliotta, A., & Domingue, J. (2008). Towards Context-aware Semantic Web Service Discovery through Conceptual Situation Spaces. *Workshop: International Workshop on Context enabled Source and Service Selection, Integration and Adaptation (CSSSIA), 17th International World Wide Web Conference (WWW2008)*, Beijing, China.

Fensel, D., Lausen, H., Polleres, A., de Bruijn, J., Stollberg, M., Roman, D., & Domingue, J. (2006). *Enabling Semantic Web Services—The Web service Modelling Ontology*. Springer.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., & Schneider, L. (2002). Sweetening Ontologies with DOLCE. In A. Gómez-Pérez, V. Richard Benjamins (Eds.), *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web: 13th International Conference*, EKAW 2002, Siguenza, Spain, October 1-4.

Gangemi, A., & Mika, P. (2003). Understanding the Semantic Web through Descriptions and Situations. In R. Meersman, Z. Tari, & et al. (Eds.), *Proceedings of the On The Move Federated Conferences (OTM'03)*, LNCS. Springer Verlag.

Gärdenfors, P. (2000). *Conceptual Spaces—The Geometry of Thought*. MIT Press.

Gärdenfors, P. (2004). How to make the semantic web more semantic. In A. C. Vieu & L. Varzi, (Eds.), *Formal Ontology in Information Systems*, (pp. 19–36). IOS Press.

Gellersen, H-W., Schmidt, A., & Beigl, M. (2002). Multi-Sensor Context-Awareness in Mobile Devices and Smart Artefacts. *ACM Journal Mobile Networks and Applications (MONET)*, *7*(5).

Gu, T., Wang, X. H., Pung, H. K., & Zhang, D. Q. (2004). A middleware for context-aware mobile services. In IEEE Vehicular Technology Conference (VT)C. Los Alamitos, CA: IEEE Computer Society Press.

Joint US/EU ad hoc Agent Markup Language Committee (2004). OWL-S 1.1 Release. http://www.daml.org/services/owl-s/1.1/.

Korpipaa, P., Mantyjarvi, J., Kela, J., Keranen, H., & Malm, E. (2003, Jul-Sept). Managing Context Information in Mobile Devices. *IEEE Pervasive Computing / IEEE Computer Society [and] IEEE Communications Society*, *2*(3), 42–51. doi:10.1109/MPRV.2003.1228526

Krause, E. F. (1987). Taxicab Geometry. Dover.

Mika, P., Oberle, D., Gangemi, A., & Sabou, M. (2004). *Foundations for Service Ontologies: Aligning OWL-S to DOLCE, WWW04*.

Motta, E. (1998). An Overview of the OCML Modelling Language. *The 8th Workshop on Methods and Languages*.

Nosofsky, R. (1992). Similarity, scaling and cognitive process models. *Annual Review of Psychology*, *43*, 25–53. doi:10.1146/annurev.ps.43.020192.000325

Raubal, M. (2004). Formalizing Conceptual Spaces. In A. Varzi & L. Vieu (Eds.), *Formal Ontology in Information Systems*, *Proceedings of the Third International Conference (FOIS 2004).Frontiers in Artificial Intelligence and Applications*, *114*, 153-164., Amsterdam, The Netherlands: IOS Press.

Sailesh, S., Pavel, D., & Trossen, D. (2006). Context Service Framework for Mobile Internet. *International Worskshop on System Support for Future* Mobile Computing Applications (FUMCA 2006), September 2006, Irvine, California, USA.

Schmidt, A., & Winterhalter, C. (2004). User Context Aware Delivery of E-Learning Material: Approach and Architecture. *Journal of Universal Computer Science (JUCS)*, *10*(1), January 2004.

Serafini, L., Giunchiglia, F., Mylopoulos, J., & ernstein, P. (2003). Local relational model: a logical formalization of database coordination. In P. Blackburn, C. Ghidini, & R. Turner (Eds.), *Context'03*.

Toivonen, S., Kolari, J., & Laakko, T. (2003). Facilitating mobile users with contextualized content. *In Proc. Workshop Artificial Intelligence in Mobile Systems.*

Weißenberg, N., Gartmann, R., & Voisard, A. (2006). An Ontology-Based Approach to Personalized Situation-Aware Mobile Service Supply. *Geoinformatica 10*, 1 (Mar. 2006), 55-90. DOI= http://dx.doi.org/10.1007/s10707-005-4886-9.

World Wide Web Consortium. W3C (2004a): Resource Description Framework, W3C Recommendation 10 February 2004, http://www.w3.org/RDF/.

World Wide Web Consortium. W3C (2004b): Web Ontology Language Reference, W3C Recommendation 10 February 2004, http://www.w3.org/TR/owl-ref/.

WSMO Working Group. (2004), D2v1.0: Web service Modeling Ontology (WSMO). WSMO Working Draft, (2004). (http://www.wsmo.org/2004/d2/v1.0/).

WSMX Working Group. (2007), The Web Service Modelling eXecution environment, http://www.wsmx.org/.

ENDNOTES

1 DIP Project: http://dip.semanticweb.org

2 http://maps.google.com/

This work was previously published in Context-Aware Mobile and Ubiquitous Computing for Enhanced Usability: Adaptive Technologies and Applications, edited by D. Stojanovic, pp. 217-234, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 7.3
Uncertainty Representation and Reasoning in the Semantic Web

Paulo Cesar G. Costa
George Mason University, USA

Kathryn Blackmond Laskey
George Mason University, USA

Thomas Lukasiewicz
Oxford University Computing Laboratory, UK

ABSTRACT

This chapter is about uncertainty representation and reasoning for the Semantic Web (SW). We address the importance, key issues, state-of-the-art approaches, and current efforts of both the academic and business communities in their search for a practical, standard way of representing and reasoning with incomplete information in the Semantic Web. The focus is on why uncertainty representation and reasoning are necessary, its importance to the SW vision, and the major issues and obstacles to addressing uncertainty in a principled and standardized way. Although some would argue that uncertainty belongs in the "rule layer" of the SW, we concentrate especially on uncertain extensions of ontology languages for the Semantic Web.

DOI: 10.4018/978-1-60566-112-4.ch013

WHY CARE ABOUT UNCERTAINTY?

After some years of SW research, the subject remains rife with controversy, and there is still some disagreement on how uncertainty should be handled in SW applications. Thus, it is no surprise that little was said on the subject in previous chapters of this book. A major reason for the present state of affairs is that the most popular technologies employed in SW applications are rooted in traditional knowledge representation formalisms that have historically ignored uncertainty. The most compelling examples are Frame Systems (Minsky, 1975), and Description Logics, which evolved from the so-called "Structured Inheritance Networks" (Brachman, 1977), and form the logical basis for the ontology language OWL.

The spotlight is not on the *status quo*, but on what the future holds. To answer this question, we begin with a comprehensive analysis of the major challenges to be faced by the SW community, including what kinds of interactions, scenarios, demands, and obstacles must be addressed to make the SW promises a reality. Next, we assess whether protocols that rely only on complete, deterministic information will suffice to address these challenges. Although much progress has been made by tackling problems in which uncertainty is inessential or can be circumvented, addressing the full range of challenges inherent in the Semantic Web vision will require optimal use of *all* available information. In this Chapter, we argue that a principled framework for representing and reasoning with incomplete information is necessary to realizing the SW vision. Because uncertainty is a ubiquitous aspect of most real-world problems, any representation scheme intended to model real-world entities, properties and processes must be able to cope with uncertain phenomena. Current SW technologies' inability to represent and reason about uncertainty in a sound and principled manner raises an unnecessary barrier to the development of new, powerful features for general knowledge application, a limitation that threatens to derail the original vision for the Semantic Web as a whole. In other words, we argue that realizing the SW as envisioned by Tim Berners-Lee (Berners-Lee & Fischetti, 2000) requires a principled framework for representing and reasoning with uncertainty.

The Semantic Web envisions effortless cooperation between humans and computers, seamless interoperability and information exchange among web applications, and rapid and accurate identification and invocation of appropriate Web services. While considerable progress has been achieved toward realization of the Semantic Web vision, it is increasingly apparent that a sound and principled technology for handling uncertainty is an important requirement for continued progress. Uncertainty is an unavoidable factor in knowledge interchange and application interoperability. Different applications have different ontologies, different semantics, and different knowledge and data stores. Legacy applications are usually only partially documented and may rely on tacit usage conventions that even proficient users do not fully understand or appreciate. Furthermore, data that is exchanged in the context of the semantic web is often incomplete, inconsistent, and inaccurate. This suggests that recent work in the application of probability, fuzzy logic, and decision theory to complex, open-world problems could be of vital importance to the success of the Semantic Web. Incorporating these new technologies into languages, protocols, and specifications for the Semantic Web is fundamental to realizing the Semantic Web vision.

Typical Problems Needing Uncertainty Representation and Reasoning. The following web-relevant reasoning challenges illustrate the kinds of problems for which reasoning under uncertainty is important.

- Information extracted from large information networks such as the World Wide Web is typically incomplete. The ability to exploit partial information is useful for identifying sources of service or information. For example, the fact that an online service deals with greeting cards may be evidence that it also sells stationery. It is clear that search tools capable of utilizing probabilistic knowledge could increase search effectiveness.
- Much information on the World Wide Web is likely to be uncertain. Common examples include weather forecasts and gambling odds. A canonical method for representing and integrating such information and the uncertainty associated with it is necessary for communicating such information in a seamless fashion.
- Web information is also often incorrect or only partially correct, raising issues related to trust or credibility. Uncertainty representation and reasoning helps to resolve

tensions amongst information sources for purposes of approximating appropriately.

- The Semantic Web will require numerous distinct but conceptually overlapping ontologies to co-exist and interoperate. Ontology mapping will benefit from the ability to represent and reason with information about partial overlap, such as likelihoods of membership in Class A of Ontology 1 given membership in Class B of Ontology 2.

Section 5 below discusses some use cases, based on the work of the W3C Uncertainty Reasoning for the World Wide Web Incubator Group (URW3-XG). These use cases exhibit the above characteristics, and are representative of the kinds of challenges that the SW must address. Despite the potential that a principled framework for representing uncertainty would have in contributing to the development of robust SW solutions, for historical reasons, research on the Semantic Web started with little support for representing and reasoning in the presence of uncertain, incomplete knowledge. As interest in and application of SW technology grows, there is increasing recognition of the need for uncertain reasoning technology, and increasing discussion of the most appropriate ways to address this need.

Should Ontologies Represent Uncertainty? A major impediment to widespread adoption of technologies for representing and reasoning with incomplete information is the dominance of the classical logic paradigm in the field of ontological engineering. There is a plethora of definitions of the term ontology in the field of information systems. Among these, a common underlying assumption is that classical logic would provide the formal foundation for knowledge representation and reasoning. Until recently, theory and methods for representing and reasoning with uncertain and incomplete knowledge have been neglected almost entirely. However, as research on knowledge engineering and applications of ontologies matures, the ubiquity and importance of uncertainty across a wide array of application areas has generated consumer demand for ontology formalisms that can capture uncertainty. Although recognition of the need for uncertainty reasoning is growing, there is disagreement about its appropriate place in the Semantic Web architecture. We have argued elsewhere (e.g., Costa, 2005; Costa and Laskey, 2006), that there is a need to represent declarative knowledge about likelihood in domain ontologies. In environments in which noisy and incomplete information is the rule, likelihood information is a key aspect of domain knowledge. Furthermore, much of the key semantic content needed to enable interoperability involves information about plausibility. For this reason, we have argued, knowledge about likelihoods should be included in formal domain ontologies.

This viewpoint is not universal. A counterargument to our position is that probability is inherently epistemic, whereas formal ontology should represent phenomena as they exist in the world. Carried to its extreme, however, this philosophical stance would preclude the use of virtually every ontology that has yet been developed. To explore this idea further, we note that if computational ontologies had existed in the 17th century, Becher and his followers might well have developed an ontology of phlogiston. We may chuckle now at their naïveté, but who among our 17th century predecessors had the foresight to judge which of the many scientific theories then in circulation would stand the test of time? Researchers in medicine, biology, defense, astronomy, and other communities have developed a plethora of domain ontologies. It is virtually certain that at least some aspects of some of these ontologies will, as human knowledge progresses, turn out in retrospect to be as well founded as the theory of phlogiston. Shall we outlaw use of all these ontologies until the day we can prove they contain only that which is ontological, and nothing that is mere epistemology? Moreover, many aspects of our common, shared knowledge of these domains are inherently probabilistic. Well-established statistical regularities are a key element of expert reasoning. A principled means of representing these probabilistic aspects of domain knowledge is needed to facilitate interoperability and knowledge sharing.

Similar questions arise with the representation of vagueness. Fuzzy logic has been applied extensively to problems of reasoning with imprecisely defined terms. For example, fuzzy reasoning might be applied to retrieve and sort responses to a query for "inexpensive" patio furniture. A fuzzy reasoner would assign each furniture set a degree of membership in the fuzzy set "inexpensive," and would sort the retrieved sets by their membership in the fuzzy set. There is an analogous question of whether it is legitimate to extend ontology formalisms to allow representation of fuzzy membership values, or whether fuzziness is inherently epistemological and does not belong in an ontology.

There is a valid, important, and as yet unresolved philosophical clash between those who believe that we live in a deterministic world in which uncertainty is entirely epistemic, and those who believe the world includes phenomena that are ontologically stochastic and/or imprecise and should be represented as such. From an engineering standpoint, we cannot wait for the debate to be resolved before we move forward with building and using ontologies.

Although our ultimate scientific objective is to seek the truth about reality as it is, this ultimate objective is unattainable in the lifetime of any human. Therefore, no "perfect ontology of all things" is reachable, regardless of one's philosophical view on uncertainty. Nevertheless, from a pragmatic perspective, it is necessary and desirable to do the best we can with the knowledge we have, even if this causes the ontology to be under-specified due to incomplete information. Formal ontology provides a useful means of communicating domain knowledge in a precise and shareable manner, and of extending and revising our descriptions as human knowledge accrues. Accepting only complete knowledge would leave us with too little information to solve most of the interesting problems that ontologies are capable of addressing.

Not surprisingly, as ontology engineering research has achieved a greater level of maturity, the need for uncertainty representation and reasoning for the Semantic Web has become more and more clear. Correspondingly, interest is increasing in extending traditional ontology formalisms to include standard mechanisms for representing and reasoning with uncertainty. Whether the ultimate consensus is that ontology formalisms should be capable of representing information about uncertainty, or that ontologies should represent the space of possibilities and that information about uncertainty should be conveyed in a different semantic layer, principled means of representing and reasoning with uncertainty are increasingly seen as necessary.

***Uncertainty in Rule Languages*.** A related stream of research has focused on augmenting SW rule languages to handle uncertainty (Damásio *et al.*, 2006; Lukasiewicz, 2005, 2006; Lukasiewicz & Straccia, 2007). Although there is as yet no standard rule language for the Semantic Web, the W3C's Rule Interchange Format (RIF) Working Group has recently released working draft documents specifying use cases, requirements, and a core design for a format that allows rules to be translated between rules languages[a]. The use cases and requirements document does not mention uncertainty, but the core design mentions the need to translate between rule languages that handle uncertainty, and makes brief mention of syntactic and semantic implications of the need to treat uncertainty. This brief treatment is far from sufficient to address the full range of issues that need to be addressed to achieve semantic interoperability between systems that express and reason with uncertainty. For space reasons, we do not address rule language research in detail in this chapter. We note, however, that augmenting ontologies to express uncertainty generates a requirement to augment rule languages to take advantage of the information expressed in uncertainty-enhanced ontologies.

Towards a Pragmatic View. Apart from the historical and philosophical issues, as research on SW leaves the conceptual level and reaches a level of maturity in which the actual challenges are better understood, realization has grown that many SW applications require a principled means for representing uncertainty and performing plausible reasoning with incomplete data. As the interest in uncertainty representation techniques grows, the focus of SW shifts from philosophical issues toward "down to earth" engineering issues. Important challenges are to identify the kinds of information management problems that would benefit most from mechanisms for reasoning with uncertainty, to assess the scalability of uncertainty representation approaches, to evaluate the suitability of different forms of representation and reasoning to solve specific use cases, and others.

This pragmatic, focused view has pushed researchers from many different domains of knowledge into an appreciation of the need for a forum to discuss the ways in which uncertainty reasoning can contribute to addressing their respective challenges, and to evaluate the strengths and weaknesses of different approaches to representing and reasoning under uncertainty. Although uncertainty-related papers were sometimes presented in other venues, the first forum explicitly geared towards answering the above issues was the workshop on Uncertainty Representation for the Semantic Web (URSW workshop), held in conjunction with the Fourth International Semantic Web Conference (ISWC 2005). The intention of the URSW workshop was to provide an open forum to all forms of uncertainty representation and reasoning, without being prejudicial in favor of any particular approach. At the second workshop (URSW 2006), a consensus was reached that the most important tasks were (1) to develop a set of use cases for uncertainty in the SW; and (2) to assess how each approach (or combination of approaches) would address appropriate challenges set out in the use cases. In the end, a much improved understanding of those issues would led to identification of best practices involving uncertainty reasoning in the SW.

The strong interest in the URSW and similar venues prompted the W3C to create, in March 2007, the Uncertainty Reasoning for the World Wide Web Incubator Group (URW3 XG), with the objective of better defining the challenge of working with incomplete knowledge. The URW3 adopted the same "approach-independent" stance as the URSW, with an initial focus on the problem itself rather than a particular approach to solving it. At the time of this writing, the URW3 is actively pursuing its development of use cases, and planning for a third URSW is underway. The next two sections present a brief view of the major approaches for uncertainty in the SW being discussed in fora such as the URW3 and URSW.

PROBABILISTIC APPROACHES TO UNCERTAINTY IN THE SEMANTIC WEB

Bayesian probability provides a mathematically sound representation language and formal calculus for rational degrees of belief, which gives different agents the freedom to have different beliefs about a given hypothesis. This provides a compelling framework for representing uncertain, incomplete knowledge that can come from diverse agents. Not surprisingly, there are many distinct approaches using Bayesian probability for the Semantic Web.

Bayesian knowledge representation and reasoning systems have their formal basis in the axioms of probability theory (e.g., Ramsey, 1931; Kolmogorov, 1960/1933). Probability theory allows propositions to be assigned truth-values in the range from zero, meaning certain falsehood, to one, meaning certain truth. Values intermediate

$$P(B|A) = \frac{P(A|B)P(B)}{P(B)}$$

between zero and one reflect degrees of likelihood of a proposition that may be either true or false. *Bayes Rule*, a theorem that can be derived from the axioms of probability theory, provides a method of updating the probability of a proposition when information is acquired about a related proposition. The standard format of Bayes rule is:

On the right side of the formula, $P(A)$ is called the prior probability of A, and represents our belief in event A before obtaining information on event B. Likewise, $P(B)$ is called the prior probability of B. There is also $P(A|B)$, which is the likelihood of event A given that event B has happened. On the left side of the formula there is $P(B|A)$, which is the posterior probability of B, and represents our new belief in event B after applying Bayes rule with the information collected from event A. Bayes rule provides the formal basis for the active and rapidly evolving field of Bayesian probability and statistics. In the Bayesian view, inference is a problem of belief dynamics. Bayes rule provides a principled methodology for belief change in the light of new information.

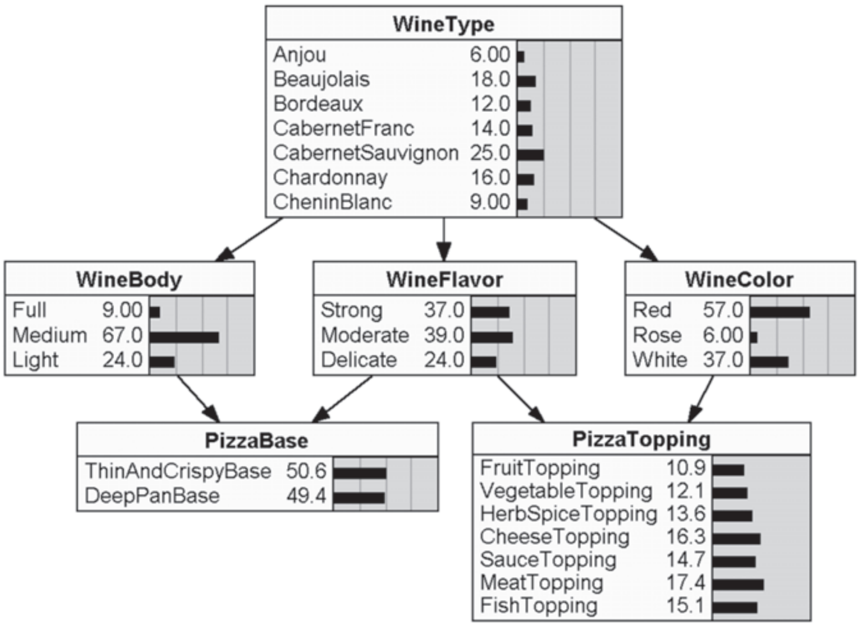
Bayesian Networks (BNs). BNs provide a means of parsimoniously expressing joint probability distributions over many interrelated hypotheses. A Bayesian network consists of a directed acyclic graph (DAG) and a set of local distributions. Each node in the graph represents a random variable. A random variable denotes an attribute, feature, or set of hypotheses about which we may be uncertain. Each random variable has a set of mutually exclusive and collectively exhaustive possible values. That is, exactly one of the possible values is or will be the actual value, and we are uncertain about which one it is. The graph represents direct qualitative dependence relationships; the local distributions represent quantitative information about the strength of those dependencies. The graph and the local distributions together represent a joint probability distribution over the random variables denoted by the nodes of the graph.

Bayesian networks have been successfully applied to create consistent probabilistic representations of uncertain knowledge in diverse fields. Heckerman *et al.* (1995) provide a detailed list of recent applications of Bayesian Networks. The prospective reader will also find comprehensive coverage of Bayesian Networks in a large and growing literature on this subject, such as Pearl (1988), Neapolitan (1990, 2003), and others. Figure 1 shows an example of a BN representing part of a highly simplified ontology for wines and pizzas.

In this toy example[b], we assume that domain knowledge about gastronomy was gathered from sources such as statistical data collected among restaurants and expertise from sommeliers and pizzaiolos. Moreover, the resulting ontology also considered incomplete knowledge to establish a probability distribution among features of the pizzas ordered by customers (i.e. type of base and topping) and characteristics of the wines ordered to accompany the pizzas.

Consider a customer who enters a restaurant and requests a pizza with cheese topping and a thin and crispy base. Using the probability distribution stored in the BN of Figure 1, the waiter can apply Bayes rule to infer the best type of wine to offer the customer given his pizza preferences the body of statistical and expert information previously linking features of pizza to wines. Such computation would be difficult when there are very many features. Bayesian networks provide a parsimonious way to express the joint distribution and a computationally efficient way to implement Bayes rule. This inferential process is shown in Figure 2, where evidence (i.e., the customer's order) was entered in the BN and its result points to Beaujolais as the most likely wine the customer would order, followed by Cabernet Sauvignon, and so on.

Figure 1. A BN for pizzas and wines

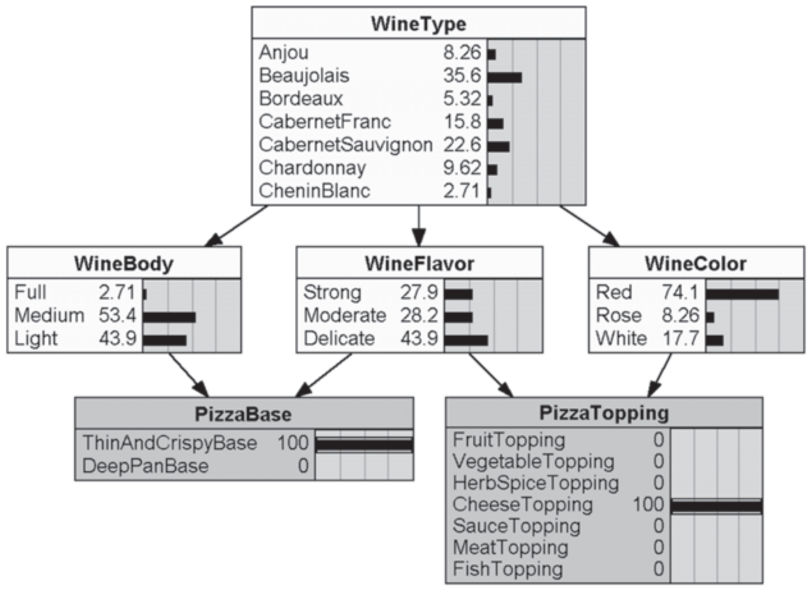


Although this is just a toy example, it is useful to show how incomplete information about a domain can be used to improve decisions. In an ontology without uncertainty, there would not be enough information for a logical reasoner to infer a good choice of wine to offer the customer, and the decision would have to be made without optimal use of all the information available.

As Bayesian networks have grown in popularity, their shortcomings in expressiveness for many real-world applications have become increasingly apparent. More specifically, Bayesian Networks assume a simple attribute-value representation – that is, each problem instance involves reasoning about the same fixed number of attributes, with only the evidence values changing from problem instance to problem instance. In the pizza and wine example, the PizzaTopping random variable conveys general information about the class of pizza toppings (i.e., types of toppings for a given pizza and how it is related to preferences over wine flavor and color), but the BN in Figures 1 and 2 is valid for pizzas with only one topping. To deal with more elaborate pizzas, it is necessary to build specific BNs for each configuration, each one with a distinct probability distribution. Figure 3 depicts a BN for a 3- topping pizza with a specific customer preference displayed. Also, the information conveyed by the BNs (i.e., for 1-topping, 2-toppings, etc.) relates to the class of pizza toppings, and not to specific instances of those classes. Therefore, the BN in Figure 3 cannot be used for a situation in which the costumer asks for two 3-topping pizzas. This type of representation is inadequate for many problems of practical importance. Similarly, these BNs cannot be used to reason about a situation in which a customer orders several bottles of wine that may be of different varieties. Many domains require reasoning about varying numbers of related entities of different types, where the numbers, types, and relationships among entities usually cannot be specified in advance and may have uncertainty in their own definitions.

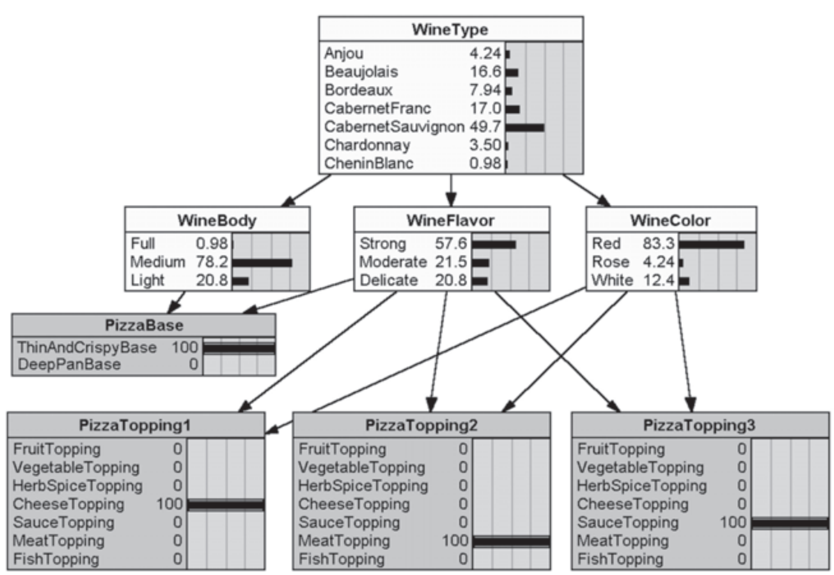
In spite of their limitations, BNs have been used in specific applications for the SW where

Figure 2. BN after entering evidence



the limitations on expressivity can be overcome by clever knowledge engineering workarounds. One example is BayesOWL (Ding and Peng, 2004; Ding, 2005), which augments OWL semantics to allow probabilistic information to be represented via additional markups. The result is a probabilistic annotated ontology that could then be translated to a Bayesian network. Such a translation is based on a set of translation rules that rely on the probabilistic information attached to individual concepts and properties within the annotated ontology. After successfully achieving

Figure 3. A BN for the 3-topping pizza configuration with evidence



the translation, the resulting Bayesian network will be associated with a joint probability distribution over the application domain. Although a full translation of an ontology to a standard BN is impossible given the limitations of the latter in terms of expressivity, the scheme can be successfully used to tackle specific problems involving uncertainty.

Also focusing on Bayesian extensions geared towards the Semantic Web is the work of Gu *et al*. (2004), which takes an approach similar to that of BayesOWL. A related effort is the set of RDF extensions being developed by Yoshio Fukushige (2004). Generally speaking, SW approaches that rely on BNs will have to compensate for their lack of expressiveness by specializing in a specific type of problem, such as the BN-focused approaches for solving the ontology mapping problem (e.g., Mitra *et al.*, 2004; and Pan *et al.*, 2005; Peng *et al.*, 2007).

Probabilistic Extensions to Description Logics. Most of the probabilistic extensions aimed at the ontology engineering domain are based on description logics (DLs), which Baader and Nutt (2003, page 47) define as a family of knowledge representation formalisms that represent the knowledge of an application domain (the "world") by first defining the relevant concepts and roles of the domain (its terminology), which represent classes of objects/individuals and binary relations between such classes, respectively, and then using these concepts and roles to specify properties of objects/individuals occurring in the domain (the world description).

Description logics divide a knowledge base into two components: a terminological box, or T-Box, and the assertional box, or A-Box. The first introduces the terminology (i.e., the vocabulary) of an application domain, while the latter contains assertions about instances of the concepts defined in the T-Box. Description logics are a subset of first-order logic (FOL) that provide a very good combination of decidability and expressiveness. In fact, an important desired property of description logics is the decidability of their reasoning tasks. Description logics are also the basis of the web ontology language OWL, whose sublanguages OWL Lite and OWL DL correspond to the expressive description logics *SHIF*(**D**) and *SHOIN*(**D**), respectively.

There are several probabilistic extensions of description logics in the literature, which can be classified according to the generalized description logics, the supported forms of probabilistic knowledge, and the underlying probabilistic reasoning formalism.

Heinsohn (1994) presents a probabilistic extension of the description logic *ALC* (a member of the *AL*-languages (Schmidt-Schauß & Smolka, 1991) obtained by including the full existential quantification and the union constructors to the basic *AL* (attributive language)), which allows representation of terminological probabilistic knowledge about concepts and roles, and which is essentially based on probabilistic reasoning in probabilistic logics. Heinsohn, however, does not allow for assertional knowledge about concept and role instances. Jaeger (1994) proposes another probabilistic extension of the description logic *ALC*, which allows for terminological and assertional probabilistic knowledge about concepts and roles and about concept instances, respectively, but does not support assertional probabilistic knowledge about role instances (but he mentions a possible extension in this direction). The uncertain reasoning formalism in (Jaeger, 1994) is essentially based on probabilistic reasoning in probabilistic logics, as the one in (Heinsohn, 1994), but coupled with cross-entropy minimization to combine terminological probabilistic knowledge with assertional probabilistic knowledge. Jaeger's recent work (2006) focuses on interpreting probabilistic concept subsumption and probabilistic role quantification through statistical sampling distributions, and develops a probabilistic version of the guarded fragment of first-order logic.

The work by Koller *et al.* (1997) gives a probabilistic generalization of the CLASSIC description logic, called P-CLASSIC. In short, each probabilistic component is associated with a set P of p-classes, and each p-class C in set P is represented using a Bayesian network. Like Heinsohn's work (1994), the work by Koller *et al.* (1997) allows for terminological probabilistic knowledge about concepts and roles, but does not support assertional probabilistic knowledge about instances of concepts and roles. However, differently from (Heinsohn, 1994), it is based on inference in Bayesian networks as underlying probabilistic reasoning formalism. Closely related work by Yelland (2000) combines a restricted description logic close to *FL* with Bayesian networks. It also allows for terminological probabilistic knowledge about concepts and roles, but does not support assertional knowledge about instances of concepts and roles.

Another DL with a probabilistic extension is *SHOQ*(**D**) (Horrocks & Sattler, 2001). *SHOQ*(**D**) is the basis of DAML+OIL (Horrocks, 2002), the language that came from merging two ontology languages being developed in the US (DAML) and Europe (OIL) and has been superseded by OWL. Its probabilistic extension is called P-*SHOQ*(**D**) (Giugno & Lukasiewicz, 2002) (see also (Lukasiewicz, 2008)) and allows for expressing both terminological probabilistic knowledge about concepts and roles, as well as assertional probabilistic knowledge about instances of concepts and roles. P-*SHOQ*(**D**) is based on probabilistic lexicographic entailment from probabilistic default reasoning (Lukasiewicz, 2002) as underlying probabilistic reasoning formalism, which treats terminological and assertional probabilistic knowledge in a semantically very appealing way as probabilistic knowledge about random and concrete instances, respectively.

Description logics are highly effective and efficient for the classification and subsump-tion problems that they were designed to address. However, their ability to represent and reason about other commonly occurring kinds of knowledge is limited. One restrictive aspect of DL languages is their limited ability to represent constraints on the instances that can participate in a relationship. As an example, a probabilistic Description Logics version of the toy example in Figures 1 to 3 would allow us to instantiate (say) three pizzas. However, suppose we want to express that for a given pizza to be compatible with another pizza in a specific type of situation (e.g., a given mixture of toppings for distinct pizzas), it is mandatory that the two individuals of class pizza involved in the situation are not the same. In DLs, making sure that the two instances of class pizza are different in a specific situation is only possible if we actually instantiate/specify the tangible individuals involved in that situation. Indeed, stating that two "fillers" (i.e., the actual individuals of class Pizza that will "fill the spaces" of concept pizza in our statement) are not equal without specifying their respective values would require constructs such as *negation* and *equality role-value-maps*, which cannot be expressed in description logics. While equality and role-value-maps provide additional useful means to specify structural properties of concepts, their inclusion makes the logic undecidable (Calvanese & De Giacomo, page 223).

First-Order Probabilistic Approaches. In recent years, a number of languages have appeared that extend the expressiveness of probabilistic graphical models in various ways. This trend reflects the need for probabilistic tools with more representational power to meet the demands of real world problems, and goes to the encounter of the needs for Semantic Web representational schemes compatible with incomplete, uncertain knowledge. A clear candidate logic to fulfill this requirement for extended expressivity is first-order logic (FOL), which according to Sowa (2000, page 41) "has enough expressive power to define all of mathematics, every digital computer that has ever been built, and the semantics of every version of logic, including itself."

FOL was invented independently by Frege and Pierce in the late nineteenth century (Frege, 1879/1967; Pierce, 1898) and is by far the most commonly used, studied, and implemented logical system. A theory in first-order logic assigns definite truth-values only to sentences that have the same truth-value (either true or false) in all interpretations of the theory. The most that can be said about any other sentence is that its truth-value is indeterminate. A logical system is *complete* if all valid sentences can be proven and *negation complete* if for every sentence, either the sentence or its negation can be proven. Kurt Gödel proved both that first-order logic is complete, and that no consistent logical system strong enough to axiomatize arithmetic can be negation complete (cf. Stoll, 1963; Enderton, 2001). However, systems based on classical first-order logic lack a theoretically principled, widely accepted, logically coherent methodology for reasoning under uncertainty. Below are some of the approaches addressing this issue.

Object-Oriented Bayesian Networks (Koller & Pfeffer, 1997; Bangsø & Wuillemin, 2000; Langseth & Nielsen, 2003) represent entities as instances of object classes with class-specific attributes and probability distributions. Probabilistic Relational Models (PRM) (Pfeffer *et al.*, 1999; Getoor *et al.*, 2000; Getoor *et al.*, 2001; Pfeffer, 2001) integrate the relational data model (Codd, 1970) and Bayesian networks. PRMs extend standard Bayesian Networks to handle multiple entity types and relationships among them, providing a consistent representation for probabilities over a relational database. PRMs cannot express arbitrary quantified first-order sentences and do not support recursion. Although PRMs augmented with DBNs can support limited forms of recursion, they still do not support general recursive definitions. Jaeger (1997) extends relational probabilistic models to allow recursion, but it is limited to finitely many random variables. Plates (Buntine, 1994; Gilks *et al.*, 1994; Spiegelhalter *et al.*, 1996) represent parameterized statistical models as complex Bayesian networks with repeated components.

DAPER (Heckerman *et al.*, 2004) combines the entity-relational model with DAG models to express probabilistic knowledge about structured entities and their relationships. Any model constructed in Plates or PRM can be represented by DAPER. Thus, DAPER is a unifying language for expressing relational probabilistic knowledge. DAPER expresses probabilistic models over finite databases, and cannot represent arbitrary first-order sentences involving quantifiers. Therefore, like other languages discussed above, DAPER does not achieve full first-order representational power.

MEBN (Laskey and Mahoney, 1997; Laskey and Costa, 2005; Laskey, 2007) represents the world as consisting of entities that have attributes and are related to other entities. Knowledge about the attributes of entities and their relationships to each other is represented as a collection of MEBN fragments (MFrags) organized into MEBN Theories (MTheories). An MFrag represents a conditional probability distribution for instances of its resident random variables given their parents in the fragment graph and the context nodes. An MTheory is a set of MFrags that collectively satisfies consistency constraints ensuring the existence of a unique joint probability distribution over instances of the random variables represented in each of the MFrags within the set. MEBN semantics integrates the standard model-theoretic semantics of classical first-order logic with random variables as formalized in mathematical statistics.

Although the above approaches are promising where applicable, a workable solution for the Semantic Web requires a general-purpose formalism that gives ontology designers a range of options to balance tractability against expressiveness. Current research on SW formalisms using first-order probabilistic logics is still in its infancy, and generally lack a complete set of publicly available tools. Examples include PR-OWL (Costa, 2005), which is an upper ontology for building probabilistic ontologies based on MEBN logic[c], and KEEPER (Pool and Aiken, 2005), an OWL-based interface for the relational probabilistic toolset Quiddity*Suite, developed by IET, Inc.

Their constructs are similar in spirit and provide an expressive method for representing uncertainty in OWL ontologies. Costa (2005) gives a definition for Probabilistic Ontologies, develops rules for constructing PR-OWL ontologies in a manner that can be translated into Quiddity*Suite, and describes how to perform the translation.

As an illustration of the expressiveness of a first-order probabilistic logic, Figure 4 presents a graphical depiction of the MFrags for the wine and pizza toy example.[d] It conveys both the structural relationships (implied by the arcs) among the nodes and the numerical probabilities (embedded in the probability distributions and not depicted in the figure). The MFrags depicted in Figure 4 form a consistent set that allows to reason probabilistically about a domain and can be stored in an OWL file using the classes and properties defined in the PR-OWL upper ontology. The MFrags can be used to instantiate situation specific Bayesian networks to answer queries about the domain of application being modeled. In other words, a PR-OWL probabilistic ontology consists of both deterministic and probabilistic information about the domain of discussion (e.g., wines and pizzas), stored in an OWL file that can be used for answering specific queries for any configuration of the instances given the evidence at hand.

In particular, the toy ontology of Figure 4 can be applied to reason about situations involving any number of pizzas with any number of toppings on each, accompanied by any number of bottles of wine, and including any possible interactions among specific instances of those. Figure 5 illustrates this concept, depicting a situation in which evidence a customer has ordered one thin and crispy pizza with three toppings (cheese, meat, and sauce) and is planning to order one bottle of wine. The BN represents the response to a request to suggest a good wine to go with the pizzas.

In MEBN syntax[e], the knowledge base is augmented by an instance of pizza (!P0), three instances of topping types (!T0, !T1, !T2), and an instance of wine (!W0). To answer the query on the wine suggestion, a probabilistic reasoner will use the evidence available to build a Situation Specific Bayesian Network (SSBN). This example was constructed to yield the same BN as Figure 3. This illustrates the point that the MFrags in Figure 4 have captured all information that is needed to

Figure 4. MFrags representing the wine and pizza example

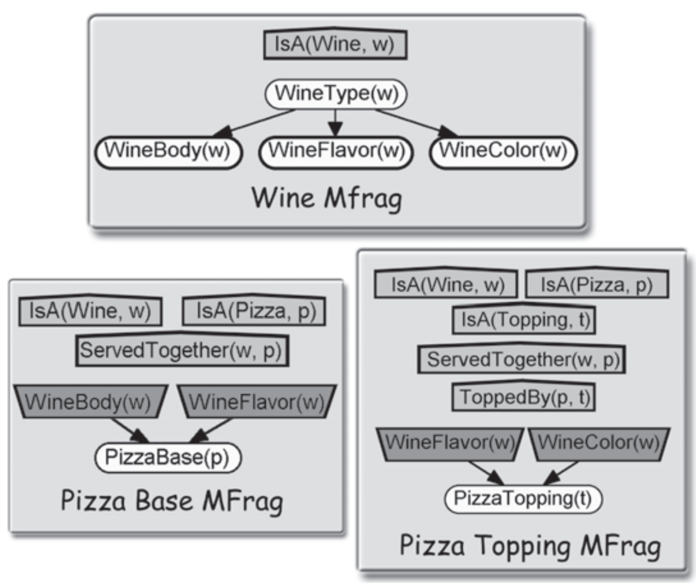
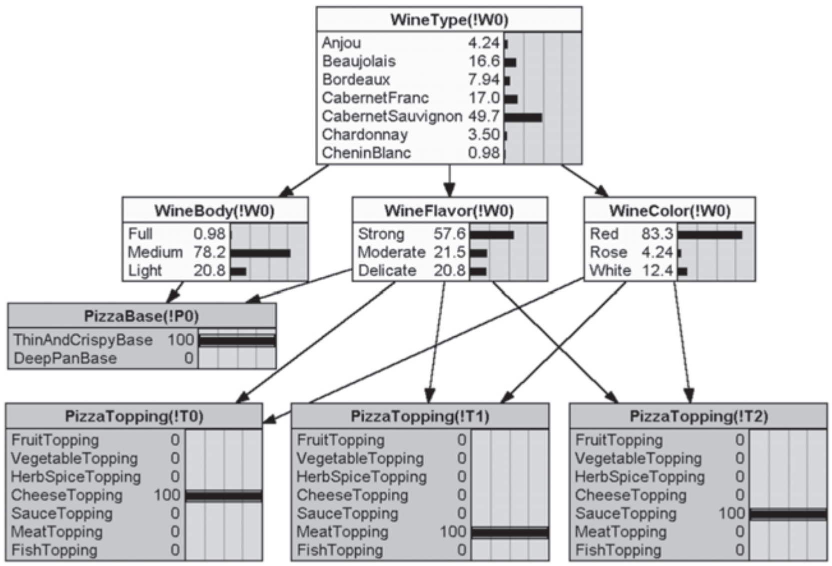


Figure 5. SSBN constructed from the MFrags of Figure 4



build SSBNs for any specific configuration of pizzas and wines for this toy example.

Clearly, this example is oversimplified, but it suffices to illustrate how PR-OWL can be used to build a probabilistic ontology combining legacy ontologies of pizzas and wines. This example illustrates the use of an expressive probabilistic language to capture knowledge that cannot be expressed with standard Bayesian networks. Probabilistic ontologies are an increasingly important topic in forums devoted to best practices in systems development. Given the nature of the domain knowledge embedded in their systems, system developers in general would profit most from the advantages of being able to convey such knowledge with a principled treatment for uncertainty.

FUZZY FORMALISMS APPLIED TO THE SEMANTIC WEB

In contrast to probabilistic formalisms, which allow for representing and processing degrees of uncertainty about ambiguous pieces of information, fuzzy formalisms allow for representing and processing degrees of truth about vague (or imprecise) pieces of information. The following examples illustrate the difference between degrees of uncertainty and degrees of truth.

Consider the statement "it will rain tomorrow". This statement is *uncertain*, that is, it is either true or false, depending on the weather conditions of tomorrow, but we generally do not have a complete knowledge about whether it will rain tomorrow or not. In probabilistic formalisms, we thus assume a set of possible worlds, each of which is associated with a probability. Intuitively, we are uncertain about which possible world is the right one. In each world, we only allow for binary truth-values, and thus in each world the statement "it will rain tomorrow" is either true or false. This way, we can quantify our ignorance about whether it will rain tomorrow or not. For example, we may say that the probability that it will rain tomorrow is 0.7, which means that the probabilities of all worlds in which it will rain tomorrow sum up to 0.7.

Consider next the statement "John is tall." This statement is *vague*, that is, it is more or less true, depending on the body height of John, but we are unable to say whether this statement is completely true or false due to the involvement of the vague concept "tall," which does not have a precise definition. In fuzzy formalisms, we assume fuzzy interpretations, which directly generalize binary interpretations by mapping elementary vague propositions into a truth value space between false and true. For example, we may say that John is tall with the degree of truth 0.7, which intuitively means that John is relatively tall but not completely tall.

It is also important to point out that vague statements are truth-functional, that is, the degree of truth of a vague complex statement (which is constructed from elementary vague statements via logical operators) can be calculated from the degrees of truth of its constituents, while uncertain complex statements are generally not a function of the degrees of uncertainty of their constituents (Dubois and Prade, 1994).

Vagueness abounds especially in multimedia information processing and retrieval. Another typical application domain for vagueness and thus fuzzy formalisms are natural language interfaces to the Web. Furthermore, fuzzy formalisms have also been successfully applied in ontology mapping, information retrieval, and e-commerce negotiation tasks.

Fuzzy Propositional Logics. Rather than being restricted to a binary truth value among false and true, *vague propositions* may also have a truth value strictly between false and true. One often assumes the unit interval [0, 1] as the set of all possible truth values, where 0 and 1 represent the ordinary binary truth values false and true, respectively. For example, the vague proposition "John is tall" may be more or less true, and it is thus associated with a truth value in [0, 1], depending on the body height of John.

To combine and modify the truth values in [0, 1], one assumes *combination functions*, namely, *conjunction*, *disjunction*, *implication*, and *negation functions*, denoted $\otimes$, $\oplus$, $\triangleright$, and $\ominus$, respectively, which are functions $\otimes, \oplus, \triangleright$: $[0, 1]\times[0, 1] \rightarrow [0, 1]$ and $\ominus$: $[0, 1] \rightarrow [0, 1]$ that generalize the ordinary logical operators $\wedge$, $\vee$, $\rightarrow$, and $\neg$, respectively, to the set of truth values [0, 1]. As usual, we assume that the combination functions have some natural algebraic properties, namely, the properties shown in Tables 1 and 2. Note that in Table 1, Tautology and Contradiction follow from Identity, Commutativity, and Monotonicity. Note also that conjunction and disjunction functions (with the properties shown in Table 1) are also called *triangular norms* and *triangular co-norms* (Hájek, 1998), respectively. The combination functions of some well-known fuzzy logics are shown in Table 3.

More formally, a *fuzzy (propositional) interpretation I* maps each elementary vague proposition *p* into the set of truth values [0,1], and is then extended inductively to all (complex) vague propositions (which are constructed from the elementary vague propositions by using the binary and unary logical operators $\wedge$, $\vee$, $\rightarrow$, and $\neg$) as follows (where $\otimes$, $\oplus$, $\triangleright$, and $\ominus$ are conjunction, disjunction, implication, and negation functions, respectively, as described above):

A fuzzy (propositional) knowledge base consists of a finite set of *fuzzy formulas,* which have one of the forms $\varphi \geq 1$, $\varphi \leq 1$, $\varphi > l$, or $\varphi < l$, where φ is a vague proposition, and l is a truth value from [0,1]. Such statements express that φ has a degree of truth of at least, at most, greater than, and lower than l, respectively. For example, $tall\_John \geq 0.6$ says that *tall_John* has a degree of truth of at least 0.6. Any such fuzzy knowledge base represents a set of fuzzy interpretations, which can be used to define the notions of *satisfiability, logical consequence,* and *tight logical consequence,* as usual. Here, it is important to point out the difference from Bayesian networks: rather than encoding one single probability distribution (over a set of binary interpretations), fuzzy knowledge bases encode a set of fuzzy interpretations.

Fuzzy Description Logics and Ontology Languages. In fuzzy description logics and ontology

Figure 6. Axioms for conjunction and disjunction functions

Axiom Name	Conjunction Function	Disjunction Function
Tautology / Contradiction	$a \otimes 0 = 0$	$a \oplus 1 = 1$
Identity	$a \otimes 1 = a$	$a \oplus 0 = a$
Commutativity	$a \otimes b = b \otimes a$	$a \oplus b = b \oplus a$
Associativity	$(a \otimes b) \otimes c = a \otimes (b \otimes c)$	$(a \oplus b) \oplus c = a \oplus (b \oplus c)$
Monotonicity	if $b \leq c$, then $a \otimes b \leq a \otimes c$	if $b \leq c$, then $a \oplus b \leq a \oplus c$

Figure 7. Axioms for implication and negation functions

Axiom Name	Implication Function	Negation Function
Tautology / Contradiction	$0 \rhd b = 1,\ a \rhd 1 = 1,\ 1 \rhd 0 = 0$	$\ominus 0 = 1,\ \ominus 1 = 0$
Antitonicity	if $a \leq b$, then $a \rhd c \geq b \rhd c$	if $a \leq b$, then $\ominus a \geq \ominus b$
Monotonicity	if $b \leq c$, then $a \rhd b \leq a \rhd c$	

languages, concept assertions, role assertions, concept inclusions, and role inclusions have a degree of truth rather than a binary truth value. Semantically, this extension is essentially obtained by (i) generalizing binary first-order interpretations to fuzzy first-order interpretations and (ii) interpreting all the logical operators by a corresponding combination function. Syntactically, as in the fuzzy propositional case, one then also allows for formulas that restrict the truth values of concept assertions, role assertions, concept inclusions, and role inclusions. Some important new ingredients of fuzzy description logics are often also fuzzy concrete domains, which include fuzzy predicates on concrete domains, and fuzzy modifiers (such as "very" or "slightly"), which are unary operators that change the membership functions of fuzzy concepts.

Figure 8. Combination functions of various fuzzy logics

	Łukasiewicz Logic	Gödel Logic	Product Logic	Zadeh Logic
$a \otimes b$	$\max(a+b-1, 0)$	$\min(a, b)$	$a \cdot b$	$\min(a, b)$
$a \oplus b$	$\min(a+b, 1)$	$\max(a, b)$	$a + b - a \cdot b$	$\max(a, b)$
$a \rhd b$	$\min(1-a+b, 1)$	$\begin{cases} 1 & \text{if } a \leq b \\ b & \text{otherwise} \end{cases}$	$\min(1, b/a)$	$\max(1-a, b)$
$\ominus a$	$1-a$	$\begin{cases} 1 & \text{if } a = 0 \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} 1 & \text{if } a = 0 \\ 0 & \text{otherwise} \end{cases}$	$1-a$

$$
\begin{aligned}
\mathcal{I}(\phi \wedge \psi) &= \mathcal{I}(\phi) \otimes \mathcal{I}(\psi) \\
\mathcal{I}(\phi \vee \psi) &= \mathcal{I}(\phi) \oplus \mathcal{I}(\psi) \\
\mathcal{I}(\phi \rightarrow \psi) &= \mathcal{I}(\phi) \rhd \mathcal{I}(\psi) \\
\mathcal{I}(\neg \phi) &= \ominus \mathcal{I}(\phi)
\end{aligned}
$$

As a fictional example, an online shop may use a fuzzy description logic knowledge base to classify and characterize its products. For example, suppose (1) textbooks are books, (2) PCs and laptops are mutually exclusive electronic products, (3) books and electronic products are mutually exclusive products, (4) PCs have a price, a memory size, and a processor speed, (5) *pc1* is a PC with the price 1300€, the memory size 3 GB, and the processor speed 4 GHz, (6) *pc2* is a PC with the price 500€, the memory size 1 GB, and the processor speed 2 GHz, (7) *pc3* is a PC with the price 900€, the memory size 2 GB, and the processor speed 3 GHz, (8) *ibm*, *acer*, and *hp* are the producers of *pc1*, *pc2*, and *pc3*, respectively. These relationships are expressed by the following description logic knowledge base:

(1) $Textbook \sqsubseteq Book$;
(2) $PC \sqcup Laptop \sqsubseteq Electronics$; $PC \sqsubseteq \neg Laptop$;
(3) $Book \sqcup Electronics \sqsubseteq Product$; $Book \sqsubseteq \neg Electronics$;
(4) $PC \sqsubseteq \exists hasPrice.Integer \sqcap \exists hasMemorySize.Integer \sqcap \exists hasProcessorSpeed.Integer$;
(5) $(PC \sqcap \exists hasPrice.1300 \sqcap \exists hasMemorySize.3 \sqcap \exists hasProcessorSpeed.4)(pc1)$;
(6) $(PC \sqcap \exists hasPrice.500 \sqcap \exists hasMemorySize.1 \sqcap \exists hasProcessorSpeed.2)(pc2)$;
(7) $(PC \sqcap \exists hasPrice.900 \sqcap \exists hasMemorySize.2 \sqcap \exists hasProcessorSpeed.3)(pc3)$;
(8) $produces(ibm, pc1)$; $produces(acer, pc2)$; $produces(hp, pc3)$.

The notions "expensive PCs", "PCs having a large memory", and "PCs having a fast processor" can then be defined as fuzzy concepts by adding the following three fuzzy concept definitions:

$ExpensivePC \equiv PC \sqcap \exists hasPrice.PCExpensive$,

$LargeMemoryPC \equiv PC \sqcap \exists hasMemorySize.MemoryLarge$,

$FastProcessorPC \equiv PC \sqcap \exists hasProcessorSpeed.ProcessorFast$.

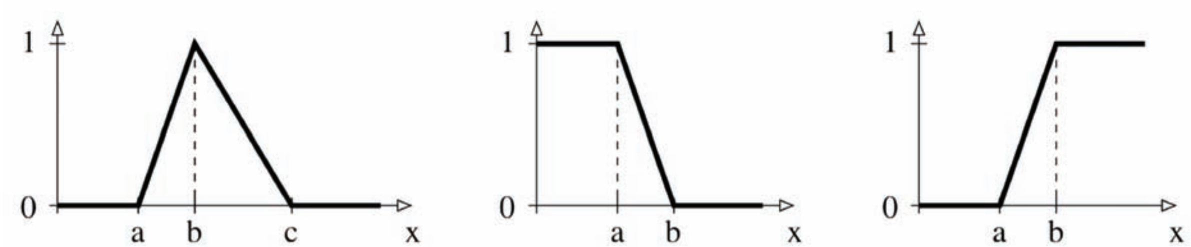
Here, *PCExpensive*, *MemoryLarge*, and *ProcessorFast* are fuzzy unary datatype predicates, which are defined by $PCExpensive(x) = rs(x; 600, 1200)$, $MemoryLarge(x) = rs(x; 1, 3)$, and $ProcessorFast(x) = rs(x; 2, 4)$, respectively, where $rs(x; a, b)$ is the so-called right-shoulder function (see Figure 9). Informally, as for the fuzzy concept "expensive PCs", every PC costing at least 1200€ (resp., at most 600€) is definitely expensive (resp., not expensive), while every PC costing between 600€ and 1200€ is expensive to some degree between 0 and 1.

Similarly, the notions "costs at most about 1000€" and "has a memory size of around 2 GB" in a buyer's request can be expressed through the following fuzzy concepts *C* and *D*, respectively:

$C \equiv \exists hasPrice.LeqAbout1000$ and $D \equiv \exists hasMemorySize.Around2$,

where $LeqAbout1000 = ls(500, 1500)$ and $Around2 = tri(1.5, 2, 2.5)$ (see Figure 9).

Figure 9. (a) triangular function tri(x; a, b, c), (b) left-shoulder function ls(x; a, b), and (c) right-shoulder function rs(x; a, b)



The literature contains many different approaches to fuzzy extensions of description logics and ontology languages. They can be roughly classified according to (a) the description logics or the ontology languages that they generalize, (b) the fuzzy constructs that they allow, (c) the fuzzy logics that they are based on, and (d) their reasoning algorithms.

One of the earliest works is due to Yen (1991), who proposes a fuzzy extension of a quite restricted sublanguage of *ALC*. Yen considers fuzzy terminological knowledge, along with fuzzy modifiers, but no fuzzy assertional knowledge, and he uses Zadeh Logic as underlying fuzzy logic. Yen's work also includes a reasoning algorithm, which allows for testing crisp subsumptions. Tresp and Molitor's work (1998) presents a more general fuzzy extension of *ALC*. Like Yen's work, it also includes fuzzy terminological knowledge, along with a special form of fuzzy modifiers, but no fuzzy assertional knowledge, and it is based on Zadeh Logic. The reasoning algorithm of Tresp and Molitor's work is a tableaux calculus for computing subsumption degrees.

Another important fuzzy extension of *ALC* is due to Straccia (1998, 2001), who allows for both fuzzy terminological and fuzzy assertional knowledge, but not for fuzzy modifiers, and again assumes Zadeh Logic as underlying fuzzy logic. Straccia's work also includes a tableaux calculus for deciding logical consequences and computing tight logical consequences. Hölldobler *et al.* (2002, 2005) extend Straccia's fuzzy *ALC* with fuzzy modifiers of the form $f_m(x) = x^\beta$, where $\beta > 0$, and present a sound and complete reasoning algorithm for the graded subsumption problem.

Straccia (2004) shows how reasoning in fuzzy *ALC* under Zadeh Logic can be reduced to reasoning in classical *ALC*. This idea has also been explored by Li *et al.* (2005a, 2005b).

Approaches towards more expressive fuzzy description logics include the works by Sanchez and Tettamanzi (2004, 2006), who consider the description logic *ALCQ*. They introduce the new notion of *fuzzy quantifiers*. As underlying fuzzy logic, they also assume Zadeh Logic. Their reasoning algorithm calculates the satisfiability interval for a fuzzy concept. Straccia (2005c) defines the semantics of a fuzzy extension of *SHOIN(**D**)*, which is the description logic that stands behind OWL DL. Stoilos *et al.* (2005a) use this semantics to define a fuzzy extension of the OWL language, and also propose a translation of fuzzy OWL to fuzzy *SHOIN*.

Other works include the one by Hájek (2005, 2006), who considers *ALC* under arbitrary t-norms and proposes especially a reasoning algorithm for testing crisp subsumptions. Bonatti and Tettamanzi (2006) provide some complexity results for reasoning in fuzzy description logics.

Recent works by Straccia (2005b, 2005a) present a calculus for *ALC(**D**)*, which works whenever the connectives, the fuzzy modifiers, and the concrete fuzzy predicates are representable as bounded mixed integer linear programs. For example, Łukasiewicz logic satisfies these conditions. The method has been extended to fuzzy *SHIF(**D**)*, which is the description logic standing behind OWL Lite, and a reasoner (called *fuzzyDL)* supporting Zadeh, Łukasiewicz, and classical semantics has been implemented and is available from Straccia's web page.

Towards reasoning in fuzzy *SHOIN(**D**)*,Stoilos *et al.* (2005, 2005b) show results providing a tableaux calculus for fuzzy *SHIN* without fuzzy general concept inclusions and under the Zadeh semantics. Stoilos et al. (2006) provide a generalization thereof that additionally allows for fuzzy general concept inclusions. In closely related work, Li *et al.* (2006) provide a tableaux calculus for fuzzy *SHI* with fuzzy general concept inclusions.

FUTURE RESEARCH DIRECTIONS

As the Semantic Web makes its transition from a vision to implementation, many of its stakeholders begin to feel the need to represent and reason under uncertainty. SW applications being developed for domains of knowledge in which uncertainty plays a significant role must include a means to store and retrieve incomplete knowledge. To cite just a few examples from the medical domain[f], statistical regularities linking a given protein to (say) Alzheimer's disease, the predisposition of patients with gene X towards developing cancer Y, or gene ontology evidence codes in support of a particular GO annotation of a gene can be all considered specific instances of more general cases in which a principled means for representing incomplete knowledge is needed. Similar situations can be observed in other domains of knowledge currently being studied in the context of the Semantic Web, which makes uncertainty representation and reasoning a rapidly growing field of SW research.

In early 2007, the W3C approved an incubator group to focus on uncertain knowledge. The URW3-XG[g] has the overall mission of better defining the challenge of reasoning with and representing uncertain information available through the WWW and related technologies. Accomplishing this mission involves identifying problems for which uncertainty is an essential aspect, to produce use cases, and to identify requirements for knowledge representation and reasoning when crisp truth-values are unknown or inappropriate. It is important to emphasize the fact that the group's scope does not include recommending a single methodology, but to investigate whether standard representations of uncertainty can be identified that will support requirements across a wide spectrum of reasoning approaches.

This stance is compatible with what we see as the future in this area. For most domains of knowledge, the task of representing the various distinct forms of uncertainty that might have a strong influence in the way knowledge is represented and applied is complex enough that the search for a "silver bullet" is inadvisable. That is, selecting a specific approach to be the definitive one for representing uncertainty would be a recipe for failure. In fact, a combination of approaches might be the best way to address the SW use cases involving uncertainty. For now, research on uncertain knowledge applied to the Semantic Web is gaining momentum but still lacks clear definitions, use cases, and applications. This state of affairs makes it difficult for developers to create useful solutions to most problems drawn from uncertainty-plagued domains, but the current rate of progress makes it clear to us that major change is near at hand.

REFERENCES

Baader, F., & Nutt, W. (2003). Basic description logics. In Baader, F., Calvanese, D., McGuiness, D., Nardi, D., & Patel-Schneider, P. (Eds.), *The Description Logics Handbook: Theory, Implementation and Applications*. 1st edition, *2*, 47-100. Cambridge, UK: Cambridge University Press.

Bangsø, O., & Wuillemin, P.-H. (2000). *Object oriented Bayesian networks: A framework for top-down specification of large Bayesian networks and repetitive structures*. Technical Report No. CIT-87.2-00-obphw1. Department of Computer Science, Aalborg University, Aalborg, Denmark.

Berners-Lee, T., & Fischetti, M. (2000). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. 1st edition. New York: HarperCollins Publishers.

Bonatti, P., & Tettamanzi, A. (2006). Some complexity results on fuzzy description logics. In Di Gesu, V., Masulli, F., & Petrosino, A. (Eds.), *Fuzzy logic and applications, 2955 of LNCS*, 19-24. Springer.

Brachman, R. J. (1977). What's in a concept: Structural foundations for Semantic networks. *International Journal of Man-Machine Studies, 9*(2), 127–152. doi:10.1016/S0020-7373(77)80017-5

Buntine, W. L. (1994). *Learning with graphical models*. Technical Report No. FIA-94-03. NASA Ames Research Center, Artificial Intelligence Research Branch.

Calvanese, D., & De Giacomo, G. (2003). Expressive description logics. In Baader, F., Calvanese, D., McGuiness, D., Nardi, D., & Patel-Schneider, P. (Eds.), *The Description Logics Handbook: Theory, Implementation and Applications*. 1st edition, *5*, 184-225. Cambridge, UK: Cambridge University Press.

Charniak, E. (1991). Bayesian networks without tears. *AI Magazine*, *12*, 50–63.

Codd, E. F. (1970). A relational model for large shared data banks. *Communications of the ACM*, *13*(6), 377–387. doi:10.1145/362384.362685

Costa, P. C. G. (2005). *Bayesian Semantics for the Semantic Web*. Doctoral dissertation. Department of Systems Engineering and Operations Research, George Mason University: Fairfax, VA, USA. p. 312.

Costa, P. C. G., & Laskey, K. B. (2006, November 9-11). PR-OWL: A framework for probabilistic ontologies. In *Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS 2006)*. Baltimore, MD, USA.

Damasio, C., Pan, J., Stoilos, G., & Straccia, U. (2006). An approach to representing uncertainty rules in RuleML. In *Proceedings of the 2nd International Conference on Rules and Rule Markup Languages for the Semantic Web (RuleML-06)*. IEEE Computer Society. Available at http://2006.ruleml.org/online-proceedings/submission_24.pdf

Ding, Z. (2005). *BayesOWL: A probabilistic framework for Semantic Web*. Doctoral dissertation. Computer Science and Electrical Engineering. University of Maryland, Baltimore County: Baltimore, MD, USA. p. 168.

Ding, Z., & Peng, Y. (2004, January 5-8). A probabilistic extension to ontology language OWL. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04)*. Big Island, Hawaii, USA.

Dubois, D., & Prade, H. (1994). Can we enforce full compositionality in uncertainty calculi? [AAAI Press.]. *Proceedings*, *AAAI-1994*, 149–154.

Frege, G. (1879). *Begriffsschrift*, 1879, translated in Jean van Heijenoort, ed., *From Frege to Gödel*, Cambridge, MA: Harvard University Press.

Fukushige, Y. (2004). Representing probabilistic knowledge in the Semantic Web, *W3C Workshop on Semantic Web for Life Sciences*. Cambridge, MA, USA.

Getoor, L., Friedman, N., Koller, D., & Pfeffer, A. (2001). *Learning probabilistic relational models*. New York, NY, USA: Springer-Verlag.

Getoor, L., Koller, D., Taskar, B., & Friedman, N. (2000). Learning probabilistic relational models with structural uncertainty. *Paper presented at the ICML-2000 Workshop on Attribute-Value and Relational Learning:Crossing the Boundaries*. Stanford, CA, USA.

Gilks, W., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modeling. *The Statistician*, *43*, 169–178. doi:10.2307/2348941

Giugno, R., & Lukasiewicz, T. (2002, September 23-26). P-SHOQ(D): A probabilistic extension of SHOQ(D) for probabilistic ontologies in the Semantic Web. In Flesca, S.; Greco, S.; Leone, N.; and Ianni, G. (Eds.), *Proceedings of the Eight European Conference on Logics in Artificial Intelligence (JELIA 2002)* Cosenza, Italy (LNCS 2424, pp. 86-97). Berlin, Germany: Springer-Verlag.

Hájek, P. (1998). *Metamathematics of fuzzy logic*. Kluwer.

Hájek, P. (2005). Making fuzzy description logics more expressive. *Fuzzy Sets and Systems, 154*(1), 1–15. doi:10.1016/j.fss.2005.03.005

Hájek, P. (2006). What does mathematical fuzzy logic offer to description logic? In Sanchez, E. (Ed.), *Capturing Intelligence: Fuzzy Logic and the Semantic Web.* Elsevier.

Heckerman, D., Mamdani, A., & Wellman, M. P. (1995). Real-world applications of Bayesian networks. *Communications of the ACM, 38*(3), 24–68. doi:10.1145/203330.203334

Heckerman, D., Meek, C., & Koller, D. (2004). *Probabilistic models for relational data*. Technical Report MSR-TR-2004-30, Microsoft Corporation, March 2004. Redmond, WA, USA.

Heinsohn, J. (1994, July 29-31). Probabilistic description logics. *Paper presented at the Tenth Conference on Uncertainty in Artificial Intelligence (UAI-94)*.Seattle, WA, USA.

Hölldobler, S., Khang, T. D., & Störr, H.-P. (2002). A fuzzy description logic with hedges as concept modifiers. In *Proceedings InTech/VJFuzzy-2002*, 25-34.

Hölldobler, S., Nga, N. H., & Khang, T. D. (2005). The fuzzy description logic *ALC*flh. In *Proceeedings DL-2005*.

Horridge, M., Knublauch, H., Rector, A., Stevens, R., & Wroedn, S. (2004). *A practical guide to building OWL ontologies using the Protégé-OWL plugin and CO-ODE tools*. The University of Manchester. Retrieved June 9, 2007, from http://www.co-ode.org.

Horrocks, I. (2002). DAML+OIL: A reasonable Web ontology language. Keynote talk at the WES/CAiSE Conference. Toronto, Canada.

Horrocks, I., & Sattler, U. (2001, August 4-10). Ontology reasoning in the SHOQ(D) description logic. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI 2001)*. Seattle, WA, USA.

Jaeger, M. (1994, May 24-27). Probabilistic reasoning in terminological logics. *Paper presented at the Fourth International Conference on Principles of Knowledge Representation and Reasoning* (KR94). Bonn, Germany.

Jaeger, M. (1997, August 1-3). Relational Bayesian networks. *Paper presented at the 13th Annual Conference on Uncertainty in Artificial Intelligence (UAI-97)*.Providence, RI, USA.

Jaeger, M. (2006). Probabilistic role models and the guarded fragment. In *Proceedings IPMU-2004*, pp. 235–242. Extended version in *Int. J. Uncertain. Fuzz.*, *14*(1), 43–60. doi:10.1142/S0218488506003820

Koller, D., Levy, A. Y., & Pfeffer, A. (1997, July 27-31). P-CLASSIC: A tractable probabilistic description logic. *Paper presented at the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*. Providence, RI, USA.

Koller, D., & Pfeffer, A. (1997). Object-oriented Bayesian networks. *Paper presented at the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI-97*). San Francisco, CA, USA.

Kolmogorov, A. N. (1960). *Foundations of the Theory of Probability.* 2nd edition. New York, NY, USA: Chelsea Publishing Co. Originally published in 1933.

Langseth, H., & Nielsen, T. (2003, July). Fusion of domain knowledge with data for structured learning in object-oriented domains. *Journal of Machine Learning Research . Special Issue on the Fusion of Domain Knowledge with Data for Decision Support*, *4*, 339–368.

Laskey, K. B. (2007). MEBN: A language for first-order Bayesian knowledge bases. *Artificial Intelligence*, *172*(2-3).

Laskey, K. B., & Costa, P. C. G. (2005). Of Klingons and starships: Bayesian logic for the 23rd Century, in *Uncertainty in Artificial Intelligence: Proceedings of the Twenty-first Conference*. Edinburgh, Scotland: AUAI Press.

Laskey, K. B., & Mahoney, S. M. (1997, August). Network fragments: Representing knowledge for constructing probabilistic models. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI-97)*. Providence, RI, USA.

Li, Y., Xu, B., Lu, J., & Kang, D. (2006). Discrete tableau algorithms for *SHI*. In *Proceeedings DL-2006.*

Li, Y., Xu, B., Lu, J., Kang, D., & Wang, P. (2005a). Extended fuzzy description logic *ALCN*. In *Proceedings KES-2005, 3684 of LNCS,* 896-902. Springer.

Li, Y., Xu, B., Lu, J., Kang, D., & Wang, P. (2005b). A family of extended fuzzy description logics. In *Proceedings COMPSAC-2005*, 221-226. IEEE Computer Society.

Lukasiewicz, T. (2002). Probabilistic default reasoning with conditional constraints. *Annals of Mathematics and Artificial Intelligence*, *34*(1/3), 35–88. doi:10.1023/A:1014445017537

Lukasiewicz, T. (2008). Expressive probabilistic description logics. *Artificial Intelligence*, *172*(6-7), 852–883. doi:10.1016/j.artint.2007.10.017

Minsky, M. L. (1975). Framework for representing knowledge. In *The Psychology of Computer Vision*. P. H. Winston (Eds.), 211-277. New York, NY: McGraw-Hill.

Mitra, P., Noy, N. F., & Jaiswal, A. R. (2004, November). OMEN: A probabilistic ontology mapping tool. *Workshop on Meaning Coordination and Negotiation at the Third International Conference on the Semantic Web (ISWC-2004)*. Hisroshima, Japan.

Neapolitan, R. E. (1990). *Probabilistic reasoning in expert systems: Theory and algorithms*. New York, NY, USA: John Wiley and Sons, Inc.

Neapolitan, R. E. (2003). *Learning Bayesian networks*. New York: Prentice Hall.

Pan, R., Ding, Z., Yu, Y., & Peng, Y. (2005, November). A Bayesian approach to ontology mapping. In *Proceedings of the Fourth International Semantic Web Conference (ISWC-2005)*. Galway, Ireland.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA, USA: Morgan Kaufmann Publishers.

Peirce, C. S. (1885). On the algebra of logic. *American Journal of Mathematics*, *7*, 180–202. doi:10.2307/2369451

Peng, Y., Ding, Z., Pan, R., Yu, Y., Kulvatunyou, B., Izevic, N., et al. (2007, May). A probabilistic framework for semantic similarity and ontology mapping. In *Proceedings of the 2007 Industrial Engineering Research Conference (IERC)*. Nashville, TN, USA.

Pfeffer, A. (2001, August 4-10). IBAL: A probabilistic rational programming language international. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001)*, *1*, 733-740. Seattle, WA, USA.

Pfeffer, A., Koller, D., Milch, B., & Takusagawa, K. T. (1999, July-August 1). SPOOK: A system for probabilistic object-oriented knowledge representation. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 541-550. Stockholm, Sweden

Pool, M., & Aikin, J. (2004, July 6-9). KEEPER and protégé: An elicitation environment for Bayesian inference tools. *Paper presented at the Workshop on Protégé and Reasoning held at the Seventh International Protégé Conference.* Bethesda, MD, USA.

Ramsey, F. P. (1931). *The Foundations of Mathematics and other Logical Essays*. London, UK: Kegan Paul, Trench, Trubner & Co.

Sanchez, D., & Tettamanzi, A. (2004). Generalizing quantification in fuzzy description logics. In *Proceedings 8th Fuzzy Days in Dortmund.*

Sanchez, D., & Tettamanzi, A. (2006). Fuzzy quantification in fuzzy description logics. In Sanchez, E. (Ed.), *Capturing Intelligence: Fuzzy Logic and the Semantic Web.* Elsevier.

Schmidt-Schauß, M., & Smolka, G. (1991). Attributive concept descriptions with complements. *Artificial Intelligence*, *48*(1), 1–26. doi:10.1016/0004-3702(91)90078-X

Spiegelhalter, D. J., Thomas, A., & Best, N. (1996). Computation on graphical models. *Bayesian Statistics*, *5*, 407–425.

Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J. Z., & Horrock, I. (2005). A fuzzy description logic for multimedia knowledge representation. In *Proceedings of the International Workshop on Multimedia and the Semantic Web*.

Stoilos, G., Stamou, G. B., Tzouvaras, V., Pan, J. Z., & Horrocks, I. (2005b). The fuzzy description logic *f-SHIN*. In *Proceedings URSW-2005*, 67-76.

Stoilos, G., Straccia, U., Stamou, G., & Pan, J. Z. (2006). General concept inclusions in fuzzy description logics. In *Proceedings ECAI-2006*, 457-61. IOS Press.

Straccia, U. (1998). A fuzzy description logic. In *Proceedings AAAI-1998*, 594-599. AAAI Press/ MIT Press.

Straccia, U. (2001). Reasoning within fuzzy description logics. *Journal of Artificial Intelligence Research*, *14*, 137–166.

Straccia, U. (2004). Transforming fuzzy description logics into classical description logics. In *Proceedings JELIA-2004*, *3229* of *LNCS*, 385-399. Springer.

Straccia, U. (2005a). Description logics with fuzzy concrete domains. In *Proceedings UAI-2005*, 559-567. AUAI Press.

Straccia, U. (2005b). Fuzzy *ALC* with fuzzy concrete domains. In *Proceeedings DL-2005*, 96-103.

Tresp, C., & Molitor, R. (1998). A description logic for vague knowledge. In *Proceedings ECAI-1998*, 361-365. J. Wiley & Sons.

Yelland, P. M. (2000). An alternative combination of Bayesian networks and description logics. In *Proceedings KR-2000*, 225–234. Morgan Kaufmann.

Yen, J. (1991). Generalizing term subsumption languages to fuzzy logic. In *Proceedings IJCAI-1991*, 472-177. Morgan Kaufmann.

ADDITIONAL READING

Costa, P. C. G., Fung, F., Laskey, K. B., Laskey, K. J., & Pool, M. (2006, November 5). *Proceedings of the ISWC Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2006)*. Athens, GA, USA. Available at http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-218/.

In addition, the literature listed below includes recent material on distinct approaches to representing and reasoning with uncertainty: Costa, P. C. G., Laskey, K. B., Laskey, K. J., & Pool, M. (2005. November 7). *Proceedings of the ISWC Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2005)*. Galway, Ireland. Available at http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-173/.

Due to the initial stage of research on the subject, there are no specifications yet for representing and reasoning with uncertainty and thus no SW applications based on commonly accepted standards. The first step towards standardization is already being taken by the W3C via the already cited URW3 XG incubator group. At the time of this writing, the group was finishing its report and listing use cases for possible uncertainty-aware SW applications.

Jousselme, A. L., Maupin, P., & Bosse, E. (2003, July 8-12). Uncertainty in a situation analysis Perspective. In *Proceedings of the Sixth International Conference of Information Fusion, 2*, 1207-1214. 2003, Cairns, Queensland, Australia.

Lukasiewicz, T. (2005). Probabilistic Description Logic Programs. In *Proceedings ECSQARU 2005*, Barcelona, Spain, July 2005, *3571* of *LNCS*, pp. 737-749. Springer. Extended version: *International Journal of Approximate Reasoning 45*(2), 288-307, 2007.

Lukasiewicz, T. (2006). Fuzzy Description Logic Programs under the Answer Set Semantics for the Semantic Web. In *Proceedings of the 2nd International Conference on Rules and Rule Markup Languages for the Semantic Web (RuleML-06)*, 89-96. IEEE Computer Society. Extended version: *Fundamenta Informaticae 82*, 1-22, 2008.

Lukasiewicz, T., & Straccia, U. (2007). Description logic programs under probabilistic uncertainty and fuzzy vagueness. In *Proceedings ECSQARU 2007*, Hammamet, Tunisia, October/November 2007, *4724* of *LNCS*, 187-198. Springer.

Mitra, P., Noy, N., & Jaiswal, A. R. (2005, November 7). Ontology mapping discovery with uncertainty. presented at the *Fourth International Semantic Web Conference (ISWC 2004*). Galway, Ireland.

Pan, J. Z., Stoilos, G., Stamou, G., Tzouvaras, V., & Horrocks, I. (2006). f-SWRL: A Fuzzy Extension of SWRL. In *Data Semantics*, special issue on Emergent Semantics, *4090*/2006: 28-46.

Parsons, S. (1996, June). Current approaches to handling imperfect information in data acknowledgement bases. [Los Alamitos, CA, USA: IEEE Computer Society.]. *IEEE Transactions on Knowledge and Data Engineering*, *8*(3), 353–372. doi:10.1109/69.506705

Sanchez, E. (2006). *Fuzzy logic and the Semantic Web*. 1st edition, April 3, 2006. Oxford, UK: Elsevier Science.

Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J. Z., & Horrocks, I. (2005a). Fuzzy OWL: Uncertainty and the Semantic Web. In *Proceedings of the International Workshop on OWL: Experience and Directions (OWL-ED2005)*.

Straccia, U. (2005c). Towards a fuzzy description logic for the Semantic Web. In *Proceedings of the Second European Semantic Web Conference, ESWC 2005*.

The interested reader is strongly encouraged to browse the use cases being studied by the XG, which convey formalized details of some of the most promising use cases analyzed by the group. The use cases are described in the XG Report, available from http://www.w3.org/2005/Incubator/urw3/. The use cases there were taken from a large list initially considered, and included various domains of knowledge such as discovery, appointment making, Healthcare and Life Sciences, ontology mapping, belief fusion and opinion pooling, shopping software agents, large-scale database retrieval and reasoning, and many others.

Yang, Y., & Calmet, J. (2005). OntoBayes: An ontology-driven uncertainty model. *Presented at the International Conference on Intelligent Agents, Web Technologies and Internet Commerce (IAWTIC2005)*. Vienna, Austria. Available at http://iaks-www.ira.uka.de/iaks-calmet/papers/IAWTIC05_yang.pdf.

ENDNOTES

[a] See http://www.w3.org/TR/2006/WD-rif-ucr-20060710/ and http://www.w3.org/TR/rif-core/.

[b] Inspired by the wine ontology available at http://protege.cim3.net/cgi-bin/wiki.pl?ProtegeOntologiesLibrary and the pizza ontology presented in Horridge *et al.* (2004).

[c] PR-OWL is available from http://www.pr-owl.org.

[d] The pentagon nodes are context nodes, representing constraints that must be satisfied for the distributions in the MFrag to apply. The trapezoid nodes are input nodes, whose probability distribution is defined outside the MFrag. The oval nodes are resident nodes, whose distributions are defined in the MFrag.

[e] In MEBN, RVs take arguments that refer to entities in the domain of application. An interpretation of the theory uses entity identifiers as labels to refer to entities in the domain. Entity identifiers are written either as numerals or as alphanumeric strings beginning with an exclamation point, e.g., !M3, 48723.

[f] See http://esw.w3.org/topic/HCLS/UncertaintyUseCases for a more comprehensive analysis.

[g] Charter available at http://www.w3.org/2005/Incubator/urw3/charter.

APPENDIX: QUESTIONS FOR DISCUSSION

1. Cite a typical problem needing uncertainty representation and reasoning within the context of the Semantic Web.
2. Figure 2 displays the Bayesian Network from Figure 1 after evidence on pizza topping and base was included. Based on the results, which wine should a waiter suggest to the customer? Why?
3. What is "under the hood" of that reasoning process? How did the BN arrive at its conclusion?
4. What is the major limitation of BNs when applied to the SW?
5. What is the major advantage of using probabilistic extensions to DLs?
6. If Probabilistic FOL approaches cannot guarantee decidability in many cases, why should anyone care about using them?
7. What is the difference between uncertainty and vagueness?
8. Which are typical application areas for fuzzy formalisms?
9. What is the difference between probabilistic and fuzzy propositional interpretations? What is the difference between the semantics of Bayesian Networks and fuzzy propositional knowledge bases?

Answers:

1. See Introduction of this chapter.
2. According to the knowledge stored in that BN and the evidence entered, the waiter should suggest a red wine with medium to light body, delicate flavor. His first choice should be Beaujolais. Expert information from sommeliers and statistical regularities on previous orders were used to build this model, which allows it to make the best choice possible with the available information. For purposes of this example, "best" means the most likely wine a customer would order given the evidence.
3. The BN model uses Bayes rule to update the beliefs displayed in each of its nodes. After evidence was entered, an algorithm performed the belief updating in real time. See Charniak (1991) for a good introduction on BNs.
4. Apart from their flexibility and inferential power, BNs have only assertional expressivity and cannot represent situations where each problem instance involves reasoning about the same fixed number of attributes, with only the evidence values changing from problem instance to problem instance. This is insufficient for most SW problems. However, BNs can be applied in very specific cases, for which the representational power of BNs is sufficient.
5. Description logics are highly effective and efficient for the classification and subsumption problems they were designed to address. They provide decidability and their probabilistic extensions allow for representation of both numerical and structural aspects of a probabilistic model.
6. First Order Probabilistic Approaches have sufficient representational power for most real world problems. Additionally, they can provide ontology designers with a range of options to balance tractability against expressiveness.
7. Uncertainty expresses the lack of knowledge about binary statements, while vague-ness expresses the inherent imprecision of many-valued statements.

8. Some typical application areas for fuzzy formalisms are multimedia information processing and retrieval, natural language interfaces to the Web, ontology mapping, information retrieval, and e-commerce negotiation tasks.
9. A probabilistic propositional interpretation maps binary interpretations to a degree of likelihood in [0, 1], while a fuzzy propositional interpretation maps elementary vague propositions to a degree of truth in [0, 1]. A Bayesian Network encodes a probability distribution over a set of binary interpretations, while a fuzzy propositional knowledge bases encodes a set of fuzzy propositional interpretations.

This work was previously published in Semantic Web Engineering in the Knowledge Society, edited by J. Cardoso; M. Lytras, pp. 315-340, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 7.4
Semantic Web-Enabled Protocol Mediation for the Logistics Domain

Oscar Corcho
Universidad Politécnica de Madrid, Spain

Silvestre Losada
Intelligent Software Components, S.A., Spain

Richard Benjamins
Intelligent Software Components, S.A., Spain

ABSTRACT

Among the problems that arise when trying to make different applications interoperate with each other, protocol mediation is one of the most difficult ones and for which less relevant literature can be found. Protocol mediation is concerned with non-matching message interaction patterns in application interaction. In this chapter we describe the design and implementation of a protocol mediation component that has been applied in the interoperation between two heterogeneous logistic provider systems (using two different standards: RosettaNet and EDIFACT), for a specific freight forwarding task.

DOI: 10.4018/978-1-60566-066-0.ch004

CURRENT SITUATION

Logistics is defined as the art and science of managing and controlling the flow of goods, energy, information and other resources like products, services and people from the source of production to the marketplace. As pointed out by Evans-Greenwood and Stason (2006) the current trend in logistics is to divide support between planning applications, which compute production plans overnight, and execution applications, which manage the flow of events in an operational environment. This disconnection forces users to deal with business exceptions (lost shipments, for example), manually resolving the problems by directly updating the execution and planning applications. However, this human-dependency problem can be ameliorated by using Web technology to create a heterogeneous composite application involving all participants in the process, providing a complete Third-Party Logistics solution, and giving users a single unified view into the logistics pipeline. This consolidated logistics solution greatly simplifies the task of identifying and correcting business exceptions (e.g., missing shipments or stock shortages) as they occur. Therefore, logistics management is a typical business problem where the use of a service oriented architecture is clearly suited.

Furthermore, Evans-Greenwood and Stason (2006) also talk about the possibility of combining multiple Third-Party Logistics solutions into a single heterogeneous virtual logistics network. With such a virtual network, each shipment is assigned a route dynamically assembled from one or more individual logistics providers, using dynamically created virtual supply chains. Most of these business functions are still manual and offline, but most of them can be automated with the use of service oriented architectures, as will be presented in this chapter. Obviously, the main advantages of using such solutions are the decreases in cost and speed in transactions, which influence in a better quality of the service provided to customers.

The main barrier to set up a business relationship with a company in the logistics domain is that it usually requires an initial large investment of time and money. This is ameliorated by the emergence of some industry standards like EDIFACT (EDIFACT), AnsiX12 (AnsiX12) or RosettaNet (RosettaNet), which ease the integration tasks between information systems that comply with them. However, given that these standards have some flexibility in what respects the content and sequencing of the messages that can be exchanged, the integration of systems is still time and effort consuming. Besides, there is sometimes a need to integrate systems that use different standards, what makes the integration task even more time and effort consuming.

This is the focus of one of the four case studies developed in the context of the EU project SWWS[1] (Semantic-Web enabled Web Services), a demonstrator of business-to-business integration in the logistics domain using Semantic Web Service technology. All the features of this demonstrator are described in detail by Preist and colleagues (2005), including aspects related to the discovery and selection of relevant services, their execution and the mediation between services following different protocols.

In this chapter we will focus on the last aspect (mediation) and more specifically on protocol mediation, which is concerned with the problem of non-matching message interaction patterns. We will describe the design and implementation of the protocol mediation component applied in this case study to show how to make logistic provider systems using two different standards (RosettaNet and EDIFACT) interoperate for a specific freight forwarding task.

The chapter is structured as follows. The rest of this section introduces a motivating example, focusing on the needs for protocol mediation, and gives some background on how the problem of mediation can be characterised in general and on the approaches for mediation proposed in the context of Semantic Web Service research. Section 2 summarises the protocol mediation approach followed for this case study and the main elements to be considered inside the approach. It also describes the ontology used for the description of the abstract and concrete protocols used by the entities involved in the message exchange. Section 3 provides an overview of the API of the protocol mediation component and gives details about how to configure it for deployment. Finally, section 4 gives some conclusions.

An Example in the Logistics Domain

Let us imagine that we have a manufacturing company in Bristol, UK, which needs to distribute goods internationally. The company outsources transportation into other companies, which offer *Freight Forwarding Services*. These companies may be providing the transportation service by themselves or just act as intermediaries, but this is not important for the manufacturing company. However, the manufacturing company still needs to manage relationships with these service providers, as a *Logistics Coordinator*, being responsible for selecting the service providers, reaching agreements with them with respect to the nature of the service that they will provide, coordinating the activity of different service providers so as to ensure that they link seamlessly to provide an end-to-end service (e.g., if a ship company transports a goods to a port, then the ground transportation company should be waiting for those goods with a truck to transport them to an inland city), etc.

The manufacturing company uses EDIFACT for its exchange of messages with the service providers. However, not all of them use this standard, but in some cases RosettaNet. So the situation can be that two different companies that can offer the same service (e.g., road transportation inside Germany) are using two different standards and the logistics coordinator should be able to use any of them, independently of the protocol that they use in their information systems, taking only into account the business requirements that the parcel delivery may have (quality of service, speed, price, insurance, etc.). In this situation there is a need for a seamless integration of a mediation component that is able to capture the EDIFACT messages sent by the Logistics Coordinator into RosettaNet ones that are sent to the corresponding Freight Forwarding Service, and vice versa, without any change to the information systems of any of the parties involved.

Mediation in Service Oriented Architectures and in Semantic Web Services

In **service oriented architectures**, mediation services are middleware services that are in charge of resolving inconsistencies between the parties involved in a sequence of message exchanges. Mediation can be considered at different levels:

- **Data mediation:** Transformation of the syntactic format of the messages.
- **Ontology mediation:** Transformation of the terminology used inside the messages.
- **Protocol or choreography mediation:** Transformation of sequences of messages, to solve the problem of non-matching message interaction patterns.

All types of mediation are important to achieve a successful communication between the services involved in an application, and each of them poses different challenges. In this chapter we will focus on aspects related to the last type of mediation, which is the one aimed at ensuring that, from a high-level point of view, the services involved in a message exchange achieve their overall goals. In other words, it aims at mapping the patterns of conceptually similar, but mechanically different interaction protocols sharing a similar conceptual model of a given domain.

The atomic types of mismatches that can be found between a set of interaction patterns are (Cimpian and Mocan, 2005):

- **Unexpected messages:** One of the parties does not expect to receive a message issued by another. For instance, in a request for the delivery of a parcel the logistics provider sends the parcel weight and size, the departure place and the arrival place, while the freight forwarding service does not expect the parcel weight and size, since it will not use this information.
- **Messages in Different Order:** The parties involved in a communication send and receive messages in different orders. In the previous case the sender may send the messages in the order specified above while the receiver expects first the arrival

place and then the departure place.

- **Messages that Need to be Split:** One of the parties sends a message with multiple informations inside it, which needs to be received separately by the other party. In the previous example, the sender sends the arrival and departure places in one message, while the receiver expects it as two messages.
- **Messages that Need to be Combined:** One of the parties sends a set of messages that the receiver expects as a single message with the multiple information. We can think of the inverse situation to the one aforementioned.
- **Dummy Acknowledgements or Virtual Messages that Have to be Sent:** One of the parties expects an acknowledgement for a certain message, but the receiver does not issue such acknowledgement; or the receiver expects a message that the sender is not prepared to send.

One of the purposes of the work on **Semantic Web Services** is the automation of some of the tasks involved in the development of applications that follow a service oriented architecture. As a result, some work on mediation has been done in the area. If we focus on protocol mediation, we can find the following two approaches:

Priest and colleagues (2005) and Williams and colleagues (2006) describe the approach followed in the context of SWWS, and which will be described in more detail in the next section. This approach is based on the use of a general abstract state machine that represents the overall state of the communication between parties, and a set of abstract machines for each of the parties in the conversation, which specify their state and the sets of actions to be performed when they receive a set of messages or when they have to send a set of messages.

In the context of the WSMO initiative, Cimpian and Mocan (2005) describe the approach taken for the design and implementation of the process mediator for the Semantic Web Service execution engine WSMX. This approach is similar to the previous one, since it is also based on the use of an abstract machine with guarded transitions that are fired by the exchange of messages and the definition of choreographies for each of the parties involved in the communication.

PROPOSED SOLUTION: THE SWWS APPROACH FOR PROTOCOL MEDIATION

This section describes briefly the main components involved in our protocol mediation approach. A more detailed explanation is provided in (Williams et al., 2006), and Figure 2 shows an example of the use of all these components in the logistics domain described in the introduction.

Communicative Acts

Communicative acts are the basic components of the communication. They are modelled as sequences of four events that are exchanged between systems and the underlying communication infrastructure when sending a message (see Figure 1), as follows:

- `.request.` The initiator sends a message to the communication infrastructure.
- `.indication.` The responder receives the message from the communication infrastructure.
- `.response.` The responder acknowledges the receipt of the message.
- `.confirm.` The initiator receives the acknowledge receipt.

Both the .response and .confirm primitives model an acknowledgement that the communication has reached its intended recipient. Any substantive response motivated by the communicative act itself is modelled as a subsequent communicative act in the opposite direction.

At the initiator, the outcome of a communicative act may be a success (the initiator knows that the communication has reached the intended recipient), an exception or failure (the initiator knows that the communication has failed to reach the intended recipient), or indeterminate (the initiator does not know the outcome of the communication).

Abstract Protocols and Roles

When we described protocol mediation, we commented that systems involved in a message exchange have conceptually similar interaction protocols. This high-level conceptual protocol is described by means of an abstract protocol.

The abstract protocol can be then defined as a multi-party choreography that describes the constraints that govern the sequencing of communicative acts between the systems engaged in an interaction. Each system takes on one or more roles (e.g., buyer, seller, logistics provider, freight forwarder, etc.) with respect to a choreography. The choreography then describes each of these roles in terms of the sequencing constraints on the exchange of primitives between the communication infrastructure and the system adopting the role.

Concrete Protocols

Each of the systems involved in a message exchange may have different mechanics by which communicative acts are managed. For each communicative act event in each system we will have then a concrete protocol that describes this behaviour.

Hence concrete protocols describe what happens at an initiating system in response to an admissible .request primitive and prior to (and after) the corresponding .confirm primitive. Likewise, at a responding system in response to the stimuli that give rise to an .indication primitive, the behaviours that occur between that and the corresponding .response and the behaviours that occur after that.

Processes as Abstract State Machines

The abstract and concrete protocols are described by means of processes, which in our approach are implemented by concurrent finite state machines.

Figure 1. A communicative act and its four events (Williams et al., 2006)

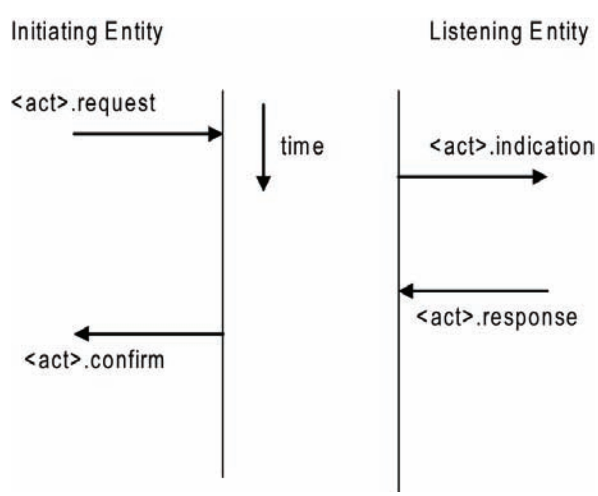
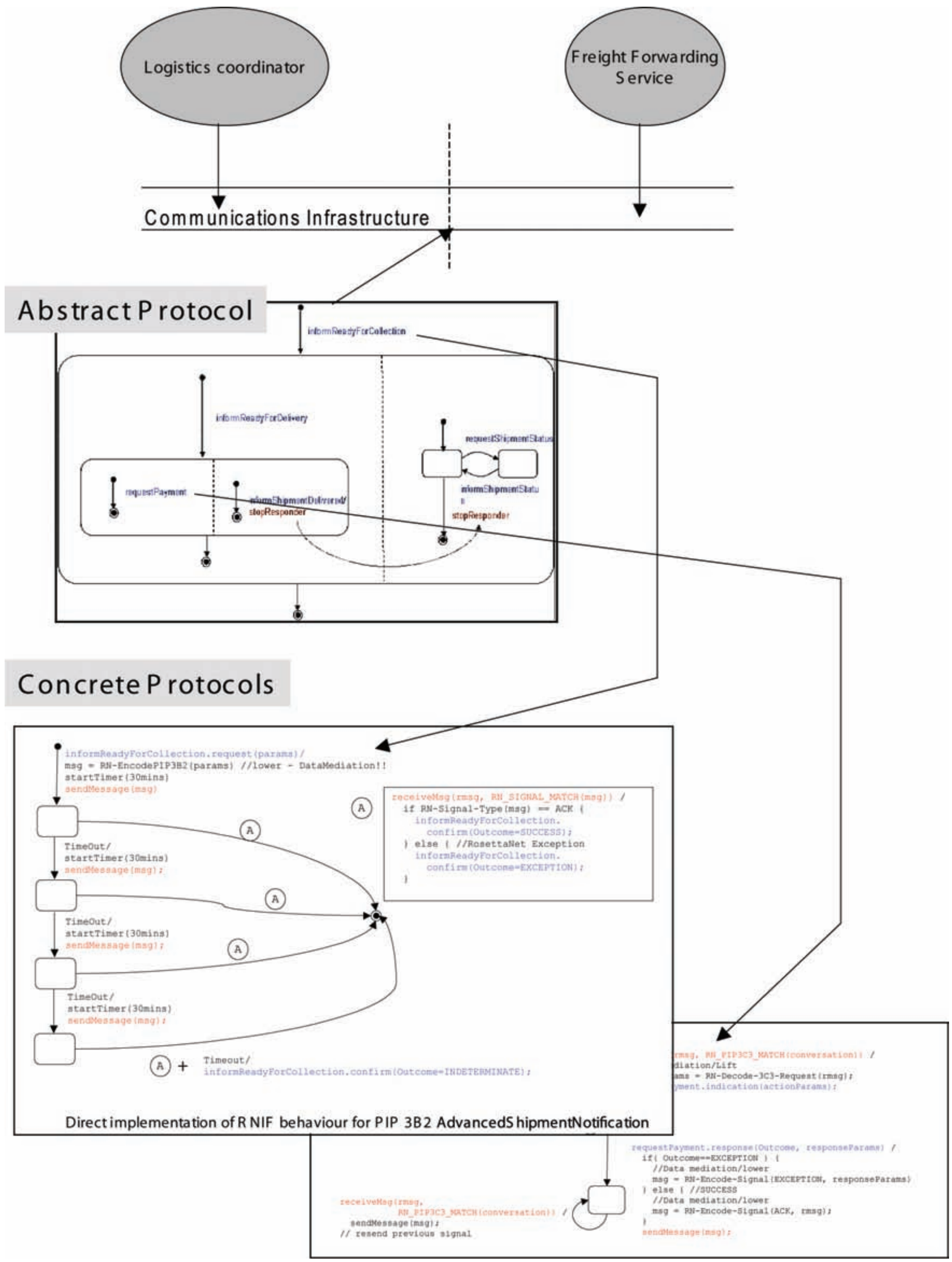


Figure 2. Abstract and some concrete protocols in the logistics domain (adapted from Williams et al., 2006)



For abstract protocols a state represents the state of the high-level conversation in the context of the common multi-party choreography (e.g., a request for payment has been issued by the freight forwarding service and received by the logistics coordinator). For concrete protocols a state represents some intermediate state in the behaviours associated with the issuing of .request and .confirm primitives or issuing .indication and .response primitives.Transitions between states may be driven by different internal and external actions, as follows:

1. *PrimitiveDriven transitions*. In abstract protocols they can be any of the primitives of a communicative act. In concrete protocols, they can be only <act>.request or <act>.response primitives, since these primitives can initiate the state machines associated to a concrete protocol.
2. *EventDriven transitions*. They are used to communicate between concurrent processes (a process may raise an event that is being waited for by another process). They are normally used in communication exchanges between more than 2 parties and in concrete protocols (e.g., two processes are waiting for the same payment under different payment procedures, credit card or cheque, and one of them is satisfied).
3. *TimeDriven transitions*. They occur on the expiry of a time interval following the entry to the state that has the time driven transition associated. They can be used in any type of protocol (e.g., in an abstract protocol, the system will have a timeout feature to send another communicative act if a response has not been received in a given time).
4. *MessageDriven transitions*. They occur only in concrete protocols, when a message is received from the communication infrastructure and filtered according to a template, so that the relevant information is extracted (e.g., for a freight forwarding service, if a request for a shipment service is broadcasted through the communication infrastructure, this could activate it so that it provides its service to the logistics provider).

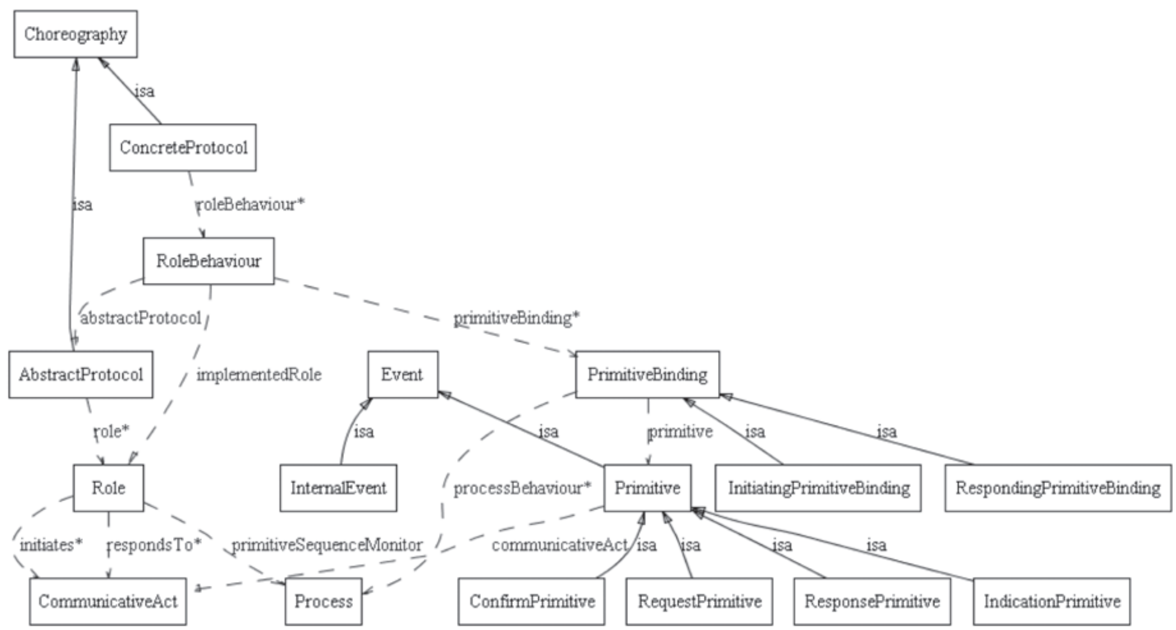
All the transitions have associated a transition condition guard (a Boolean expression that determines whether the transition can be actually performed given the state where the state machine is and the external and internal conditions) and a transition behaviour. Transition behaviours model the actual transition logic to be done besides moving from one state to another. They include (both for abstract and concrete protocols): raising .indication or .confirm primitives, raising events to concurrent processes, and instantiate concurrent processes. For concrete protocols they may also include: perform transformations on received message structures, generate message structures for transmission, and extract, maintain and manipulate information taken from message fields.

AN ONTOLOGY FOR DESCRIBING ABSTRACT AND CONCRETE PROTOCOLS

Figure 3 and Figure 4 show different parts of the very simple choreography language (VSCL) ontology, which is available at http://swws.semanticweb.org/. This ontology can be used to describe the abstract and concrete protocols presented in the previous section, together with all their components, and is used to configure the protocol mediation component described in the next section.

As shown in Figure 3, choreographies are divided into abstract and concrete protocols. An abstract protocol specifies a set of roles that identify the role that a system is playing in an exchange of messages (logistics coordinator, freight forwarding service, etc.). Each role contains a set of communicative acts that are considered in the shared abstract protocol and that allow defining the

Figure 3. Ontology excerpt related to abstract and concrete protocols and communicative acts



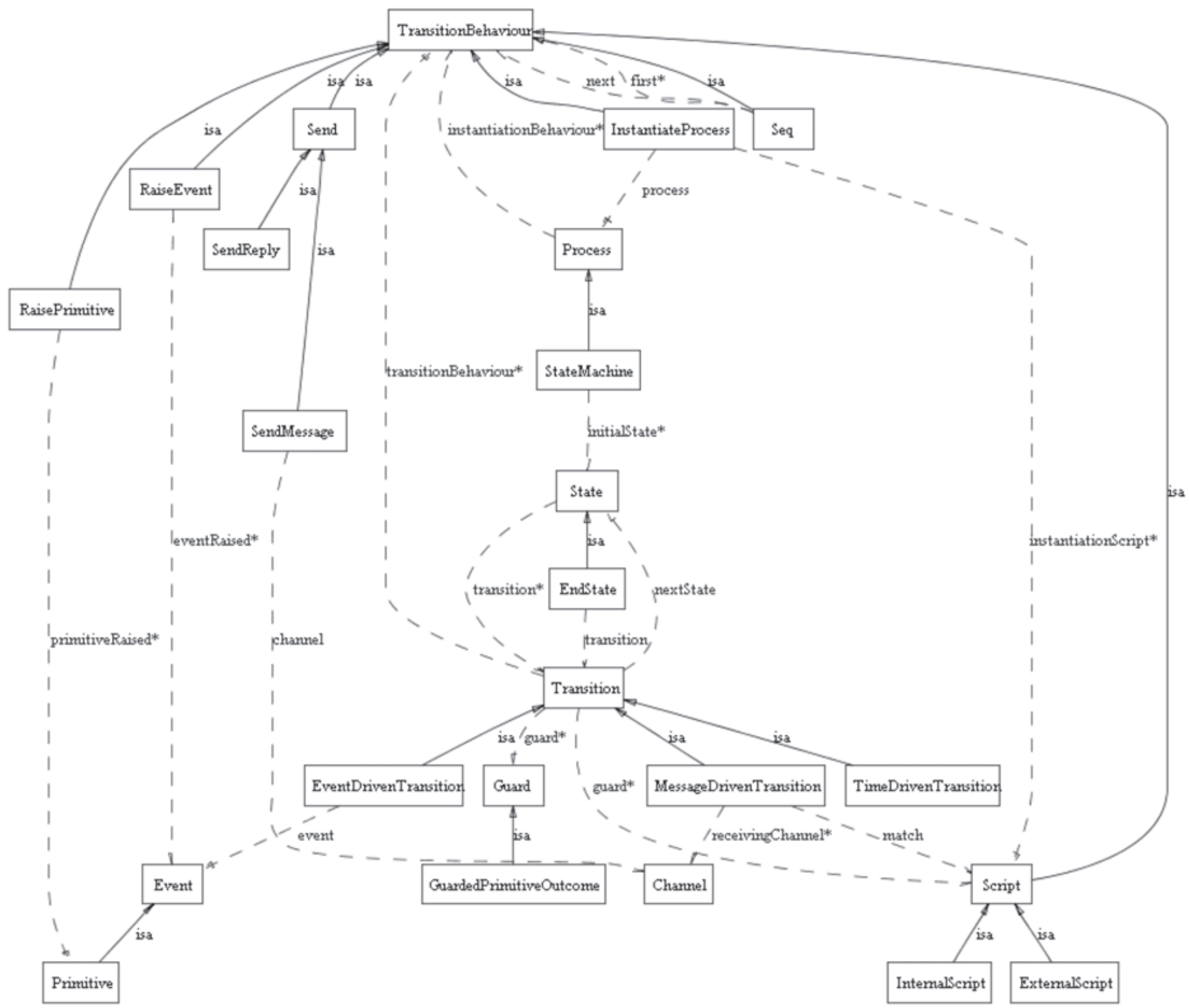
shared conceptual model of the message exchange patterns to be followed by all the systems participating in a conversation. For each of these roles in each abstract protocol and with each specific implementation of any of the systems involved there is one role behaviour, that implements a set of concrete protocols that correspond to the behaviour that the actual system for the different communicative acts that are defined for it.

The admissible sequences of communicative acts are specified in what we call a process, whose common implementation will be a state machine, as we will see in the next figure. The primitives that are considered are those that were described when we discussed communicative acts: request, indication, confirm and response.

Finally, each concrete protocol contains one or more instances of RoleBehaviour. Each instance of RoleBehaviour declare a role that may be adopted by a peer to interact with the service provider agent via its interface. Each RoleBehaviour and carries a PrimitiveBinding for each RequestPrimitive and IndicationPrimitive associated with the role. This divides PrimitiveBinding into two subclasses, InitiatingPrimitiveBinding for binding instances of RequestPrimitive and ListeningPrimitiveBinding for bindings associated with instances of IndicationPrimitive. Each instance of PrimitiveBinding associates an instance of Process with the corresponding primitive. The Process(es) associated with an InitiatingPrimitiveBinding are instantiated when an admissible invocation of the corresponding RequestPrimitive occurs. The Process(es) associated with a ListeningPrimitiveBinding are instantiated either when the corresponding conversation is instantiated or as the conversation progresses and the IndicationPrimitive associated with the binding becomes admissible.

Figure 4 illustrates the classes used to represent state machines in VSCL. A state machine is a type of process that is composed of a set of states (some of which can be end states). Each state can have a set of associated transitions, which specify the next state, a set of guards and a set of transition behaviours. Transitions can be of different types, as described in the previous section (event driven, time

Figure 4. Ontology excerpt related to the state machine descriptions



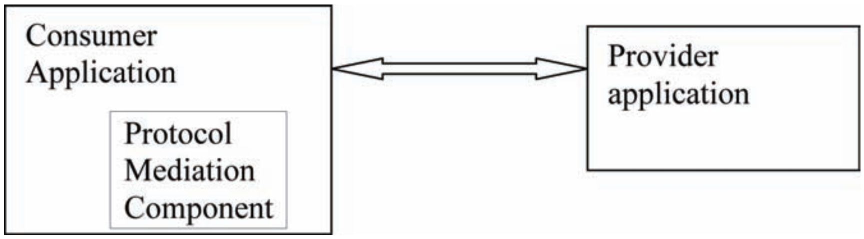
driven, or message driven). The primitive driven transitions were already specified in Figure 3 as initiating and responding primitive bindings, since they are responsible for starting a state machine.

Transitions behaviours are of different types, as pointed out in the previous section. From them, the most relevant is the script, which can be provided by a reference to a URL (external) or as part of the instance values (internal). We will analyse them in more detail later, when we discuss the component API.

In our logistics application we have a state machine for each of the protocols aforementioned. In summary, in our logistics application we have the following instances of this ontology (available at http://swws.semanticweb.org/):

- One abstract protocol with two roles defined for it: FreightForwardingServiceConsumer and FreightForwardingServiceProvider.
- 14 processes (state machines) for concrete protocols.
- Six communicative acts: InformReady-ForCollection, RequestShipmentStatus, InformShipmentStatus, InformReady-ToDeliver, InformShipmentDelivered, and

Figure 5. Location for the deployment of the protocol mediation component



RequestPayment, with their corresponding primitives (four for each of them).

- 10 event driven transitions with 20 scripts for their transition behaviours.

SOLUTION DETAILS: THE SWWS PROTOCOL MEDIATION COMPONENT

Here we provide a general description of the protocol mediation component architecture and of important implementation details, including a broad description of the component API, so that it can be used in other similar applications with protocol mediation needs.

Though the usual deployment of a protocol mediation component would be as part of the communication infrastructure between services in a service-oriented application, in our case this component has been deployed as shown in Figure 5: A consumer application incorporates the protocol mediation component inside its environment in order to control the exchange of messages with the provider application. In our logistics application, the selection of one system or another as consumer or provider is arbitrary. Our decision has been to use the logistics coordinator as a consumer and the freight forwarding service as a provider.The protocol mediation component has 5 essential subcomponents, which are described in detail in the next sections:

- **Local agent** (package com.isoco.swws.conversation.local_agent). It is the subcomponent directly used by the final user. Basically, the component allows creating conversations, initiating them in an active or a passive mode and later, by means of the ConversationManager, explicitly invoking the different CommunicativeActs and tracing the interactions with the remote conversation partner.
- **Protocol** (package com.isoco.swws.conversation.abstractprotocol). It is the internal representation of the protocols (either abstract or concrete) that rule the conversation. This is based on the ontology described in the previous section.
- **ChoreographyHandler** (package com.isoco.swws.conversation.mediation.vscl). It is the bridge between the application and the external choreography that is included in the VSCL ontology.
- **Message transfer plugin** (package com.isoco.swws.conversation.plugins). Internally, a specific communication protocol (HTTP, SMTP, etc.) is used for the communication between the consumer and the provider. This plugin serves as an interface for the protocol. This implementation of the component includes an HTTP plugin, but other plugins could be easily created and deployed.
- **Rhino facilites** (package com.isoco.swws.conversation.environment). They are used to execute the Javascript scripts included

in the choreography. The mechanism used in the component is Rhino (Mozilla) and there is an abstraction layer to ease its use and to adapt it to the application needs.

Local Agent

The local agent groups the collection of classes that the Consumer needs to create and control a conversation. A conversation is initiated with the creation of a *ConversationManager*, which receives the following parameters in its constructor:

- A set of roles (the systems involved in a conversation). The InterfaceRole contains the *remoteInterface*, the URL that holds the address of the conversation's partner, and the *localRole*, the URL of the role adopted by the local agent with respect to the choreography and this conversation.
- The URL where to find the choreography (that is, the place where the VSCL ontology instances are stored).
- An indication handler, which is used in the case that an external system has to contact this system or send it and event. Normally this handler is used when the system receives a message from the provider that causes a <CommunicativeAct>.indication. This is the way that the protocol mediation component has to inform an application that an indication has arrived. It is also responsibility of the IndicationHandler to respond to the indication of the CommunicativeAct. Responding to the .indication means to model the .response. The user must calculate the outcome and the results of that CommunicativeAct.

The implementation of the IndicationHandler is robust enough to deal with situations where it could be blocked or fail, where the response will be launched again.

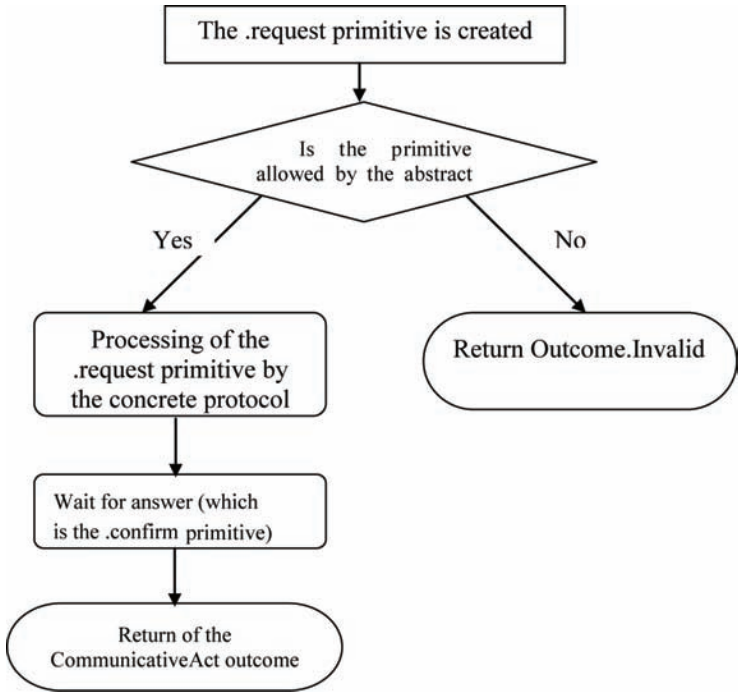
A conversation consists in the coordinated exchange of communicativeActs. The local agent can send CommunicativeActs either in a **synchronous** or an **asynchronous** way. In the synchronous exchange the communicative act is sent and the system waits for the confirmation of the remote partner. In the asynchronous exchange the communicative act is launched and the control is returned back to the application. When the confirmation from the user is received, the *confirm* method of the ConfirmHandler interface that has been specified as a parameter is invoked.

The creation of a new ConversationManager implies the following tasks: initializing the abstract and concrete protocols and initializing of the ChoreographyHandler for the successive uses of the choreography. A conversation that has been created can be initiated in two modes: **active** and **passive**:

- In the active mode, the Consumer can invoke the *synchSay* and the *asynchSay* methods (for synchronous and asynchronous exchanges of messages) to start the exchange of CommunicativeAct with the remote partner.
- In a passive mode, the *listen* method must be invoked to initiate a conversation in a passive mode. This action prevents the use of the *synchSay* and the *asynchSay* methods and the conversation waits for an indication from the remote partner. It should be noted that once the *listen* method is invoked, the conversation will only be activated by a remote message from the partner. There is no explicit method to transfer the conversation to the active mode.

Figure 6 shows how this works in an active mode: the .request primitive is created for a CommunicativeAct. This primitive is sent to the abstract protocol to know if the CommunicativeAct can be initiated in the current context of the conversation.

Figure 6. Usual process followed for a communicative act being sent



- If it cannot be initiated, the execution is aborted and Outcome.INVALID is returned to the entity to inform that it is impossible to execute that action in the current situation.
- If it can be initiated, the primitive is sent to the concrete protocol in order to execute the set of scripts and other relevant actions associated to this primitive. It is important to emphasize that the primitive is sent, that is, there is no explicit communication from the abstract protocol to the concrete protocol. The idea is that the abstract protocol allows executing the primitive but it does not consume it. Afterwards, we wait to receive the .confirm primitive and the Outcome associated to the CommunicativeAct of the primitive is returned. The outcome can be: Outcome. SUCCESS, Outcome.EXCEPTION, or Outcome.INDETERMINATE.

When the entity is waiting for an indication, the process is different. When a message arrives, it is evaluated in the MessageDrivenTransitions of the active processes of the concrete protocol. If any of them matches, that transition is executed and it will be its responsibility, among other responsibilities, to launch an .indication primitive to the abstract protocol to check if in the context of this conversation that primitive is allowed. If the primitive is allowed, the entity will be informed about it by the indication method of the IndicationHandler.

Multiple Conversations

The exchange of messages between the consumer and the provider is executed in a multiple simultaneous conversations scenario. To know which conversation should process each message, the protocol mediation component associates a unique conversation id to each ConversationManager.

Whenever a conversation is initiated by a partner, a message is sent with a parameter that informs that it is a new conversation. A new conversation id is created for this conversation and the following messages of this conversation must include this id.

The ConversationDispatcher class is responsible for registering all the existing conversations. Basically there are two lists: a list of active conversations and a list of potential conversations (those that are waiting to receive a message to start, by the invocation of the method *listen*). When a message to start a conversation arrives, all the conversations that are waiting are checked to inquire which one can process it. If a conversation can process it, that conversation is moved to the list of active conversations.

The ConversationDispatcher is also responsible for initializing all the available plugins once a conversation starts.

Protocols

Conversations are ruled by a choreography, which contains two types of protocols (abstract and concrete). Both protocols are specified by means of the ontology described in section 2.5. For each class in the ontology there is a Java class in this package, including the states and transitions.

Each ConversationManager has references to its own abstract and concrete protocols. When a conversation is created, the ConversationManager loads the initial processes with all their associated structure, using those Java classes (as explained in the following section). The list of processes and active states is updated as the transitions are executed.Transitions are modelled with the Transition Java class and its subclasses. The following methods are called for a transition:

- **Evaluate initTransition.** This function must be redefined by all the subclasses of Transition. It has two responsibilities: verify that the object that it receives is the instance that it knows how to process. For example, the EventDrivenTransition must guarantee that the object type is 'Event'. Not only must it guarantee that it has the correct type, but also that it is the instance that sets off the transition (for example, that it is the RequestShipmentStatus.request primitive). Its other responsibility is to initiate whatever is necessary to execute the transition. For example, to set some variable in the RhinoEnvironment or some other task.
- **Evaluate initGuard.** The transitions can have an associated guard that must be satisfied to continue with the transition. In general, it is a method that does not have to be redefined by the subclasses.
- **Execute doBehaviours.** As a consequence to the execution of the transition, a set of TransitionBehaviours must be executed. These behaviours represent what the transition does. This method should not be modified. As we will see, transition behaviours are specified in Javascript and executed by the RhinoEnvironment.
- **Execute advanceToNextState.** A change to the new state is performed in order to end the execution of a transition. This process entails several tasks such as the loading of all the structure of the new state from the choreography, the initialization of the associated TimeDrivenTransitions, etc.

Choreography Handler

It serves as a bridge between the application and the choreography. It is used to create instances of the classes included in the Protocols package from the choreography information available in a URL. As aforementioned, the whole choreography is not loaded completely from the start but incrementally according to the transitions done through the abstract and concrete protocols. Two significant methods from this class are:

- *createProcessByName*, which creates a state machine from the information available in its location (URL). It returns the state machine and all the structure associated to it (states, transitions, transition behaviours, scripts, etc.).
- *createStateByName*, which creates a state from its name (URI). It returns the state and all the structure associated to it (transitions, transition behaviours, scripts, etc.).

This component uses the KPOntology library[2] to navigate the RDF graph that models the choreography.

Message Transfer Plugin

This component deals with the specific communication protocol (HTTP, SMTP, etc.) used for the communication between consumers and providers. An HTTP plugin is included with the current implementation, and other plugins can be also created.

The HTTP plugin provided is made up of the HttpPlugin class and an auxiliary Web application that manages the queue of received messages, with two services:

- Receive a message. This service is used when a remote partner, e.g. the provider, must send a message to the consumer. The Web application receives the message and puts it in the queue of received messages.
- Recover the message. This service allows the HttpPlugin class to recover the messages received from the Web application.

The HttpPlugin class has two main objectives:

- Send messages to remote partners, using the sendMessage method. This method receives a remote address where to send the message, the conversation id, and the message.
- Transfer messages from the Web application to the component. The HTTP plugin has a thread that is constantly polling the Web application for the arrival of new messages.

The Web application always responds to the petition of messages by means of an XML that contains the following elements:

- **conversationId:** id of the conversation under way.
- **newConversation:** it indicates if it is a new conversation.
- **Message:** Depending on the types of message, it will have different types of structures. For instance, in the case of the RosettaNet messages, it will be divided into: “Preamble”, “DeliveryHeader”, “ServiceHeader” and “Content”.

It is the responsibility of the plugin to find the appropriate Conversation Manager from the conversation id, to build the internal structure of the protocol for the representation of the messages and to send the resulting message to the Conversation Manager for its processing.

Messages and Filters

All messages are vectors of XML structures, so that they can accommodate multi-part messages that are typical in B2B interactions. The underlying messaging system plugins are responsible for encoding/decoding between the internal XML structures (typically XML DOMs or more abstractly XML Infosets) and the packaged and encoded wire format - this includes XML validation of inbound messages against the relevant DTDs and/or XML schema. Directly or indirectly the concrete interface descriptions MUST provide message DTD/

Schema and lift/lower transformations.

In addition, received message structures also carry low-level connection and endpoint information. Typically this will not be used directly in processing the message, but is essential for the plugins to correctly formulate a response message - in particular if a response/reply needs to be returned on the same connection as a given received message.

Message are filtered and classified according to the various pieces of typing information that they carry: internet media type, XML DOCTYPE and XML root element type of the primary part of the message; and identification of the endpoint via which they were received. This associates a received message with a collection of processes which serve messages of a given kind. Concrete Role behaviour descriptions contain a static description of the message types they are able to receive.

Messages with the same conversation id are bound to a particular conversation and queued to be processed by the concrete role behaviours associated with that process - in particular messages are consumed by message driven transitions. When a message matches a message filter in the initial transition of a listening role behaviour, a factory behaviour is invoked which instantiates a new instance of a conversation (controller) and passes that new message to that controller - a new conversation id value becomes associated with the new conversation.

So coarse filtering is used to associate messages with a class of conversational role where they may either be queued at an existing conversation or used to instantiate a new conversation. Messages queued at a conversation are then visible to the processes that realise the concrete role behaviours for that conversation. As discussed earlier these may or may not be processed in strict arrival order.

Message Filtering

This component eases the use of Rhino, the Javascript interpreter used by the protocol mediation component to express message filters, transition pre-conditions and some (scripted) transition behaviours. Each process specified in the choreography has a Rhino Environment, and each environment will have a defined scope. This scope has a set of variables and functions defined in the scripts. In this way, the processes do not share the execution environment when they execute the scripts.

The abstraction layer of Rhino is achieved through the RhinoEnvironment class. The most distinguishable of its functions are:

- *execute*, which receives a script as a parameter and executes it.
- *match*, which receives a script that returns a Boolean value, executes it and returns that Boolean value.
- *setMessage*, which receives a variable name and its value, and is in charge of creating in the Javascript environment a variable with that value.
- *getMessage*, which returns the value of a variable name in the Javascript environment.

Deployment and Installation

The protocol mediation is a framework designed to be used by a client application. The typical scheme for its use would be:

- Initialize the ConversationDispatcher.
- Create a ConversationManager, specifying the choreography, the participating agents and the Indicationhandler. The implementation of the IndicationHandler must guarantee that all the possible .indication communicative acts that the remote partner can send are processed and for each one

of them, it must compute the Outcome and the adequate results.

- Initiate the exchange of CommunicativeActs with the remote partner.

Next, we show an example on how to use the component in Box 1. The objective of this example is to give a guidance on the use of the component. The typical use must be by means of an application that should keep the evolution of the conversation as well as the CommunicativeActs that have been sent and received by the remote partners.

The first thing to do is the initialization of the ConversationDispatcher. This initialization also includes the initialization of the plugins. In the previous example, the URL is the address of the local Web application that uses the HTTP-Plugin.

The second thing to do is the creation of the ConversationManager. In the previous example we talk to the partner that we can reach at "http://provider:8080/". In the conversation we adopt the role of the FreightForwardingServiceConsumer. The choreography is found in http://swws.semanticweb.org/logistics.owl. We also have the IndicationHandlerImpl which is an implementation of the IndicationHandler.

Afterwards, a CommunicativeAct is created (in this case: InformReadyForCollection) and we send it in a synchronous way.

To keep the example simple we do not send any parameter in the comunicativeAct, but it would be usual practice.

ALTERNATIVES, COST AND BENEFITS

The proposed solution to protocol mediation between heterogeneous applications can be applied not only to the logistics domain, which is the one that has been described in this chapter, but also to other similar domains where applications are already deployed and have to interoperate with each other in order to support a specific set of added-value functionalities.

While work on the area of data mediation in service exchanges is quite widespread and there are tools available in the mainstream market for solving these issues, most of the approaches for protocol mediation have been based on ad-hoc

Box 1. Partial OCML code defining location parameter instance and respective MSS member.

```
(def-instance brighton-location location
    ((has-instance-title "Brighton")
    (defined-by p2-location-brighton)))

String logisticsNamespace = "http://swws.semanticweb.org/logistics#"
ConversationDispatcher.init("http://consumer:8080/");
interfaceRole = new InterfaceRole(new URI("http://provider:8080/"),
    new URI(logisticsNamespace + "FreightForwardServiceConsumer"));
IndicationHandlerImpl indct = new IndicationHandlerImpl();
ConversationManager conversationManager = new ConversationManager(
    new InterfaceRole[]{interfaceRole},
      new URI("http://swws.semanticweb.org/logistics.owl"), indct);
CommunicativeAct communicativeAct = new CommunicativeAct(new URI(logisticsNamespace +
"InformReadyForCollection"));
conversationManager.synchSay(communicativeAct);
communicativeAct = new CommunicativeAct(new URI(logisticsNamespace + "RequestShipment-
Status"));
conversationManager.synchSay(communicativeAct);
```

solutions that are tightly related to the applications where they are being applied. No easy configurable toolkit exists yet for solving this problem, hence the main alternative for the work proposed here is to create an ad-hoc solution that solves the interaction problem between applications or services for a specific set of functionalities.

Though our approach still requires a lot of effort to be done, and requires more maturity and further evaluations to be applied in production systems, the main advantages with respect to the current state of the art are related to the reusability of the abstract representations of message exchanges for each of the systems involved, as well as the reusability of message filters across different types of applications, what can benefit the agility of developing new added-value applications in the future. Besides, the model is easily extensible and fully declarative, what influences in the lowering of maintenance costs.

CONCLUSION AND FUTURE TRENDS

In this chapter we have motivated the need to use some form of protocol mediation to make it possible to different systems in the logistics domain to communicate successfully with each other, even if they use different protocols (RosettaNet and EDIFACT). Furthermore, we have described the approach for protocol mediation developed in the context of the SWWS project, including the ontology used to describe the choreography (that is, how the systems interact with each other) and the software that implements the component that has been developed.

Though this is a first approach to solve the protocol mediation problem between systems, there is still much work to be done in the future to convert this prototype into a production-quality component. Among them, we have to add new message transfer plugins to allow message transfer using other communication protocols, such as SMTP, FTP, etc., which is what it is used by many of the current systems. Besides, a tighter integration and evaluation with existing systems has to be provided, and a library of common interaction patterns should be also implemented, so that the task of protocol mediation is as simple as possible for those developers that want to develop a mediation solution for their systems.

ACKNOWLEDGMENT

This work was supported by the EU under the SWWS and DIP consortia. In addition we want to thank Stuart Williams, for his ideas that have been the basis for this work and for providing good comments to improve this chapter, to Jorge Pérez Bolaño for making part of the component implementation, to Juan Miguel Gómez for the work on the VSCL ontology, and the other members of the consortium, who contributed to the use case and to the ideas presented here.

REFERENCES

AnsiX12 (n.d.). *National Standards Institute Accredited Standards Committee X12*.

Cimpian, E., & Mocan, A. (2005). *Process Mediation in WSMX*. WSMO Working Draft D13.7 v0.1.

EDIFACT. ISO 9735. (2002). *Electronic data interchange for administration, commerce and transport (EDIFACT) – Application level syntax rules*. International Standards Organisation.

Evans-Greenwood P, Stason M (2006). Moving Beyond Composite Applications to the Next Generation of Application Development: Automating Exception-Rich Business Processes. *Business Integration Journal*, May/June 2006.

Preist, C., Esplugas-Cuadrado, J., Battle, S. A., Grimm, S., & Williams, S. K. (2005). Automated Business-to-Business Integration of a Logistics Supply Chain Using Semantic Web Services Technology. In Gil et al. (eds), *Proceedings of the 4th International Semantic Web Conference (ISWC2005)*. Lecture Notes in Computer Science, Volume 3729, Oct 2005, Pages 987-1001

RosettaNet Implementation Framework: Core Specification Version 2.00.01. March 2002.

Williams, S. K., Battle, S. A., & Esplugas-Cuadrado, J. (2006). Protocol Mediation for Adaptation in Semantic Web Services. In Domingue and Sure (eds), *Proceedings of the 3rd European Semantic Web Conference (ESWC2006)*. Lecture Notes in Computer Science, Volume 4011, June 2006, Pages 635-649.

ADDITIONAL READING

We recommend reading WSMO deliverables about mediation, in general, and about process mediation in particular. They can be found at http://www.wsmo.org/. Efforts on process mediation are also being done in the context of the SUPER project (http://www.ip-super.org/).

ENDNOTES

1 http://swws.semanticweb.org/
2 http://kpontology.sourceforge.net

This work was previously published in Semantic Web for Business: Cases and Applications, edited by R. García, pp. 65-82, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 7.5
Probabilistic Models for the Semantic Web:
A Survey

Livia Predoiu
University of Mannheim, Germany

Heiner Stuckenschmidt
University of Mannheim, Germany

ABSTRACT

Recently, there has been an increasing interest in formalisms for representing uncertain information on the Semantic Web. This interest is triggered by the observation that knowledge on the web is not always crisp and we have to be able to deal with incomplete, inconsistent and vague information. The treatment of this kind of information requires new approaches for knowledge representation and reasoning on the web as existing Semantic Web languages are based on classical logic which is known to be inadequate for representing uncertainty in many cases. While different general approaches for extending Semantic Web languages with the ability to represent uncertainty are explored, we focus our attention on probabilistic approaches. We survey existing proposals for extending semantic web languages or formalisms underlying Semantic Web languages in terms of their expressive power, reasoning capabilities as well as their suitability for supporting typical tasks associated with the Semantic Web.

DOI: 10.4018/978-1-60566-028-8.ch005

INTRODUCTION

The Semantic Web is an extension of the World Wide Web that allows for expressing the semantics and not only the markup of data. By means of the representation of the semantics of data, new and not explicitly stated information can be derived by means of reasoners. In this way, software agents can use and integrate information automatically. As common web languages like (X)HTML and XML are not enough for this purpose (Decker et al., 2000), Semantic Web languages have been standardised (RDF, RDF Schema and OWL), proposed (e.g. WRL, SWRL) and new ones are still being devised. However, most languages that are intended for usage on the Semantic Web are deterministic and cannot represent uncertainty. Currently, there is a growing interest in probabilistic extensions of Semantic Web languages. People start to realize that there is inherently probabilistic knowledge that needs to be represented on the Semantic Web. In the following, we briefly describe five areas where probabilistic information plays a role in the context of the Semantic Web.

Representing inherently uncertain Information: Not all of the information that needs to be represented on the Semantic Web is given in terms of definite statements. E.g. statistical information can provide insights to data to be shared on the Semantic Web. Ontological information attached with statistical values like the percentage of people in a population that are of a certain age can help answer queries about the correlation between this age and a certain chronic disease. There are many situations in which the use of this statistical information could be used to improve the behaviour of intelligent systems. An example would be a recommender System that points the user to certain information based on information about the age group.

Ontology Learning: The manual creation of ontologies has been identified as one of the main bottlenecks on the Semantic Web. In order to overcome this problem several researchers are investigating methods for automatically learning ontologies from texts. Existing approach normally use a combination of NLP and text mining techniques (Maedche & Staab, 2004). Typical tasks are the detection of synonyms and of subclass relations using clustering techniques and association rule mining. In both fields, the result of the mining process can be interpreted in terms of a probabilistic judgement of the correctness of the learned relation.

Document Classification: Document Classification can be seen as a special case of ontology learning called Ontology population. Today a major part of the information on the web is present in terms of documents (Web Pages, PDF Documents etc.). A common way of linking documents to knowledge encoded in ontologies is to assign individual documents to one or more concepts representing its content. Different machine learning techniques have been applied to this problem (Sebastiani, 2002). The most commonly used is the use of naïve Bayes classifiers that estimate the probability of a document belonging to a topic based on the occurrence of terms in sample documents.

Ontology Matching: Different sources often use different ontologies to organize their information. In the case of documents, these are often classified according to different topic hierarchies. In order to be able to access information across these different sources, semantic correspondences between the classes in the corresponding ontologies have to be determined and encoded in mappings that can be used to access information across the sources. Recently, a number of approaches for automatically determining such mappings have been proposed (Euzenat & Shvaiko, 2007). Some of the most successful ones use machine learning techniques to compute the probability that two classes represent the same information.

Ontology Mapping Usage for Information Integration: The usage of the mappings that have been found by matchers as explained in the paragraph above is currently mainly deterministic. Although the mappings are attached with a confidence that expresses how sure the matcher is that the mapping holds, the usage of those mappings consists of a preprocessing step: All mappings that have a confidence value above that threshold are considered deterministically true and all mappings that have a confidence value below that threshold are considered deterministically false. However, there is evidence that this kind of usage is error prone, especially when mappings are composed over several ontologies.

The five examples above clearly demonstrate the importance of probabilistic information in the context of the Semantic Web. E.g., in order to use the learned structures effectively, we need ways to represent and reason about the probabilities assigned to them. Most existing Semantic Web languages are mainly based on classical, deterministic logic and do not support this aspect. In the following, we review a number of general approaches for combining logical languages with probabilistic models and discuss existing proposals for extending semantic web languages with probabilistic information in more details.

Aim and Scope

In this paper, we review existing proposals to extend Semantic Web languages with the capability to handle uncertain information to better deal with the situations mentioned above. There are many ways of representing and dealing with uncertainty. In this paper, we restrict our attention to approaches that use probabilistic methods for representing uncertain information. In particular, we will not cover recent proposals for fuzzy-logic based extensions of semantic web languages. We will also not discuss nonmonotonic and non-standard logics for representing uncertainty unless they are based on a probabilistic semantics. We focus on these approaches, because we believe that probabilistic methods are a natural choice for representing the kinds of uncertainty we often find on the web. A strong motivation is the awareness that Semantic Web technology could greatly benefit from a tighter integration with machine learning and information retrieval techniques which are mostly based on probabilistic models. Probabilities have been criticised mostly due to the fact that people are very bad in providing correct judgements of the probability of events. We think that on the Semantic Web, this argument does not apply, because the aim here is not to use subjective judgements of probability but to provide mechanisms to represent inherently statistical information found on the web or produced by machine learning and matchers. The five examples above clearly demonstrate the importance of probabilistic information in the context of the Semantic Web. In this paper, we review a number of proposals for extending logical languages with probabilistic information in more details. We focus on

1. approaches that directly extend Semantic Web languages, in particular RDF and OWL
2. approaches that extend formalisms that have a very close connection to Semantic Web languages or that have explicitly designed to be used on the Semantic Web by the authors.

In the latter category, on the one hand, we cover probabilistic extensions of Description Logics which are commonly accepted as being the formal basis of OWL. Even though most approaches only cover logics that are much weaker than OWL, the methods proposed can directly be applied to the corresponding subset of OWL. The second kind of languages we consider are rule languages. Although there is not yet an official rule language for the Semantic Web, it is clear, however that rule languages have an important role to play on the Semantic Web. As the area of rule languages is also very broad, we focussed on approaches that have been developed for the Semantic Web. Due to the fact that ontologies and thus Description Logics play a very important role in the Semantic Web, all approaches that extend are that combine rules and ontologies in some way. We restrict ourselves to probabilistic logics that allow combinations of rule and ontologies also with our application example in the area of ontology matching and ontology mapping usage in mind. This application example is presented below.

When talking about the different approaches, we will distinguish between the logical language which is used to describe knowledge and the probabilistic model used to assign probabilities to certain assertions of the logical language. Based on this distinction, we discuss the following issues of the different approaches:

- the general probabilistic model used
- Expressiveness of the logical language
- kind of logical sentences that can be assigned a probability
- reasoning support and expected efficiency for large scale models

In order to evaluate the applicability of the respective approaches, we also consider an example scenario from the area of ontology matching and ontology mapping usage. This example illustrates also the inherent uncertainty of mappings and why this uncertainty needs to be taken into account for reasoning. Our example is based on two ontologies used in the Ontology Alignment Evaluation Challenge[1]. Assume a situation where a user is looking for publications about AI based on two ontologies O_1 and O_2.

Let O_1 be specified by the following axiom which specify that for each publication there is a keyword which is a subject. Furthermore, there is a publication about the Semantic Web which has the keyword Artificial Intelligence.

1. Publication v 8keyword.Subject
2. (SW, AI): keyword
3. SW: Publication

Let O_2 be specified by the following axioms which specify that reports are always publications and every concept in the knowledge base is about some topic. Furthermore, there is one report about Logic Programming and a publication about Description Logics. Both are about Logics.

4. Report v Publication
5. > v 8about.Topic
6. BN:Report
7. DL: Publication
8. (BN, Probability): about
9. (DL, Logics): about

Without loss of generality, we can assume that O_1 is the local ontology, i.e. the ontology being queried explicitly by the user. In order to integrate the information which is stored in both ontologies, mappings are needed. With a probabilistic matcher like GLUE (Doan et al., 2003) mappings can be found which map the second ontology O_2 to our local ontology O_1.

10. O1: Publication(x) ← O2: Publication(x) with probability 0.8
11. O1: Publication(x) ← O2: Report(x) with probability 0.9
12. O1: Subject(x) ← O2: Topic(x) with probability 0.9
13. O1: keyword(x, y) ← O2: about(x, y) with probability 0.8

The mapping (10) basically says that all instances that are belonging to the concept Publication in O_2 are also belonging to the concept publication of O_1 with the probability 0,9. Due to the Kolmogorov axioms of probability theory, the probability that instances belonging to O_2:Publication do not belong to O_1:Publication is 0,1. For completeness, the probability instances that do not belong to O_2:Publication belong to O_1:Publication need to be derived by a matcher as well. GLUE can be modified such that it conforms to this requirement. Let's assume that for (10) this probability is 0.2, for (11) it is 0.4, for (12) it is 0.3 and for (13) it is 0.4.

If we pose a query, we want to get an answer that integrates the information of both ontologies. So, in our example, if we query our local ontology for all publications:

Publication(x) $\wedge$ keyword(x, AI)we want to get also all relevant publications mentioned only in the second ontology. The answer here is

- the publication about Semantic Web with probability 1.0 because it is mentioned in the local ontology and no mapping has been used for deriving it
- the publication about Logic Programming in the second ontology which was derived

by two mappings (10) and (11) and thus gets an accumulated probability of 0,75
- the publication about Description Logics which was derived by only one mapping (10) and has only the probability 0,44

The computation is based on the semantics of the Bayesian Description Logic Programming formalism and shows nicely the importance of the consideration of uncertainty in the area of information integration in the Semantic Web. Without the consideration of the uncertainty each mapping is associated with, all 3 answers would be needed to be treated in the same way although the publication about the Semantic Web is much more relevant than the one about Description Logics. Furthermore, if mapping composition is considered with mapping chains over several ontologies, mappings with rather low probabilities can contribute to an answer with a rather high probability.

Another requirement for a mapping language is the possibility to express mappings between individuals. E.g. in our example, publications of O_2 that are about probabilities are less probable to be publications in O_1 that deal with AI than publications of O1 that are about logics:

14. O1:keyword(x, AI) ← O2: about(x, Probability) e.g. with probability 0.7
15. O1:keyword(x, AI) ← O2: about(x, Logics) e.g. with probability 0.9

It is immediately clear that such mappings need to be expressed for a comprehensive handling of mappings in the Semantic Web area. Mappings that do not involve any variables might also be necessary to be expressed in certain scenarios. When we investigate the different probabilistic extensions of Semantic Web languages, we also have a look at the applicability of the formalisms for the area of Information Integration as presented in this example.

This chapter is structured as follows. In the next section, we present an overview of current Semantic Web languages and related formalisms that are the basis for the logical languages used in the different approaches discussed later in the paper. We also provide a brief introduction to some basic probabilistic models that are used in the different approaches. Based on these basic methods, we discuss proposals for probabilistic languages for the Semantic Web, in the section "probabilistic extensions of Semantic Web languages" below in this chapter. We start with proposals for extending RDF and OWL. Afterwards, we discuss approaches for extending related formalisms with notions of probability, namely Description Logics and different Rule Languages. We conclude the chapter with a critical review of the state of the art and an analysis of directions for future research.

PRELIMINARIES AND BACKGROUND

In this section, we introduce the reader to the state of the art in current Semantic Web languages and the background on the probabilistic models used in the probabilistic extensions surveyed below in the section "probabilistic extensions of Semantic Web languages" below in this chapter.

Current Semantic Web Languages

So far, the development of languages for the Semantic Web was dominated by traditional views on metadata models and logic-based knowledge representation. The major languages that have been developed are RDF/RDF Schema (Lassila & Swick, 1999; Manola & Miller, 2004) for representing metadata and the Web Ontology language OWL (Bechhofer et al., 2004) for representing terminological knowledge in terms of ontologies. The Web Ontology language OWL has its root in the formalism of Description Logics, a decidable subset of first-order logic that contains special constructs for defining classes in terms of necessary and sufficient conditions based on predicates representing binary relations between instances of different classes. More specifically, OWL corresponds to particular Description Logic variants (OWL Lite corresponds to SHIF(D) and OWL DL corresponds to SHOIN(D) (Horrocks et al., 2003)) in the sense that reasoning in OWL can be reduced to checking satisfiability in this logic (Horrocks & Patel-Schneider, 2004). Similarly, the semantics of RDF can be modelled with First-Order Logics, Description Logics and Datalog (Fikes & Guiness, 2001), (de Bruijn & Heymans, 2007).

Recently the need for rule languages on the Semantic Web has been recognized. Rule languages complement Description Logics as they allow to represent kinds of axioms not expressible in SHIF and SHOIN (e.g. property chaining (cf. e.g. (Horrocks, 2005)). Thus, several rule language proposals for the Semantic Web have emerged, examples being the Semantic Web Rule Language SWRL (Horrocks et al., 2005) and the Web Rule Language WRL (Angele et al., 2005) for describing domain-dependent inference rules. The Semantic Web Rule language allows the definition of conjunctive rules over the concepts and binary relations or roles, respectively, which are contained in an OWL ontology (Horrocks et al., 2005). Finally, similar to OWL, WRL is a layered language consisting of three languages, one being a superset of the other. WRL-Core which is the least subset of the WRL language family corresponds to a subset of OWL which lies in the language of Logic Programming (also known as the DLP fragment (Grosof et al., 2003)). WRL-Flight contains WRL-Core and is a Datalog-based rule language. WRL-Full contains WRL-Flight and is a rule language with function symbols and negation under the Well-Founded Semantics (Angele et al., 2005).

Description Logics which is represented by OWL in the Semantic Web and Logic Programming which is represented by a couple of W3C rule language proposals have both nice orthogonal properties and expressivity. Ways for combing both have been and still are investigated. Description Logics and Logic Programming have been found to have a common subset called Description Logic Programs (Grosof et al., 2003). Therefore, Description Logic Programs have a Logic Programming and a Description Logic syntax and wrappers can be used to translate them. Another subset of Description Logics and Logic Programming has been recently proposed that is called Horn-SHIQ (Hustadt et al., 2005) and is a strict superset of Description Logic Programs. Besides the investigation of the intersection of Description Logics and Logic Programming, a lot of research aims at a more comprehensive integration of both formalisms. Several approaches for enabling an interaction between logic programs and description logics exist. Usually, they consist of a Description Logics knowledge base and a Logic Program and the latter is equipped with special features for interacting with the Description Logics knowledge base. An example of such an approach where the Description Logic knowledge is an OWL Lite or OWL DL knowledge base is the formalism of Description Logic Programs under the answer set semantics by (Eiter et. al, 2004).

All of the languages mentioned above are logical languages with a classic model-theoretic semantics that makes a statement either true or false and have no means to represent uncertainty in any way.

Probabilistic Languages and Models

In the following, a short overview of the probabilistic models used for the languages in the sections below is presented. Those models are *Bayesian Networks*, *Bayesian Logic Programs*, *Independent Choice Logic*, *Probabilistic Datalog* and *Multi-Entity Bayesian Networks*. Some of these models are related to each other, e.g. Bayesian Networks can be considered as a subset of Bayesian Logic Programs because the latter provide a compact representation of the former in the same way like first-order logic does with sentential logic. Independent Choice Logic is a generalization and a superset of the formalism of Bayesian Logic Programs. The relationship of Bayesian Networks, Bayesian Logic Programs and Independent choice Logic with probabilistic Datalog is unclear. Multi Entity Bayesian Networks are comprising Bayesian Networks in the same way like Bayesian Logic Programs do. Multi Entity Bayesian Networks are more expressive than Bayesian Logic Programs, but it is unclear whether there is a semantical subset relationship. The relationship between Multi Entity Relationship Programs and Independent Choice Logic has not been investigated yet either. Multi Entity Relationship Programs differ from Probabilistic Datalog by the usage of negation. Probabilistic Datalog uses well-founded negation and the closed world assumption while Multi Entity Relationship Programs model probabilistic First-Order Logic knowledge bases and employ classical negation as well as the open world assumption.

Bayesian Networks (BNs) - One of the best understood models for representing the joint probability distribution of a domain of interest is the model of Bayesian Networks (BNs) (Jensen, 2001). A BN is a compact representation of the joint probability distribution among a set of random variables in a domain under consideration. More precisely, a BN is a directed, acyclic graph with the random variables as nodes and direct influence relationships as arcs. Several exact or approximate algorithms for reasoning in Bayesian Networks exist. Exact inference has been proven to be NP-complete in the maximal number of parents of nodes in the network. A considerable amount of research effort has been spent on different issues like learning of the conditional probability tables/distributions of the nodes in the BN, learning the structure of a BN, etc. (Castillo et al., 1997; Cowell et al., 1999; Jensen, 2001). A BN has been found to correspond to a probabilistic extension of sentential definite clauses.

In the area of the Semantic Web where the same or similar knowledge can happen to be represented on different and independent peers and integrated reasoning and information usage requires mappings, cycles in the complete representation may occur. Unfortunately, BNs are not allowed to have directed cycles. For reasoning with BNs, a huge amount of free and commercial software tools and implementations exist.

Bayesian Logic Programs (BLPs) - Bayesian Logic Programs (Kersting & De Raedt, 2001) are an extension of Bayesian Networks to first-order definite clause logic and a probabilistic extension of definite first-order logic at the same time. A BLP consists of a set of rules and facts, i.e. a definite clause logic program. Each fact is associated with an a-priori probability and each rule with a conditional probability where the probability of the head atom is conditioned on the states of the body atoms. Each ground atom of the Herbrand Model of the definite clause logic program corresponds to a node in a corresponding Bayesian Network. The arcs are defined through the rules. For each valid ground rule, an arc from each node representing a body atom to the node representing the head atom exists in the corresponding Bayesian Network. Additionally, combining rules are defined in order to enable the combination of conditional probabilities of different valid ground rules with the same head atom. BLPs are defined to be acyclic. Therefore, the corresponding Bayesian Networks are acyclic as well. Reasoning with BLPs corresponds to deriving the Herbrand Model or the part of it which is relevant to the query and building the corresponding Bayesian Network. For BLPs, no complexity results have been published, yet. Currently, only one tool for reasoning with BLPs exists: the Balios engine (Kersting & Dick, 2004).

Independent Choice Logic (ICL) - Independent Choice Logic (Poole, 1997) is a logic that is built upon a given base logic that conforms to some restrictions and determines truth in the possible worlds defined by choice spaces. Possible worlds are built by choosing propositions from sets of independent choice alternatives. As base logic, Poole suggests acyclic logic programs under the stable model semantics. However, as we will see later in the subsections on probabilistic (disjunctive) description logic programs below, the approach works for other base logics as well.

An independent choice logic theory on a base logic is a pair (C, F) where C is a so-called choice space and F is a knowledge base in the base logic. C is a set of sets of ground atomic formulae from the language of the base logic such that for two choices $c_1, c_2 \in$ C, if $c_1 \neq c_2$ then $c_1 \cap c_2 = \emptyset$. The elements of C are called alternatives and are basically random variables. The elements of an alternative c are called atomic choices and are basically possible values for the random variable c. The semantics of ICL is defined in terms of possible worlds. A possible world corresponds to the selection of one element from each alternative. Such a selection is called total choice. The atoms that follow using the consequence relation of the base logic from these selected atoms together with the knowledge base of the base logic are true in this possible world. Reasoners for ICL are conceivable but depend on the base logic used. Also, the complexity for deciding consistency and query answering depends on the base logic used.

Multi-Entity Bayesian Networks (MEBNs) - Multi-entity Bayesian Networks (Laskey & Costa, 2005) extend the Bayesian Network model to full First-Order logic. In this way, graphical models with repeated sub-structures can be represented and a probability distribution over models of any consistent, finitely axiomatizable first-order theory can be expressed. With MEBN logic, entities that have attributes and are related to other entities can be represented. Features of entities and relationships among entities are random variables. The knowledge about attributes and relationships is expressed as a collection of MEBN fragments (MFrags) organized into MEBN theories (MTheories). An MFrag represents a conditional probability distribution and an MTheory is a set of MFrags that collectively satisfies consistency constraints ensuring the existence of a unique joint probability distribution over instances of the random variables represented in the MTheory. Possible queries are queries for the degree of belief in specific random variables given evidence random variables. The response to a query is computed by constructing a so-called situation-specific Bayesian Network that can be processed by a usual tool for Bayesian Networks. We are not aware of the existence of general complexity results for reasoning with the MEBN formalism. There are proposals for reasoning algorithms (Laskey, 2006) but no direct implementation of a reasoner for MEBN logic. But there is a translation of a subset of the MEBN formalism into probabilistic relational models implemented in the Quiddity*Suite (cf. http://www.iet.com/quiddity).

Probabilistic Datalog (pDatalog) - Probabilistic Datalog (Fuhr, 2000) is Datalog where each fact and each rule is extended with a probability which states the certainty of it being true. An important underlying assumption is that each element of the probabilistic Datalog program (i.e. every fact and every rule) is probabilistically independent from the other elements. Probabilistic Datalog has been equipped with a well-founded semantics. According to Nottelmann (2005), the probability of a rule can be seen as a conditional probability like with Bayesian Logic Programs. However, while Bayesian Logic Programs allow an arbitrary set of states for the ground atoms in the Herbrand Base, probabilistic Datalog envisions just Boolean states for the atoms. Bayesian Logic Programs do not allow any negation while probabilistic Datalog allows negation under the well-founded semantics. As yet, it is unclear whether probabilistic Datalog programs can be represented as Bayesian Networks. Probabilistic Datalog has been implemented in the HySpirit system (Roellecke et al, 2001) and query answering and the computation of probabilities is a two step process. First, the answers to the Datalog component of the query are computed by means of bottom-up evaluation that employs magic sets. Afterwards, the inclusion-exclusion principle is used to compute the probability of the resulting expressions in Disjunctive Normal Form. (Fuhr, 2000) states "Practical experimentation with HySpirit has shown that the evaluation of about 10 or more conjuncts is not feasible". However, recently, in (De Raedt et. al, 2007) an algorithm has been proposed that is able to perform approximate probabilistic reasoning by combining iterative deepening with binary decision diagrams and is very efficient. (De Raedt et. al, 2007) claims that "one can deal with up to 100000 conjuncts".

PROBABILISTIC EXTENSIONS OF SEMANTIC WEB LANGUAGES

In this section, we survey probabilistic extensions of RDF, RDF Schema and OWL which are W3C recommendations and thus correspond to a standard. We also have a look at probabilistic extensions of subsets of the Description Logics corresponding to OWL, i.e. SHIF(D) and SHOIN(D), and RDF (Schema).

Extensions of RDF

RDF can be considered as the most widely accepted Semantic Web language as it provides the syntactic basis for other Semantic Web languages. A proof for its success is the huge amount of software for processing RDF data that has been implemented up to now. Quite naturally, also some approaches for combining probabilities with RDF have been proposed. Fukushige (2005) proposes an RDF vocabulary for representing Bayesian Networks. In (Udrea et al., 2006), a probabilistic extension of acyclic RDF statements with a model-theoretic semantics and a fixpoint semantics has been proposed. While the first work concentrates on representation issues, the second work can be considered as probabilistic logic on its own.

Representing Probabilistic Information in RDF

In (Fukushige, 2005), a vocabulary extension of RDF has been proposed that is capable of representing the different elements of a Bayesian Network and link them to regular RDF statements. The vocabulary consists of a set of classes (*prob:Partition*, *prob:ProbabilisticStatement*, *prob:Clause*, *prob:Probability*,) and a set of predicates (e.g. *prob:predicate*, *prob:condition*, *prob:case*, *prob:about*) that can represent a Bayesian Network. This vocabulary allows to link statements to their probabilities, express conditional probabilities and more complex probabilistic statements.

Expressiveness
The vocabulary can solely represent Bayesian Networks and can basically be considered as a syntactical interchange format for Bayesian Networks. Thus, as with Bayesian Networks, cyclic probabilistic descriptions cannot be represented. We deem this as a clear disadvantage, because we think that cyclic descriptions cannot be forbidden or avoided in such an open and unstructured environment like the web.

Reasoning and Efficiency
As yet no reasoning support has been implemented. However, after having implemented a parser and wrapper for this vocabulary, in principle any tool for reasoning with Bayesian Networks can be used for reasoning.

Applicability to Information Integration
In principle, the vocabulary can be used for representing mappings between ontologies in a similar way as done with Bayesian Description Logic Programs (see a more detailed presentation on Bayesian Description Logic Programs in the subsection entitled likewise below in this chapter). A huge disadvantage, however, is that Bayesian Networks are not properly integrated with RDF

on the meta level: the vocabulary for representing Bayesian Networks uses RDF for its syntax without a tight coupling to the logical model of RDF. Therefore, RDF ontologies cannot be integrated with mappings expressed in this vocabulary properly. Clearly, with OWL ontologies, it is not possible either.

pRDF

In contrast to the former formalism that is intended to just provide a vocabulary for representing Bayesian Networks, pRDF is a formal probabilistic extension of RDF which corresponds to a of probabilistic logic on its own.

Expressiveness

pRDF is a probabilistic extension of a subset of RDF and consists of a pair (S, I) with S being a pRDF schema and I being a pRDF instance base. A pRDF schema S is defined as a finite set consisting of probabilistic quadruples extending the RDF Schema built-in predicate *rdfs:subClassOf* and non-probabilistic triples using the RDF Schema built-in predicates *rdfs:subPropertyOf*, *rdfs:range* and *rdfs:domain*. This means that in pRDF neither the subproperty relationship nor domain and range restrictions can be defined probabilistically. A pRDF instance base I is a finite set of quadruples extending the RDF built-in *rdf:type* and arbitrary properties $p \in P$.More precisely, pRDF allows the following kinds of probabilistic definitions:

- A sequel of axioms: $C(x) \rightarrow D_1(x), ..., C(x) \rightarrow D_n(x)$ and a probability distribution over the axioms in the sequel where $C \neq D_1 \neq ... \neq D_n$.
- A sequel of axioms: $P(inst, inst_1), ..., P(inst, inst_n)$ and a probability distribution over the axioms in this sequel, where $inst \neq inst_1 \neq ... \neq inst_n$ and P being either the RDF built-in ***rdf:type*** or an arbitrary user-defined property.

Furthermore, the following deterministic expressions are allowed:

- $R(x, y) \rightarrow R_2(x, y)$,
- $R(x, y) \rightarrow C(x)$,
- $R(x, y) \rightarrow C(y)$.

A disadvantage of this approach is that only a very small subset of RDF/S is supported by pRDF yielding a very low expressivity. Furthermore, pRDF instances are required to be acyclic, which again can only be realized in small and closed environments, but not on the Web as it is.

Reasoning and Efficiency

A model theoretic semantics and a fixpoint operator has been defined basing on a t-norm (Fagin, 1999). Furthermore, a reasoner has been implemented that evaluates the fixpoint operator until the least fixpoint has been reached. The properties of a t-norm allow certain pruning strategies that are employed in the reasoning algorithms. Queries to pRDF instances are atomic, i.e. conjunctions cannot be dealt with. A query is a quadruple (*i*, *p*, *S*, *P*) where *i* can be an instance, *p* can be a property, *S* can be a set of instances *i* is related to via *p* and *P* can be a probability distribution for this sequel of property axioms. At most one of the elements of the quadruple is allowed to be a variable.

Unfortunately, for pRDF schema no query answering facility has been defined yet. The reasoning engine supports only reasoning with pRDF instances.

Applicability to Information Integration

This formalism can be used for information integration with mappings. Mappings that map classes from one ontology to classes of the other ontology can be expressed. Also, mappings that map instances from one ontology to instances of another ontology can be expressed. But no mappings can be expressed that capture partly uninstancialized axioms like the ones in (14) and (15). However, the uncertainty attached to each mapping can be used for integrated reasoning with the mappings and the ontologies. But, due to the limited RDF support, not only the mappings but especially also the RDF ontologies which are to be mapped have a very limited expressivity.

Extensions of OWL

Quite naturally a number of proposals for using probabilistic knowledge on the Semantic Web focus on the extension of the Web Ontology Language as the central mechanism of representing complex knowledge in semantic web applications. When looking at the existing proposals, we see two fundamentally different approaches for combining OWL with probabilistic information.

The first kind of approach implements a loose coupling of the underlying semantics of OWL and probabilistic models. In particular these approaches use OWL as a language for talking about probabilistic models. An example of this approach is the work of Yang and Calmet (2006) that propose a minimal OWL ontology for representing random variables and dependencies between random variables with the corresponding conditional probabilities (Yang & Calmet, 2006). This allows the user to write down probabilistic models that correspond to Bayesian networks as instances of the OntoBayes Ontology. The encoding of the model in OWL makes it possible to explicitly link random variables to elements of an OWL ontology, a tighter integration on the formal level, however, is missing. A similar approach is proposed by Costa and Laskey (2006). They propose the PR-OWL model which is an OWL ontology for describing first order probabilistic models (Costa & Laskey, 2006). More specifically, the corresponding ontology models Multi-Entity Bayesian networks (Laskey & Costa, 2005) that define probability distributions over first-order theories in a modular way. Similar to OntoBayes, there is no formal integration of the two representation paradigms as OWL is used for encoding the general structure of Multi-entity Bayesian networks on the meta-level.

The second kind of approaches actually aims at enriching OWL ontologies with probabilistic information to support uncertain reasoning inside OWL ontologies. These approaches are comparable with the work on probabilistic extensions of Description Logics also presented in this section. A survey of the existing work reveals, however, that approaches that directly address OWL as an ontology language are less ambitious with respect to combining logical and probabilistic semantics that the work in the DL area. An example is the work of Holi and Hyvönnen (2006) that describe a framework for representing uncertainty in simple classification hierarchies using Bayesian networks. A slightly more expressive approach called BayesOWL is proposed by Ding and others (Ding et. al, 2006). They also consider Boolean operators as well as disjointedness and equivalence of OWL classes and present an approach for constructing a Bayesian network from class expressions over these constructs. An interesting feature of BayesOWL is some existing work on learning and representing uncertain mappings between different BayesOWL ontologies reported in (Pan et al., 2005) which is an interesting alternative to existing matching tools.

In the following, we discuss PR-OWL and BayesOWL which are the most interesting representatives of the two general approaches to combining OWL and probabilistic models in more details.

PR-OWL

Expressiveness

As mentioned above PR-OWL is an OWL Ontology that describes Multi-Entity Bayesian Networks. OWL is mainly used as a basis for a Protégé plugin for modelling MEBNs and as a language for representing MEBNs and linking them to domain ontologies encoded in OWL. On the other hand, MEBNs can be translated into

Bayesian networks. This means that PR-OWL can be used to link OWL ontologies to Bayesian networks through the MEBN formalism. The Question about the expressiveness of PR-OWL therefore boils down to an analysis of the expressiveness of MEBNs as the actual representation model for uncertainty provided by the approach. According to the authors, MEBNs are capable of representing and reasoning about probabilistic information about any sentence in first-order logic by compiling it into a Bayesian network[2] but they define some restrictions on the nature of the theory, especially on the use of quantifiers. MEBNs specify random variables representing terms and organize them in so-called fragments that describe a certain aspect of the world. Fragments have an interface that defines the terms covered by the fragment. Each fragment defines the joint distribution over the random variables in terms of conditional probabilities encoded as part of a Bayesian network. Variables in terms can be instantiated with multiple constants each instantiation leading to a unique node in the resulting network. Logical formulas are modelled by special fragments that encode the semantics of Boolean operators, quantifiers and instantiation. Fragments are linked via shared terms and additional constraints ensure that only wanted instantiations take place.

It is quite hard to say whether MEBNs are expressive enough to capture probabilistic information about OWL ontologies. In principle it should be possible to translate each OWL ontology into first order logic and assign probabilities to conditional probabilities of the resulting model by encoding it as an MEBN. So far, it has not been investigated whether the restrictions on the use of quantifiers in MEBNs affect the representation of Ontologies. The language should be expressive enough top represent mappings between terms from different ontologies that go beyond simple concept-to concept mappings because it allows to combine terms from different ontologies using arbitrary logical operators as well as the conditional probability of one given the other. It is less clear whether the representation of the mappings can be integrated with the definitions in the semantically consistent way that goes beyond simple reference to parts of the ontologies. In the same way, we could also represent the result of ontology learning methods in terms of conditional probabilities between terms. As fragments in MEBN need input in terms of instantiations of the interface, probabilistic information about instances (e.g. the probability that a paper is about a certain topic) cannot directly be encoded in MEBNs, we could, however find a workaround by explicitly representing a Bayesian classifier as a fragment.

Reasoning and Efficiency

Reasoning in MEBNs is performed by constructing a Bayesian network from the instantiations of fragments. Inside each fragment, a network fragment is created that includes random variables and conditional probabilities for all input objects based on the network pattern specified in the fragment. Here, the actual conditional probability values depend on the number of input objects. The independent network fragments are then combined into a so-called situation-specific network, a Bayesian network that is customized to the given situation in terms of fragments actually instantiated and input objects. The basic reasoning task supported by this network is to compute the probability of one or more random variables given some evidence in terms of instantiations of some input random variables. This means that we can ask for the probability that certain terms are true or false given some knowledge about the truth or falseness of some other terms.

The basic problem of MEBNs when it comes to efficiency is the complexity of the logical language supported. In particular, this has an impact on the size of the situation specific network created as this network represents probabilistic information about all instances simultaneously instead of re-evaluating a standard network multiple times. In information retrieval applications, we often assume that information objects are independent of each other and do not have to be treated in parallel. Although, the bottom-up creation of this network ensures that only the part of the network that is actually needed to answer the query is constructed, this network can still have an infinite size. It would be interesting to identify tractable subsets of MEBNs that correspond to more tractable fragments of first order logic.

Applicability to Information Integration

Again, as with the formalism of Fukushige (2005) presented above, the vocabulary can be used for representing very expressive mappings between ontologies in the MEBN formalism. However, as PR-OWL does not provide a proper integration of the formalism of MEBN and the logical basis of OWL on the meta level, OWL ontologies cannot be integrated with mappings expressed in this vocabulary properly. More specifically, as the connection between a statement in PR-OWL and a statement in OWL is not formalized, it is unclear how to perform the integration of ontologies that contain statements of both formalisms.

BayesOWL

Expressiveness

BayesOWL is an approach for representing probabilistic information about class membership within OWL ontologies. The approach can be seen as an extension of Holi & Hyvönnen (2006). Both approaches support the representation of degrees of overlap between classes in terms of conditional probabilities of the form P(C|D) where C and D are class names. These statements denote the probability that an instance that is a member of D is also a member of C. The main feature of BayesOWL is that it does not only support simple class hierarchies but also class definitions of the following form

- Equivalence: $C(x) \leftrightarrow D(x)$
- Complement: $C(x) \leftrightarrow \neg D(x)$
- Disjointness: $C(x) \rightarrow \neg D(x)$
- Intersection: $C(x) \leftrightarrow D(x) \wedge E(x)$
- Union: $C(x) \leftrightarrow D(x) \vee E(x)$

This means that BayesOWL is actually a probabilistic extension of propositional logic rather than more expressive description logics. This is a quite a strong restriction as it means that we cannot represent probabilistic information about any relations except the subsumption relation. This limits the applicability to scenarios where we are only interested in the classification of information objects and not in relations between them. This means that the approach is not suitable to support the reasoning about structured information which plays an important role in many semantic web applications.

Reasoning and Efficiency

The basic reasoning task associated to BayesOWL is given some evidence for an object in terms of its classification to determine membership probabilities for all the classes in the ontology. For this purpose a Bayesian network is constructed from the definitions in the model. As in PR-OWL network nodes with a predefined conditional probability table are used to represent Boolean Operators. This computation is done using iterative proportional fitting, a special technique from statistics that selects a probability distribution that best fits the conditional probabilities given in the network. This approach is quite different from the other approaches presented in this survey as the inference is not guided by a specific query. This can be an advantage if many queries about different aspects of the model are issued; we can expect it to be unnecessarily complex if we are only interested in very specific aspects of the model as the method will also compute probabilities that do not have an influence on the variable. Despite this fact, the use of Bayesian networks for implanting probabilistic reasoning can be expected to be relatively efficient. A special feature of BayesOWL is that it allows including probabilistic mappings between different ontologies into the inference procedure (Pan et al., 2005). Mappings are represented in terms of conditional probability statements that include concepts from different ontologies. The probabilistic influence of these statements on the distributions is used to update the distribution in the mapped ontologies. The conditional probabilities used in the mappings can be created using statistical learning methods.

In summary, the approach is well suited for applications that use rather simple classifications of information items such as documents that are classified according to a topic hierarchy. It supports the representation and semi-automated mapping of such hierarchies. As soon as the application demands for more structural information such as document metadata, the approach reaches it limits in terms of the inability to represent information about relations.

Applicability to Information Integration

This formalism provides an integration between Bayesian Networks and OWL and thus it can be used for expressing uncertain mappings between OWL ontologies and for using those mappings for integrating the information distributed over the ontologies. As only class definitions are supported, however, neither the mappings nor the ontologies themselves can contain instances which is a severe drawback of this approach. Also, the expressivity on the schema level is very low in general and thus only a very small subset of OWL can be used for expressing ontologies to be mapped (and mappings).

Extensions of Description Logics

There have been a number of approaches for extending description logics with probabilistic information in the earlier days of description logics. Heinsohn (Heinsohn, 1991) was one of the first to propose a probabilistic notion of subsumption for the logic ALC. Jaeger (Jaeger, 1994) investigated some general problems connected with the extension of T-Boxes and A-Boxes with objective and subjective probabilities and proposed a general method for reasoning with probabilistic information in terms of probability intervals attached to Description logic axioms. Recently, Giugno and Lukasiewicz proposed a probabilistic extension of the logic SHOQ along the lines sketched by Jäger (Giugno & Lukasiewicz, 2002). A major advantage of this approach is the integrated treatment of probabilistic information about Conceptual and Instance knowledge based on the use of nominals in terminological axioms that can be used to model uncertain information about instances and relations. An alternative way of combining description logics with probabilistic information has been proposed by Koller et al. (1997). In contrast to the approaches mentioned above, the P-CLASSIC approach is not based on probability intervals. Instead it uses a complete specification of the probability distribution in terms of a Bayesian network which nodes correspond to concept expressions in the CLASSIC description logic. Bayesian networks have also been used in connection with less expressive logics such as TDL (Yelland, 2000). The approaches for encoding probabilities in concept hierarchies using Bayesian networks described in the section “preliminaries and background” can be seen as a simple special case of these approaches.

We can see two general approaches for extending description logics with probabilistic information. The first is based on probability intervals describing the validity of concept inclusion axioms, the other one is based on the use of Bayesian networks for assessing and relating the probability of different features of the terminological model. In the following, we will restrict our discussion to representative approaches of these different strategies, namely P-SHOQ and P-CLASSIC.

P-SHOQ(D)

Expressiveness

P-SHOQ(D) is based on the description logics SHOQ(D) which is very close to the description logic which provides the semantics of OWL. The only feature of OWL that is not contained in the language is the use of inverse roles. In particular, the language also supports datatypes in the same way as OWL does. Probabilistic information is represented by statements of the form (C|D)[l,u] where C and D are concept expressions in SHOQ(D) and l and u are the maximal and the minimal probability that an instance of D is also an instance of C. Using this general scheme, different kinds of knowledge can be represented, for instance:

1. The probability that C is subsumed by D P(C(x)|D(x))
2. The probability that a particular individual o is a member of a concept C P(C(o))
3. The probability that an individual o is related to an instance of a concept C P(R(o,x)|C(x))
4. The probability that two individuals o and o' are related P(R(o,o'))

From a representational point of view, P-SHOQ(D) offers a lot of possibilities for supporting the task mentioned in the motivation. For the case of overlapping ontologies uncertain mappings between concepts in different ontologies can be represented using probabilistic subsumption statements of the form P(i:C(x)| j:D(x)) where C is a concept from ontology i and D a concept from ontology j. Concerning the task of ontology learning, the language is expressive enough to capture typical information that is determined in the learning process such as the concept hierarchy. We can also represent uncertain information about the range of concepts. The lack of inverse relations in the language, however, makes it impossible to represent domain restrictions. The use of nominals allows us to represent the results of instance learning both for concept and relation instances using statement 3 and 4 mentioned above.

Reasoning and Efficiency

Reasoning in P-SHOQ is based on a function μ that maps every instance of the interpretation domain Δ on a number in [0,1] such that the value of this function for all elements in Δ sum up to 1. The Probability Pr(C) of a concept expression C is defined as the sum of all μ values of the instances of C. Based on this semantics a number of reasoning task have been defined that can be solved using appropriate inference procedures. At the most basic level, the tasks supported by the language are to determine whether a given knowledge base is consistent and to compute the upper and lower bounds l and u of a conditional probability statement P(C(x)|D(x))∈[l,u]. Computing these bounds in based on independent choice logic. Different choices are specified by the possible semantic relations that could hold between any pair of concepts. This definition of choices leads to two linear equation systems whose solutions are the upper and the lower bound of the probability. Solving the equation system involves reasoning in SHOQ(D) for determining the possible choices.

Based on this general method for computing upper and lower bounds a number of reasoning tasks that generalize standard reasoning tasks in Description Logics can be defined. In particular, the approach supports the following tasks

- **Concept satisfiability:** in particular decide whether P(∃x:C(x))∈[0,0] does not follow
- **Concept Subsumption:** given two concepts C and D compute l and u such that P(C|D)∈[l,u] follows from the knowledge base
- **Concept Membership:** given an instance o and a concept C compute l and u such that P(C(o))∈ [l,u] follows from the knowledge base.
- **Role Membership:** given two instances o and o' and a relation R compute l and u such that P(R(o,o'))∈[l,u] follows from the knowledge base

These reasoning tasks provide a suitable basis for supporting tasks such as probabilistic data retrieval across different ontologies. In particular, we can formulate queries as concept expressions in SHOQ(D) and compute the probabilities that certain instances are members of this query concept. Probabilistic information originating from uncertain mappings and classifications provide background constraints for this reasoning task. A potential problem of the approach with respect to the retrieval scenario is the ability to use the probabilities as a basis for ranking. As the approach is based on intervals rather than exact probabilities, there is no total order on the results that could be used for this purpose. Another potential problem is the complexity of the approach which has not been investigated in detailed. It is clear however, that reasoning in SHOQ(D) is likely to be highly intractable.

Applicability to Information Integration

P-SHOQ can be used for expressing all the mappings mentioned in the introduction. The ontologies, however, are not allowed to contain inverse roles. Furthermore, RDF ontologies whose semantics cannot be described solely with the Description Logics paradigm cannot be integrated, because the Logic Programming paradigm which is needed for describing the RDF semantics as well, is not covered by P-SHOQ.

P-CLASSIC

Expressiveness

P-CLASSIC is a probabilistic extension of the CLASSIC Description Logics. Different from SHOQ, the CLASSIC description logics is designed for efficiency of reasoning rather that for expressive power. In particular, CLASSIC does only contain conjunction, negation on atomic concepts, universal and number restrictions as well as role fillers. As a result, deciding subsumption in CLASSIC can be computed in polynomial time based on structural comparison of concept expressions. P-CLASSIC extends the language with probabilistic information about properties of typical instances in terms of a Bayesian network. The corresponding network contains random variables indicating the following information.

- Membership in atomic concepts A
- For each Property R
    - A distribution over possible fillers o in expressions of the form P(R(x,o))
    - A distribution over possible ranges C in expressions of the form $(R(x,y) \rightarrow C(y))$ where C is specified in terms of a separate Bayesian network.
    - A distribution over the number of fillers n in equations of the from $(\exists^{=n} y{:}R(x,y))$

Additionally, the network represents an efficient encoding of the joint probability over these random variables in terms of conditional probabilities between kinds of assertions mentioned above. This means that P-CLASSIC can be used to represent probabilistic information about terminological knowledge. In particular, we can represent probabilistic subsumption relations between atomic concepts that can be used to represent uncertain mappings and the results of learning subsumption relations. The other features of the language can also be used to represent the result of ontology learning especially distributions over property fillers and ranges are useful for this purpose.

Reasoning and Efficiency

The basic reasoning service in P-CLASSIC is to compute the probability of a complex concept expression based on the definition of the joint probability distribution over atomic classes and features of relations. The inference algorithm given in (Koller et al., 1997) takes a concept expression and a P-CLASSIC knowledge base as input and returns the probability of the concept expression.

This probability is computed by bottom-up construction of a Bayesian network that represents the concept and using it to infer the probability that an arbitrary object is member of this concept expression. This method can be used to implement probabilistic data retrieval by computing the probability of a class description using a Bayesian network that has been initialized with evidence that corresponds to the properties of the individual we want to test. The fact that P-CLASSIC is based on exact probabilities rather than probability intervals means that the probability defines a natural ranking function for answers.

The major advantage of P-CLASSIC is the fact that reasoning is relatively efficient compared to other formalisms. This is due to the fact that both, the logical and probabilistic formalism have been chosen with efficiency in mind. The algorithm for constructing the Bayesian Network of a class description is defined as a direct extension of the structural subsumption algorithm of P-CLASSIC that is known to be polynomial. Additional complexity is added by the need to evaluate the network. This problem is known to have an exponential complexity, but only in the maximal the number of parents of a node. Further, the reuse of results for certain class expressions improve the time needed for actually compute the probability. This means that P-CLASSIC has relatively nice properties with respect to the computational complexity.

Applicability to Information Integration

When P-CLASSIC was devised, its application in the area of information integration was not intended. Mainly, it was intended to express and reason about the degree of overlap between concepts of an ontology. P-CLASSIC works with probabilistic formalizations of so-called p-classes each of which describes a certain class of individuals. Except of the expressibility of a the probability distribution over the role fillers of a role, the probabilistic expressions formalize concepts. The possibility to express a probability distribution over the role fillers of a role is not enough for the area of information integration. Therefore, this formalism is too restricted for being used in the area of information integration.

Extensions of Logic Programming Formalisms

Several approaches for extending Logic Programming formalisms with probabilities have been proposed. However, most of them have not been designed with the Semantic Web in mind. In the following, we discuss only those probabilistic logic programming approaches that have been designed for the Semantic Web and involve ideas about how to connect rule bases with ontologies represented in OWL or related formalisms. Two kinds of such approaches can be distinguished. The first kind integrates OWL with Logic Programming by allowing to specify a logic program and a description logics knowledge base at the same time and allowing them to interact in some way. In general, the logic program is used for querying both knowledge bases. For this purpose, the logic program can contain atoms that query the Description Logics knowledge base. We survey two approaches of this kind, (Lukasiewicz, 2005) (and a restricted version thereof by Lukasiewicz (2006)) and (Cali et al, 2008). The other kind of approaches base on a subset OWL and Logic Programming have in common and on a translation from OWL to Logic Programming formalisms that have been extended with probabilities. The subset of OWL and Logic Programming, that these approaches consider is Description Logic Programs (DLP) which is very close to Datalog (Grosof et al., 2003). (Predoiu, 2006; Predoiu & Stuckenschmidt, 2007) translates OWL ontologies that lie in the DLP fragment to a probabilistic Datalog formalism that is close to Bayesian Logic Programs (Kersting & De Raedt, 2001) while (Nottelmann & Fuhr, 2005) translate a slight extension of the DLP fragment, namely DLP with equality, to probabilistic Datalog (Fuhr, 2000).

In the following, we present a short overview on Description Logic Programs: As they are a subset of the Description Logics underlying OWL and the Logic Programming paradigm and thus have a Description Logics and a Logic Programming syntax. In the logic programming syntax, they correspond to pure Datalog without negation, equality and integrity constraints. I.e. as with Datalog, a Description Logic Program consists of facts and rules. Each rule has the form $H \leftarrow B_1, ..., B_n$, where H and the B_i are atomic formulae and $n \geq 1$. An atomic formula consists of a predicate symbol *p* followed by a bracketed n-tuple of terms ti, $p(t_1, . . ., t_n)$ with $n \geq i \geq 0$. A term can be either a constant (i.e. an instance) or a variable (i.e. a placeholder for an instance). If all terms in an atomic formula are constants, the atomic formula is called a ground atom. The left hand side of a rule, H, is called head and the right-hand side of a rule, $B_1 \wedge . . . \wedge B_m$, is called body. All variables in rules are universally quantified, although this is not explicitly written. For i = 0, the rule is called a fact. Only ground atoms are allowed in facts.

In the DLP language, the predicates are only allowed to be 2-ary and the variable graph of the body of each rule is connected and acyclic. Semantically, Description Logic Programs in the logic programming syntax do not differ from them having been specified in the description logics syntax. As reasoning is concerned with syntactical manipulations, however, Description Logic Programs in the logic programming syntax are restricted to fact-form inference with logic programming reasoners, i.e. only facts can be derived and no axioms like with description logics reasoners that reason with the description logics syntax of Description Logic Programs.

In the following we compare two formalisms that are based on the Description Logic Programming fragment and 2 formalisms that are more expressive.

Bayesian Description Logic Programs

In (Predoiu, 2006), Description Logic Programs have been embedded into the Bayesian Logic Programming formalism (Kersting & De Raedt, 2001). In this approach, the probabilistic extension has the purpose of information integration and has been proposed in order to represent uncertain mappings between ontologies and rules. Also, a means to reason with the mappings and the ontologies and rules having been mapped in an integrated way has been proposed.

Expressiveness

Bayesian Description Logic Programs (BDLPs) are a probabilistic extension of the logic programming syntax (and semantics) of Description Logic Programs (Grosof et al., 2003). In Bayesian Description Logic Programs, facts are attached with an apriori probability and rules are attached with a conditional probability where the states of the head atom are conditioned on the states of the body atoms. Like a Bayesian Logic Program, a Bayesian Description Logic Program encodes a Bayesian Network.

Reasoning and Efficiency

The basic reasoning task associated with Bayesian Description Programs is querying for the probability density of a conjunction of ground atoms given a conjunction of ground evidence atoms. In (Predoiu & Stuckenschmidt, 2007), the semantics has been extended to allow non-ground query atoms in order to enable information retrieval by deriving all ground atoms that satisfy the query and rank them by means of their probabilities. There are no complexity results known yet for Bayesian Description Logic Programs and no inference engine is available yet. However, the inference engine for Bayesian Logic Programs, Balios (Kersting & Dick, 2004) which calls Sicstus Prolog for deriving the least Herbrand Model, can be used for reasoning with Bayesian Description Logic programs as well, because Bayesian Description Logic Programs are a subset of Bayesian Logic Programs.

Applicability for Information Integration
Bayesian Description Logic Programs have been devised in order to enable Information Integration and they are able to cover all representational issues mentioned in the introduction. However, the ontologies to be mapped are restricted to the Description Logic Programming fragment and this is often a too severe expressivity restriction.

pOWL Lite$^-$ and pOWL Lite^{EQ}

(Nottelmann & Fuhr, 2005) have presented probabilistic extensions of two OWL Lite subsets. One of these subsets corresponds to Description Logic Programs and the other one to Description Logic Programs with equality. The probabilistic extensions are both based on probabilistic Datalog (c.f. the section on probabilistic models above in this chapter). OWL formulae that can be translated to Datalog can each be provided with probabilities and processed afterwards by a pDatalog system.

Expressiveness
As mentioned above, two OWL Lite subsets have been extended with probabilities. One corresponds to Description Logic Programs, its probabilistic extension being called pOWL Lite$^-$.[3] The other one corresponds to Description Logic Programs extended with equality, its probabilistic extension being called pOWL Lite^{EQ}. A translation of OWL formulae in the Description Logic Programming fragment (possibly with equality) into the Logic Programming syntax is provided and these can be attached with probabilities in the way that pDatalog allows. These probabilistic Datalog rules are processed afterwards by a pDatalog system.

Possible pOWL Lite$^-$ expressions are listed below. Note that α (α ∈ [0, 1]) which is written in front of each uncertain expression is the probability for the complete expression which is written behind it.

- **Class membership axioms:** α C(a) with α ∈ [0, 1]

This expression corresponds to the statement that a is an instance of class C with probability α

- **Complex class membership assertions:** α C(y) ← R(a, y)
- **Role assertions:** α R(a, b)
- **class inclusions:** α_1 B_1(x) ← A(x). and **…** and α_n B_n(x) ← A(x). with n ≥ 1. This expression corresponds to the OWL expression Class(A partial B_1 … B_n) and its probabilistic extension allows to express for each B_i a certainty with which A is a subclass of B_i.
- **Class inclusions with a restriction:** α B(y) ← A(x), R(x, y). This expression corresponds to the OWL expression Class(A partial restriction(R allValuesFrom B)) and its probabilistic extension allows to express the probability for A being a subclass of the class of elements that have a relation with elements of B.
- **Role inclusions:** α R(x, y) ← S(x, y).
- **Symmetric role axioms:** α R(x, y) ← R(y, x).
- **Transitive role axioms:** α R(x, z) ← R(x, y), R(y, z).
- **Domain restrictions:** α B(x) ← R(x, y).
- **Range restrictions:** α B(y) ← R(x, y).

Additionally, OWL Lite^{EQ} allows the expression of the following axioms:

- **Individual equivalence expressions:** α a = b ← U(a), U(b).
- **Maximal Cardinality of 1 expressions:** α y = z ← A(x), R(x, y), R(x, z).
- **Functional role axioms:** α y = z ← R(x, y), R(x, z).
- **Inverse functional role axioms:** α x = y

$\leftarrow$ R(x, z), R(y, z).

Y is a predicate which contains all individuals that are available in the pOWL Lite$^-$ or pOWL Lite^{EQ} knowledge base.

Additionally, in order to deal with pOWL Lite$^-$/ pOWL Lite^{EQ} more easily, a language for stating probabilistic horn rules basing on the SWRL syntax has been added. For the purpose of reasoning, however, this language is translated to pDatalog as well. Clearly, with this addition, the expressivity goes beyond Description Logic Programs. Although the supported fragment of OWL is not extended, much more of the Logic Programming fragment is covered. It is unclear whether full pDatalog or only a subset is supported.

Reasoning and Efficiency

In (Nottelmann & Fuhr, 2005), an implementation, i.e. a wrapper for a pDatalog reasoner like HySpirit, has not been provided. Efficiency for reasoning with pOWLLite$^-$ and pOWLLite^{EQ} can be considered promising due to its limited expressivity. However, with the addition of the capability for stating horn rules basing of the SWRL syntax, one might end up with the full expressivity of pDatalog. Then, the general empirical complexity results of pDatalog mentioned in the section "probabilistic languages and models" above in this chapter is carried forward to pOWLLite$^-$ and pOWLLite^{EQ} with the addition of probabilistic horn rules in the SWRL syntax.

Applicability for Information Integration

This formalism is applicable for information integration and can express all kinds of mappings suggested in the introduction. But again, the restriction of the ontologies to the Description Logic Programming fragment is often too severe. Note that, although the formalism has been additionally equipped with horn rules basing on the SWRL syntax, the integration with the translation of the OWL ontologies in the DLP fragment has not been formalized explicitly and thus cannot be considered concerning the expressivity of the ontologies.

Probabilistic Description Logic Programs with Special DL-Atoms

In (Lukasiewicz, 2005) and (Lukasiewicz, 2006), probabilistic description logic programs[4] (pdl programs) are presented that base on a loose query-based coupling of a Logic Program and a Description Logic knowledge base. The non-probabilistic formalism that pdl programs are based on has been published in (Eiter et al., 2004) as a combination of answer set programming with Description Logics. This non-probabilistic formalism has been combined with independent choice logic yielding a probabilistic extension of the base formalism.

Expressiveness

By means of the non-probabilistic base logic, a knowledge base KB = (*L*, *P*) can be specified. *L* corresponds to a classical SHIF(D) or SHOIN(D) knowledge base and *P* corresponds to a Logic Program which may contain queries to *L*. While *L* can be specified in the typical Description Logics syntax and has the typical Description Logics semantics, the Logic Program consists of a finite set of rules of the form

$$a \leftarrow b_1, ..., b_k, \text{not } b_{k+1}, ..., \text{not } b_m \text{ with } m \geq k \geq 0.$$

Here, a and the b_i are atomic formulae. An atomic formula consists of a predicate symbol *p* followed by a bracketed n-tuple of terms ti, p(t_1, . . ., t_n) with n ≥ i ≥ 0. A term can be either a constant (i.e. an instance) or a variable (i.e. a placeholder for an instance). Two kinds of negated atoms are distinguished: classically negated atoms ¬*a* and default-negated atoms *not a*. Furthermore, there are special kinds of atoms called dl-atoms that are allowed to be one of the $_{bi}$ with k ≥ i. I.e. the dl-atoms are only allowed to occur in the positive, unnegated part of the body . Such dl-atoms form a query to *L* with additional constraints that extend or shrink the instance set associated with concepts and roles occurring in L. The logic program *P* has been given a well-founded and an answer-set semantics in (Eiter et. al, 2004).

Basing on this formalism, in (Lukasiewicz, 2005) and (Lukasiewicz, 2006), a probabilistic extension has been proposed that combines this formalism with independent choice logic. A probabilistic description logic program is a knowledge base KB = (L, P, C, μ) where

- (L, P) is a dl program as explained above. Note that in (Lukasiewicz, 2005), a well-founded and an answer-set semantics have been defined for *P*.
- *C* is a choice space that corresponds to a set of sets whose union is a subset of the Herbrand Base HB_P of *P*. Alternatives, atomic choices and total choices are defined analogously to independent choice logic (c.f. the section "probabilistic languages and models" above in this chapter). No atomic choice is allowed to occur in the head of rule in *P*, but in anywhere in the body.
- μ is a probability distribution on the choice space C, i.e. μ: $\cup C \rightarrow [0, 1]$ such that $\Sigma_{a \in A} \mu(a) = 1$ for all alternatives $A \in C$ and $\mu(B) = \Pi_{b \in B} \mu(b)$ for all total choices B of C. Note that the probability of total choices imposes probabilistic independence between the alternatives of C or, differently worded, the random variables specified by C.

Reasoning and Efficiency

Probabilistic queries to a pulp knowledge base as specified above can be either atomic or complex:

1. an atomic probabilistic query queries for the probability of a formula ψ given another formula ϕ: $(\psi \mid \phi)[l, u]$. Here, l, u are placeholders for reels in the interval [0, 1] and stand for the lower bound and the upper bound of the probability. Formulas can be arbitrary contain of negation and conjunction.
2. (complex) probabilistic queries F are inductively defined as follows: each atomic probabilistic query A (with l, u being instatiated, however) is a probabilistic query. If G and H are probabilistic queries, then so are $\neg G$ and $G \wedge H$.

The correct answer to a complex probabilistic query F is defined to be the set of all substitutions θ such that $F\theta$ is a consequence of the knowledge base. With the answer set semantics, it is distinguished between answer set consequences and tight answer set consequences. For answer set consequences, every model of the knowledge base has to be a model of $F\theta$ as well. For tight answer set consequences, furthermore, *l* (resp. *u*) have to be the infimum (resp. supremum) of $Pr(\psi\theta \mid \phi\theta)$ subject to all models of KB given that $Pr(\phi\theta) > 0$.

With the well-founded semantics, $F\theta$ is a consequence of KB if $F\theta$ is true in the well-founded model. Again, a query $(\psi \mid \phi)[l, u]\theta$ is a tight well-founded answer, is *l* (resp. *u*) are the infimum (resp. supremum) of $Pr(\psi\theta \mid \phi\theta)$ given that $pr(\phi\theta) > 0$. Note that $Pr(\psi\theta \mid \phi\theta)$ is a probabilistic interpretation either under the answer-set semantics or under the well-founded semantics as defined in (Lukasiewicz, 2005), depending on the context. More specifically, Pr is a probabilistic distribution over all models.

The computation of tight answers to queries $(\psi \mid \phi)[L, U]\theta$ under the answer-set semantics involves classical logical deduction (according to the semantics used) and solving two linear optimization problems. The complexity of solving these linear optimization problems has not been discussed, yet. However, deduction under the answer set semantics has a very high complexity. More specifically, for *L* being a SHIF(D) knowledge base (resp. a SHOIN(D) knowledge base) query answering is in the complexity class co-NEXP (resp. co-NP^{NEXP}) (Eiter et. al, 2004). Query answering under the well-founded semantics is for L being a SHIF(D) knowledge base (resp. SHOIN(D) knowledge base) complete for EXP (resp. P^{NEXP}) (Eiter et. al, 2004). In (Lukasiewicz, 2006), for the same syntax as shown above for both, knowledge bases and queries, a stratified semantics based on a (local) stratification of the knowledge base has been defined. Complexity for this semantics has not been considered at all. However, query answering in stratified logic programs in general, i.e. without integrating Description Logic knowledge bases, has a much lower complexity than in those that go beyond stratification and lie in the well-founded semantics, but is still intractable in the worst case.

Applicability to Information Integration

This formalism is the first one mentioned in this chapter that is able to fully integrate full OWL and a huge part of RDF. Concerning the expressivity, this formalism is therefore very suitable for the representation of OWL (i.e. the OWL-Lite and OWL-DL fragments) and a huge part of RDF in the same syntax. However, as dl-atoms are not allowed to occur in the head of the rules, only a Logic Program can be the target of a mapping. Therefore, it cannot be used for information integration on the Semantic Web where OWL ontologies can be the target of mappings.

Probabilistic Disjunctive Description Logic Programs

In (Cali et al., 2008), a tighter integration of Logic Programs and the Description Logics underlying OWL has been combined with independent choice logic. This approach is called probabilistic disjunctive description logic programs (pddl programs) and differs from the formalism mentioned above in the fact that there are no special dl-atoms necessary for the flow of information between *L* and *P*. In fact, concepts and roles of L can occur as unary or binary predicates in P as well. Furthermore, the logic programming component *P* is allowed to have rules with disjunction in the head while with probabilistic description logic programs with special DL-atoms mentioned above, *P* was only allowed to consist of rules with a single, positive atom in the head[5]. Note also that classical negation is not allowed to occur in probabilistic disjunctive description logic programs in contrast to probabilistic description logic programs with special dl-atoms described above.

Expressiveness

As before, in the section above, a non-probabilistic base logic is combined with independent choice logic yielding probabilistic disjunctive description logic programs. The non-probabilistic logic used is disjunctive description logic programs (Lukasiewicz, 2007). It allows to specify a knowledge base KB = (*L*, *P*) with *L* being either a SHIQ(D) or a SHOIN(D) knowledge base and *P* being a logic program. *P* is a finite set of disjunctive rules of the form

$$\alpha_1 \vee \ldots \vee \alpha_k \leftarrow \beta_1, \ldots, \beta_n, \text{not } \beta_{n+1}, \ldots, \text{not } \beta_{n+m}$$

with $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_{n+m}$ being atoms built with the predicate, role and concept symbols of *P* an *L* in the usual way. The logic program *P* has been given an answer set semantics in (Lukasiewicz, 2007).

Basing on this formalism, in (Cali et. al, 2008), a probabilistic extension has been proposed that combines this formalism with independent choice logic. A pddl program is a knowledge base KB = (L, P, C, μ) where

- (*L*, *P*) is a ddl program as explained above
- *C* is a choice space that corresponds to a set of sets whose union of its elements A $\in$ C corresponds to a subset of the set $HB_P \setminus DL_P$. Here, HB_P is the Herbrand base of *P* and DL_P is the subset of the Herbrand base

of *P* that is built with predicates that occur in *L* as concepts or roles, too. Alternatives, atomic choices and total choices are defined analogously to independent choice logic (c.f. the section "probabilistic languages and models" above in this chapter).

- μ is a probability distribution on the choice space C as defined in the section above.

Reasoning and Efficiency

A probabilistic query to a pddl knowledge base has the form $\exists\ (c_1(\mathbf{x}) \vee \ldots \vee c_n(\mathbf{x}))[r, s]$ where **x**, r, s are tuples of variables, $n \geq 1$, and each $c_i(\mathbf{x})$ is a conjunction of atoms constructed from predicate and constant symbols in P and variables in **x**. Similarly to probabilistic description logic programs with special dl-atoms, it is distinguished between correct and tight answers to such a query. Given a probabilistic query $\exists\ (q(\mathbf{x}))[r, s]$, a formula $(q(\mathbf{x}))[l, u]$ with $l, u \in [0, 1]$ is a *correct consequence* of the knowledge base iff the probability of it lies always in the interval [0, 1] for every answer set of KB and every variable assignment σ. A formula $(q(\mathbf{x}))[l, u]$ with $l, u \in [0, 1]$ is a *tight consequence* of the knowledge base iff l (resp. u) is the infimum (resp. supremum) of the probability of the formula subject to all answer sets of the knowledge base and all variable assignments σ.

The consistency and the query processing problem are decidable in pddl programs. For a pddl knowledge base KB = (*L*, *P*, *C*, μ) with *L* being either a SHIF(D) or a SHOIN(D) knowledge base, deciding whether KB is consistent is complete for $NEXP^{NP}$ given that the size of C is bounded by a constant. For a pddl knowledge base KB = (*L*, *P*, *C*, μ) with *L* being either a SHIF(D) or SHOIN(D) knowledge base, deciding whether (q)[l, u] with q being a ground atom from HB_P and $l,u \in [0, 1]$ is a consequence of KB is complete for co-$NEXP^{NP}$.

In (Cali et al, 2008), a subset of pddl knowledge bases with strictly limited expressivity has been presented which allows for deciding consistency and query processing in polynomial time. However, for this purpose, the Description Logics part *L* must be in DL-Lite (Calvanese et. al, 2005) and the logic programming part *P* extended with additional rules modelling basic inclusion in *L* must be normal, i.e. only one non-negated atom in the head is allowed, and locally stratified.

Applicability to Information Integration

This formalism is capable of representing full OWL (i.e. full OWL-Lite and OWL-DL ontologies) and a huge part of RDF in the same syntax and is therefore capable for integrated query answering and reasoning with both formalisms. Furthermore, as predicates representing concepts and roles in the ontology can occur freely in the rule, i.e. also in the head, mappings can be represented with the formalism straightforwardly. Furthermore, as disjunction in the head is allowed, inconsistent mappings can be dealt with more easily that with pure horn rules that allow only one atom in the head of a rule. The representation of mappings with this formalism has been investigated and described in detail by Cali & Lukasiewicz (2007).

DISCUSSION AND CONCLUSION

We conclude the chapter with a discussion of the benefits and drawbacks of the different approaches for extending Semantic Web languages with probabilistic information that we have surveyed above. It turns out that there exist *two different kinds of probabilistic extensions*. The first kind of extensions is a rather loose coupling between an existing Semantic Web language and a probabilistic model. There, the Semantic Web Language is just used syntactically as a vocabulary for exchanging knowledge bases specified in the probabilistic model. The second kind of extensions provides a tight integration on the formal level between a Semantic Web Language or a subset of it and a probabilistic model. The second kind of extensions encompasses as well the formalisms that integrate a Semantic Web language with logic programming and combine the resulting formalisms with a probabilistic model. These extensions provide also a tight formal integration of a Semantic Web language which usually is OWL-Lite/OWL-DL or the Description Logic which underlies these OWL fragments with a logic programming formalism and a probabilistic model.Extensions of the first kind that are mentioned in this survey are the approaches of

- (Fukushige, 2005) which proposes a vocabulary for encoding Bayesian Networks with RDF,
- (Yang & Calmet, 2006) which proposes a vocabulary for encoding Bayesian Networks with OWL and
- (Costa & Laskey, 2006) which proposes a vocabulary for encoding Multi-Entity Bayesian Networks with OWL.

These approaches are rather unsatisfying because they do not consider the semantics of Semantic Web languages but rather focus at a special kind of probabilistic model, i.e. Bayesian Networks or Multi-Entity Bayesian Networks, and provide a Semantic Web based syntactical interchange format for these probabilistic models and their semantics. By means of these approaches uncertainty can only be represented on the Semantic Web but no Semantic Web statement is extended by some kind of uncertainty. Thus, from the five areas mentioned in the introduction where a consideration of uncertainty is needed on the Semantic Web, only the needs of the first area are met. I.e. only the requirements for representing statistical information are met. The area of the Semantic Web itself does not benefit substantially from these extensions. It is even arguable whether the probabilistic models represented benefit from using a vocabulary basing on a Semantic Web language without any formal integration. Note that currently no reasoning support for these vocabularies has been implemented yet, i.e. no wrappers exist that is able to parse the Semantic Web language vocabulary defined for the particular probabilistic models and feed it to a reasoner that is capable to deal with them. However, for PR-OWL, a reasoner implementation effort has recently been started.

Extensions of the second kind naturally fulfill the requirements for representing statistical information. Additionally, because of the much tighter integration on the formal level, they are also much more appropriate for Ontology matching and aligning and also for ontology learning by means of Bayesian machine learning methods. The same holds for ontology population or document classification, respectively. E.g. (Straccia & Troncy, 2006) have proposed methods for learning probabilistic mappings between OWL ontologies that are represented as very simple pDatalog rules. These methods have been implemented in the oMAP framework. The pDatalog rules that can be learned in the oMAP framework are contained in pOWLLite^{-} as well. Thus, those mappings are very much related to POWLLite^{-} and pOWLLite^{EQ}. Probabilistic disjunctive description logic programming as described above has also been proposed for usage in the area of the usage of ontology mappings and information integration. These considerations have been theoretical and no implementation has been provided, yet, but is considered as future work. In (Predoiu, 2006), Bayesian Description Logic Programs have been proposed solely for the representation of mappings and the uncertainty inherently associated with any automatically discovered mapping. An implementation, however, is not yet provided, but under development. The only further formalism for which a mapping scenario has been considered is BayesOWL. As each BayesOWL ontology corresponds to a Bayesian Network, in the mapping scenario, Bayesian Networks are mapped to each other. Hence, this scenario is computationally very expensive. The formalism which has been identified as being the most appropriate for information integration is probabilistic disjunctive description logic programming because of its expressivity concerning the ontologies to be mapped and the mappings and the possibility to deal with inconsistencies introduced by mappings to a certain extent which needs to be further investigated. For the other probabilistic extensions surveyed in this paper, no mapping scenario has been considered. Most of them have been proposed without the area of ontology mapping and information integration in mind and therefore they all have drawbacks concerning their usage in this area. Furthermore, no research on learning or using mappings has been performed yet in any of the formalisms except of pOWLLite^{-}.The probabilistic extensions that integrate Semantic Web languages or subsets thereof tightly with a probabilistic model, can be distinguished as follows:

- Extensions that consider not only the semantics but also the syntax of established Semantic Web languages, examples being pRDF and BayesOWL. Both support only a small subset of the languages they extend probabilistically. pRDF extends basically only the three RDF built-in predicates for specifying subclass relations, instance and role membership with probabilities. Furthermore, RDF built-in predicates around properties (the subproperty relation, the definition of the range and the domain of properties) are allowed to be used classically in deterministic triples. BayesOWL has an even more limited expressivity than pRDF because it does not even allow to express uncertainty of properties and instances.
- Extensions that consider subsets of the Description Logics underlying OWL, examples being P-SHOQ(D) and P-CLASSIC. P-CLASSIC has a rather limited expressivity as it combines the description logic CLASSIC that has been designed for efficiency of reasoning and suffers thus of a limited expressivity with the probabilistic model of Bayesian Networks. CLASSIC is a very small subset of SHOQ(D). For P-CLASSIC no reasoning tools have been devised. P-SHOQ(D) has the full expressivity of SHOQ(D) and is very near to OWL-DL which corresponds to SHOIN(D). The only difference is that inverse roles cannot be specified. However, for P-SHOQ(D) no reasoning tools exist either. Furthermore, the proposed reasoning algorithm can be expected to have a very high complexity because it involves solving a linear equation system.
- Extensions that consider integrations of a Logic Programming variant and a Description Logic underlying OWL. Such extensions are Bayesian Description Logic Programs, pOWLLite^{-} and pOWLLite^{EQ}, probabilistic Description Logic Programs and probabilistic Disjunctive Description Logic Programs. We think that probabilistic extensions of integration formalisms that integrate Description Logics and Logic Programs are very important also because Logic Programming is a very important paradigm especially present in the database area. Furthermore, as shown by the Rule-Interchange-Format working group at the W3C[6] that intends to carry over the Logic Programming paradigm into the Semantic Web, there is a huge interest in representing rules on the Web. In the next paragraph we will shortly summarize a comparison of the form of integration between DL and LP, the expressivity of the formalisms and the tightness of the combination between the deterministic logical model and the probabilistic model.

Two of the probabilistic approaches that integrate Logic Programming with Description Logics, integrate only a subset of OWL. These approaches are Bayesian Description Logic

Programs and pOWLLite^{-}/pOWLLite^{EQ}. Bayesian Description Logic Programs combine pure Description Logic Programs, i.e. Datalog without equality and negation, a common subset that is shared by the Description Logics underlying OWL and the Logic programming paradigm, with Bayesian Logic Programs. The integration of the deterministic and the probabilistic model is very tight and yields even a subset of the probabilistic model. pOWLLite^{-} and pOWLLite^{EQ} are intended to be a probabilistic extension of Description Logic Programs as well (the latter extends them also with equality). Besides a probabilistic extension of Description Logic Programs (possibly extended with equality) also probabilistic Horn rules are supported that increase the expressivity and it is unclear whether the expressivity ends up in full pDatalog. However, as negation is allowed and also equality, pOWLLite$^{-/EQ}$ seems to support a larger expressivity of the deterministic model. The probabilistic models used in Bayesian Description Logic Programs and pOWLLite$^{-/EQ}$ differ as well. Bayesian Logic Programs do not support negation and are a compact representation of a Bayesian Network. pDatalog supports negation under the well-founded semantics and until now no relation to Bayesian Networks has been found.

Differently from Bayesian Logic Programs and pOWLLite$^{-/EQ}$, probabilistic Description Logic Programs and probabilistic Disjunctive Description Logic Programs support full OWL-Lite and OWL-DL and integrate them with stratified logic programs, logic programs under the well-founded and under the answer set semantics. These approaches have a strong theoretical basis and all of them combine the deterministic model with independent choice logic as probabilistic model. The query language supports differently form Bayesian Logic Programs and pOWLLite^{-}/pOWLLite^{EQ} queries for probabilistic intervals. The query language is very expressive and reasoning is very complex because it involves solving a linear equation system like with P-SHOQ. However, for a restricted subset of probabilistic Disjunctive Description Logic programs, a polynomial complexity has been shown. This subset consists of a Description Logics knowledge base lying in a subset of the Description Logic programming fragment and of a Logic Program that corresponds to Datalog with negation that is locally stratified.

Most of the approaches that probabilistically integrate the Logic Programming paradigm with the Description Logics paradigm, provide own reasoners. For Bayesian Description Logic Programs, the reasoner Balios (Kerstin & Dick, 2004) that has been implemented for its probabilistic model which is a superset of itself can be used. For pOWLLite$^{-/EQ}$, HySpirit or Pire which are reasoners for full pDatalog which is their underlying probabilistic model can be used. In fact, an implementation for pOWLLite$^{-/EQ}$ basing on PIRE exists. For probabilistic Description Logic Programs and probabilistic Disjunctive Description Logic Programs no reasoners exist yet.

FUTURE RESEARCH DIRECTIONS

As overall conclusion, we can summarize that until recently, research has not paid much attention to uncertainty in the area of the Semantic Web. However, it gains more and more interest and new approaches considering uncertainty tend to emerge. Still, many of these approaches are rather half-baked and a lot of things are missing:

- **Reflections on gathering probabilities.** *Where do the probabilities used in the web come from? What kinds of probabilities exist?* Cali & Lukasiewicz (2007) make the first proposal to distinguish between mapping trust, mapping error or plain mapping probabilities. However, we think that this is just a very first step and might be a beginning for new insights into the types and usages of probability usage, depending on the event space and intended semantics.

How can those probabilities be gathered? (Straccia & Troncy, 2006) make proposals for learning very simple pDatalog rules. Investigations of methods for learning more complex structures of different probabilistic models would enable the Semantic Web community to anticipate in which forms a Semantic Web where automatic information integration would be possible.

- **Reflections on which probabilistic models are suitable for which subareas of the Semantic Web.** I.e. investigations of the applicability and usefulness of probabilistic extensions of Semantic Web languages in the different areas that need to consider uncertainty have to be done. E.g. it has to be seen whether a probabilistic Logic Programming approach is better suited for discovering and representing mappings than a purely probabilistic Description Logic one when only OWL ontologies and no rules are involved. This requirement is interweaved with the requirement above because the investigations on the different kinds of probabilities might lead to usefulness results. Furthermore, investigations on methods for learning those different probabilistic Semantic Web extensions, might naturally lead to further insights of the usability of the different formalisms in the different areas by means of complexity results and learnability results.
- **Reflections on cyclic probabilistic representations:** None of the above mentioned probabilistic extensions of Semantic Web languages can deal with cyclic representations. We deem this as a severe drawback because of the open and free nature of the Semantic Web. If ontologies, logic programs and mappings between them are considered as a whole, cyclic descriptions are very likely to occur and are not avoidable. Only in small toy worlds, cycles can be avoided. It has to be investigated in which ways cyclic probabilistic representations can be dealt with.
- **Reasoning methods and implementations:** Reasoning tools in general are not provided for the languages themselves, only for related logical formalisms which can used by means of wrappers but are not optimized for the languages at hand. If there are reasoning tools that are specialized for the languages themselves, then they support only a part of the language like in the case of pRDF. Research needs to focus on the development of optimized reasoning methods and reasoning tools need to be implemented in order to enable the usage of uncertain statements in the Semantic Web and in order to make reasoning feasible facing the huge amount of ontologies and data that can be expected to be present in the future of the Semantic Web. For example research on approximate and distributed reasoning would enable feasible query answering with large-scale knowledge bases and instance bases like imposed by the Semantic Web. None of the approaches above employ or consider currently any form of approximate or distributed reasoning.

REFERENCES

Angele, J., Boley, H., de Bruijn, J., Fensel, D., Hitzler, P., Kifer, M., et al. (2005). *Web Rule language (WRL).* W3C Member Submission,

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., & Stein, L. A. (2004). *OWL Web Ontology Language Reference*. W3C

Cali, A., & Lukasiewicz, T. (2007). *Tightly Integrated Probabilistic Description Logic Programs*. Technical Report, Institut für Informationssysteme, TU Wien

Cali, A., Lukasiewicz, T., Predoiu, L., & Stuckenschmidt, H. (2008). *Tightly Integrated Probabilistic Description Logic Programs for Representing Ontology Mappings.* In Proceedings of the International Symposium on Foundations of Information and Knowledge Systems (FOIKS).

Calvanese, D., de Giacomo, G., Lembo, D., Lenzerini, M., & Rosati, R. (2005). *DL-Lite: Tractable description logics for ontologies*. In Proceedings 20th AAAI conference on Artificial Intelligence.

Castillo, E., Gutierrez, J. M. & Hadi. A. S. (1997). *Expert systems and probabilistic network models*. Springer-Verlag

Costa, P. C. G., & Laskey, K. B. (2006). *PR-OWL: A Framework for Probabilistic Ontologies*. In Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS).

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., & Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag.

de Bruijn & J. Heymans, S. (2007). *Logical Foundations of (e)RDF(S): Complexity and Reasoning*. In Proceedings of the International Semantic Web Conference (ISWC).

De Raedt, L., Kimmig, A., & Toivonen, H. (2007). *ProbLog: A Probabilistic Prolog and Its Application in Link Discovery.* In Proceedings of the 20th International Joint Conference on Artificial Intelligence.

Decker, S., Melnik, S., van Harmelen, F., Fensel, D., Klein, M., & Broekstra, J. (2000). The Semantic Web: The roles of XML and RDF. *IEEE Internet Computing*, *4*(5), 63–67. doi:10.1109/4236.877487

Ding, L., Kolari, P., Ding, Z., & Avancha, S. (2006). *BayesOWL: Uncertainty Modeling in Semantic Web Ontologies*, Soft Computing in Ontologies and Semantic Web, Springer Verlag.

Doan, A., Madhavan, J., Domingos, P., & Halevy, A. (2003). *Ontology Matching: A Machine Learning Approach.* Handbook on Ontologies in Information Systems (pp 397-416), Springer-Verlag.

Eiter, T., Lukasiewicz, T., Schindlauer, R., & Tompits, H. (2004). *Combining answer set programming with description logics for the Semantic Web*. In Proceedings of the 9th international conference on the Principles of Knowledge Representation and Reasoning (KR-2004).

Euzenat, J., & Shvaiko, P. (2007). *Ontology Matching*. Springer Verlag.

Fagin, R. (1999). Combining fuzzy information from multiple systems. *Journal of Computer and System Sciences*, *58*, 83–99. doi:10.1006/jcss.1998.1600

Fikes, R., & McGuinness, D. (2001). *An Axiomatic Semantics for RDF, RDF-S and DAML+OIL*. W3C Note,

Fuhr, N. (2000). Probabilistic Datalog: Implementing Logical Information Retrieval for Advanced Applications. *Journal of the American Society for Information Science American Society for Information Science*, *51*(2), 95–110. doi:10.1002/(SICI)1097-4571(2000)51:2<95::AID-ASI2>3.0.CO;2-H

Fukushige, Y. (2005). *Representing Probabilistic Knowledge in the Semantic Web,*

Fukushige, Y. (2005). *Representing Probabilistic Relations in RDF*. In Proceedings of Workshop on Uncertainty Reasoning for the Semantic Web (URSW).

Giugno, R., & Lukasiewicz, T. (2002). *P-SHOQ(D): A Probabilistic Extension of SHOQ(D) for Probabilistic Ontologies in the Semantic Web*. In Proceedings Logics in Artificial Intelligence, European Conference, JELIA.

Grosof, B., Horrocks, I., Volz, R., & Decker, S. (2003). *Description Logic Programs: Combining Logic Programs with Description Logic*. In Proceedings of 12th International Conference on the World Wide Web.

Heinsohn, J. (1991). *A Hybrid Approach for Modeling Uncertainty in Terminological Logics*. In Proceedings of the European Conference on Symbolic and Qualitative Approaches to Reasoning with Uncertainty.

Holi, M., & Hyvönen, E. (2006). *Modeling Uncertainty in Semantic Web Taxonomies*. In Z. Ma (Ed.), Soft Computing in Ontologies and Semantic Web, Springer-Verlag.

Horrocks, I. (2005). *OWL Rules, OK?* In Proceedings of W3C Workshop on Rule Languages for Interoperability.

Horrocks, I., & Patel-Schneider, P. F. (2004). Reducing OWL entailment to description logic satisfiability. *Journal of Web Semantics*, *1*(4), 345–357. doi:10.1016/j.websem.2004.06.003

Horrocks, I., Patel-Schneider, P. F., Bechhofer, S., & Tsarkov, D. (2005). OWL rules: A proposal and prototype implementation. *Journal of Web Semantics*, *3*(1), 23–40. doi:10.1016/j.websem.2005.05.003

Horrocks, I., Patel-Schneider, P. F., & van Harmelen, F. (2003). From SHIQ and RDF to OWL: The making of a web ontology language. *Journal of Web Semantics*, *1*(1), 7–26. doi:10.1016/j.websem.2003.07.001

Hustadt, U., Motik, B., & Sattler, U. (2005). *Data complexity of reasoning in very expressive description logics*. In Proceedings of the 19th International Joint Conference on Artificial Intelligence.

Jaeger, M. (1994). *Probabilistic Reasoning in Terminological Logics*. In Proceedings of the 4th international Conference on Principles of Knowledge Representation and Reasoning.

Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs*. Springer-Verlag.

Kersting, K., & De Raedt, L. (2001). *Bayesian Logic Programs*. Technical Report No. 151, Institute for Computer Science, University of Freiburg, Germany.

Kersting, K., & Dick, U. (2004). *Balios – The Engine for Bayesian Logic Programs*. In Proceedings of Knowledge Discovery in Databases (PKDD).

Koller, D., Levy, A., & Pfeffer, A. (1997). *P-CLASSIC: A tractable probabilistic description logic.* In Proceedings of the 14th AAAI Conference on Artificial Intelligence (AAAI-97).

Laskey, K. B. (2006). *MEBN: A Logic for Open-World Probabilistic Reasoning.* Technical Report C4I06-01, George Mason University, USA.

Laskey, K. B., & Costa, P. C. G. (2005). *Of Klingons and Starships: Bayesian Logic for the 23rd Century.* In Proceedings of the 21st Conference of Uncertainty in AI (UAI).

Lassila, O., & Swick, R. (1999). *Resource Description Framework (RDF) Model and Syntax Specification*. W3C

Lloyd, J. W., & Topor, R. W. (1984). Making Prolog more Expressive. *The Journal of Logic Programming*, *3*, 225–240. doi:10.1016/0743-1066(84)90011-6

Lukasiewicz, T. (2005a). *Probabilistic Description Logic Programs*. In Proceedings of the 8th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty.

Lukasiewicz, T. (2005b). *Stratified Probabilistic Description Logic Programs*. In Proceedings of the ISWC-2005 Workshop on Uncertainty Reasoning for the Semantic Web (URSW).

Lukasiewicz, T. (2007). *A Novel Combination of Answer Set Programming with Description Logics for the Semantic Web* In Proceedings of the 4th European Semantic Web Conference (ESWC 2007).

Maedche, A., & Staab, S. (2004). *Ontology Learning*. In Staab, S. & Studer, R. (Eds.) *Handbook on Ontologies*. Springer 2004.

Manola, F., & Miller, E. (2004). *RDF Primer*. W3C

Nottelmann, H. (2005). *Inside PIRE: An extensible, open-source IR engine based on probabilistic logics.* Technical Report, University of Duisburg-Essen, Germany.

Nottelmann, H., & Fuhr, N. (2006). *Adding Probabilities and Rules to OWL Lite Subsets based on Probabilistic Datalog.* In International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 14(1):17-41.

Pan, R., Ding, Z., Yu, Y., & Pen, Y. A. (2005). *A Bayesian Network Approach to Ontology Mapping*, In Proceedings of the Fourth International Semantic Web Conference (ISWC).

Poole, D. (1997). The independent choice logic for modelling multiple agents under uncertainty. *Artificial Intelligence*, *94*(1-2), 7–56. doi:10.1016/S0004-3702(97)00027-1

Predoiu, L. (2006). *Information Integration with Bayesian Description Logic Programs*. In 3rd IIWeb Interdisciplinary Workshop for Information Integration on the Web.

Predoiu, L., & Stuckenschmidt, H. (2007). *A probabilistic Framework for Information Integration and Retrieval on the Semantic Web*. In Proceedings of the 3rd International Workshop on Database interoperability (InderDB).

Roelleke, T., Lübeck, R., & Kazai, G. (2001). *The HySpirit Retrieval Platform.* In Proc. of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. doi:10.1145/505282.505283

Straccia, U., & Troncy, R. (2006). *Towards Distributed Information Retrieval in the Semantic Web: Query Reformulation Using the oMAP Framework.* In Proceedings of the 3rd European Semantic Web Conference (ESWC).

Udrea, O., Subrahmanian, V. S., & Maijkic, Z. (2006). *Probabilistic RDF*. In Proceedings of the Conference on Information reuse and integration.

Yang, Y., & Calmet, J. (2006). *OntoBayes: An Ontology-Driven Uncertainty Model.* In Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06).

Yelland, P. M. (2000). *An Alternative Combination of Bayesian Networks and Description Logics.* In Proceedings of of the 7th international Conference on Knowledge Representation (KR).

ADDITIONAL READINGS

General Logic

Ben-Ari, M. (2001). *Mathematical Logic for Computer Scientists*. Springer-Verlag.

Ebbinghaus, H.-D., Flum, J., & Thomas, W. (2007) *Mathematical Logic*. Springer-Verlag

Enderton, H. B. (2002). *A mathematical Introduction to Logic*, 2nd edition, Academic Press.

Schoening, U. (1994). *Logic for Computer Scientists*. Birkhaeuser Verlag

Logic Programming

Abiteboul, S., Hull, R., & Vianu, V. (1995) *Foundations of Databases: The Logical Level*. Addison Wesley.

Baral, C. (2003). *Knowledge Representation, Reasoning and Declarative Problem Solving*. Cambridge University Press.

Lloyd, J. W. (1987). *Foundations of Logic Programming*. Springer-Verlag.

Ullman, J. D. (1988). *Principles of Database and Knowledge-Base Systems, Volume I*. Computer Science Press.

Ullman, J. D. (1989). *Principles of Database and Knowledge-Base Systems, Volume II*. Computer Science Press.

Zaniolo, C., Ceri, S., Faloutsos, C., Snodgrass, R. T., Subrahmanian, V. S., & Zicari, R. (1997). *Advanced Database Systems*. Morgan Kaufmann.

General Probability Theory

Georgii, H.-O. (2008). *Stochastics*. de Gruyter Verlag

Ross, S. M. (2005). *A first course in Probability*. Prentice Hall.

Bayesian Networks and Graphical Models

Castillo, E., Gutierrez, J. M. & Hadi. A. S. (1997). *Expert systems and probabilistic network models*. Springer-Verlag

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., & Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag.

Edwards, D. (2000). "Introduction to Graphical Modelling", 2nd ed. Springer-Verlag.

Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs*. Springer-Verlag.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Bayesian Logic Programming

Kersting, K. (2006). *An Inductive Logic Programming Approach to Statistical Relational Learning*. Frontiers in Artificial Intelligence, Volume 148, IOS Press, Amsterdam, The Netherlands.

Kersting, K., & De Raedt, L. (2001). *Bayesian Logic Programs*. Technical Report No. 151, Institute for Computer Science, University of Freiburg, Germany.

Kersting, K., & De Raedt, L. (2007). *Bayesian Logic Programs: Theory and Tool*. In Getoor, L. & Taskar, B. (Ed.), Introduction to Statistical Relational Learning. MIT Press.

Independent Choice Logic

Poole, D. (1997). The independent choice logic for modelling multiple agents under uncertainty. *Artificial Intelligence*, *94*(1-2), 7–56. doi:10.1016/S0004-3702(97)00027-1

Multi-Entity Bayesian Networks

Laskey, K. B. (2006). *MEBN: A Logic for Open-World Probabilistic Reasoning.* Technical Report C4I06-01, George Mason University, USA.

Laskey, K. B., & Costa, P. C. G. (2005). *Of Klingons and Starships: Bayesian Logic for the 23rd Century*. In Proceedings of the 21st Conference of Uncertainty in AI (UAI).

Probabilistic Datalog

Fuhr, N. (2000). Probabilistic Datalog: Implementing Logical Information Retrieval for Advanced Applications. *Journal of the American Society for Information Science American Society for Information Science*, *51*(2), 95–110. doi:10.1002/(SICI)1097-4571(2000)51:2<95::AID-ASI2>3.0.CO;2-H

RDF and RDF Schema

A selection of documents on RDF and RDF Schema (Specification, Use Cases, Recommended Readings, Tools, Related Technologies, etc

de Bruijn & J. Heymans, S. (2007). *Logical Foundations of (e)RDF(S): Complexity and Reasoning*. In Proceedings of the International Semantic Web Conference (ISWC).

OWL

Horrocks, I., Patel-Schneider, P. F., & van Harmelen, F. (2003). From SHIQ and RDF to OWL: The making of a web ontology language. *Journal of Web Semantics*, *1*(1), 7–26. doi:10.1016/j.websem.2003.07.001

Smith, M. K., Welty, C., & McGuiness, D. L. (2004). *OWL Web Ontology Language Guide*. W3C

ENDNOTES

1. For the sake of simplicity we are considering here only a part of the ontologies. The complete ontologies can be found at http://oaei.ontologymatching.org/.
2. Due to the semi-decidability of First-order logic this can only be true if the translation allows for networks of infinite size.
3. Note that Description Logic Programs are called OWL Lite$^-$ in (Nottelmann & Fuhr, 2005). This is the reason for calling its probabilistic extension pOWL Lite$^-$.
4. Note that although the formalism is called description logic programs like the formalism in (Grosof et al., 2003), it is a completely different language as it goes beyond the common subset of Description Logics and Logic Programming. In order to hint the difference, we are using lower case letters for this formalism while we call the formalism from Grosof et al. (2003) Description Logic Programs.

5 Note that conjunction in the head is allowed with probabilistic description logic programs with special DL-atoms as well, because rules with conjunction in the head can be split to regular horn rules by means of the Lloyd-Topor-Transformation (Lloyd & Topor, 1984).

6 http://www.w3.org/2005/rules

This work was previously published in The Semantic Web for Knowledge and Data Management: Technologies and Practices, edited by Z. Ma and H. Wang, pp. 74-105, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 7.6
Estimating the Privacy Protection Capability of a Web Service Provider[1]

George O.M. Yee
Institute for Information Technology, National Research Council, Canada

ABSTRACT

The growth of the Internet has been accompanied by the growth of Web services (e.g., e-commerce, e-health, etc.), leading to important provisions put in place to protect the privacy of Web service users. However, it is also important to be able to estimate the privacy protection capability of a Web service provider. Such estimates would benefit both users and providers. Users would benefit from being able to choose (assuming that such estimates were made public) the service that has the greatest ability to protect their privacy (this would in turn encourage Web service providers to pay more attention to privacy). Web service providers would benefit by being able to adjust their provisions for protecting privacy until certain target capability levels of privacy protection are reached. This article presents an approach for estimating the privacy protection capability of a Web service provider and illustrates the approach with an example.

INTRODUCTION

This work considers Web services to be: a) Web-based services that employ Extensible Markup Language (XML), Web service Definition Language (WSDL), Simple Object Access Protocol (SOAP), and Universal Description, Discovery, and Integration (UDDI) in a Service-Oriented Architecture (SOA) (O'Neill, Hallam-Baker, MacCann, Shema, Simon, Watters, et al., 2003); and b) existing and previous generations of Web-based applications that involve Web browsers interacting with Web servers that do not employ XML, WSDL, SOAP, or UDDI. This work applies to all Web services described above.

Numerous Web services targeting consumers have accompanied the rapid growth of the

Internet. For example, Web services are available for banking, shopping, learning, healthcare, and government online. However, most of these services require a consumer's personal information in one form or another, leading to concerns over privacy. For Web services to be successful, privacy must be protected. Various approaches have been used to protect personal information, including data anonymization (Iyengar, 2002; Kobsa & Schreck, 2003) and pseudonym technology (Song, Korba, & Yee, 2006). Approaches for privacy protection that are in the research stage include: treating privacy protection as an access problem and then bringing the tools of access control to bear for privacy control (Adams & Barbieri, 2006); treating privacy protection as a privacy rights management problem using the techniques of digital rights management (Kenny & Korba, 2002); and considering privacy protection as a privacy policy compliance problem, verifying compliance with secure logs (Yee & Korba, 2004).

It is also important to estimate the privacy protection capability of a Web service provider. Suppose such estimates for similar Web services A, B, and C are made available to consumers. This leads to the following benefits. If the consumer has to choose one service from among A, B, and C, then the estimates can help the consumer decide which service to select (probably the service that has the highest capability for privacy protection). In addition, the fact that consumers have access to these estimates may encourage service providers to pay more attention to protecting consumer privacy and result in higher levels of consumer trust and acceptance of Web services. Alternatively, Web service providers can use such estimates to implement services that meet predefined goals of privacy protection. Predefined levels of the estimates could be expressed as quality-of-service requirements. The estimates could then be evaluated for incremental versions of a service until the predefined levels are achieved.

The objectives of this article are to a) define estimates of the privacy protection capability of a Web service provider, b) show how the estimates can be calculated, and c) illustrate the calculation of the estimates using a Web service example.

This article extends the work of Yee (2006) by: a) improving the practicality of the approach by refocusing on estimating privacy protection capability rather than measuring how well privacy is protected; b) updating the definition of the estimates; c) updating the method for calculating the estimates; d) updating and extending the application example; e) enlarging the related works section; f) adding an evaluation section; and g) improving the clarity of the writing in all sections.

The rest of this article is organized as follows. Section "Estimates of Privacy Protection Capability" introduces the privacy protection model and defines the estimates. "Calculation of the Estimates" shows how to calculate the estimates. The section called "Application Example" illustrates the calculation of the estimates. A discussion of related work then follows. "Evaluation of Approach" discusses the strengths and weaknesses of the approach. Finally, the article ends with conclusions and directions for future research.

ESTIMATES OF PRIVACY PROTECTION CAPABILITY

Privacy

In order to define estimates of a Web service provider's capability to protect consumer privacy, it is necessary first to examine the nature of personal privacy. As defined by Goldberg, Wagner, and Brewer (1997), privacy refers to the ability of individuals to *control* the collection, retention, and distribution of information about themselves. This leads to the following definitions for this work.

DEFINITION 1: *Privacy refers to the ability of individuals to control the collection, use,* *retention, and distribution of information about themselves.*

DEFINITION 2: *A provider's protection of user privacy refers to the provider's use of provisions to give a user desired control over the provider's collection, retention, and distribution of information about the user.*

Definition 1 is the same as given by Goldberg et al. (1997) except that it also includes "use." To see that "use" is needed, consider, for example, that one may agree to give out one's credit card number (private information) to pay for one's own purchase but not to pay for someone else's purchase. The "provisions" in Definition 2 refer to whatever means or technologies are needed to give the user the required control (uphold the user's privacy); for example, access control mechanisms, policy negotiation mechanisms, and policy compliance mechanisms. These provisions depend on the nature of the control required by the user. For example, if the user specifies that the user's information is not to be kept past a certain date, the provider must have in place a provision to track how long the information has been in its possession.

It follows from Definition 2 that if the service provider is to make provisions to protect the user's privacy, it needs to know how the user wishes to control personal information. Thus, there must be a means of communicating the nature of this control, from the user to the service provider. This communication is normally carried out using a statement of privacy preferences called a *privacy policy*. Figure 1 is an example of a user privacy policy for e-learning from Yee and Korba (2005). In Figure 1, each item of information about the user corresponds to a "privacy rule" that spells out who can collect the information, how the item is to be used (purpose), how long it can be retained, and who it can be disclosed to. For example, the information item "name, address, tel" is to be used for identification; it may be collected by E-Learning Inc., it can be retained by E-Learning Inc. indefinitely, and E-Learning Inc. must not disclose it to any other party. Figure 2 illustrates the use of a privacy policy to express how the user wishes to control private information. The arrows from the user are numbered to show that the privacy policy must come first. The bidirectional arrow between the user and the privacy policy indicates that the user both specifies the policy and complies with it. On the provider side, the provisions for user control (blue box) must comply with the privacy rules in the user's privacy policy, giving the user desired control over the user's personal information. The service provider would have to agree to comply with the user's privacy policy before it can receive any of the user's private information. Where the service provider does not agree to the user's policy, the user can negotiate with the provider (Yee & Korba, 2003a, 2003b) until there is agreement, or the user can try a different provider.

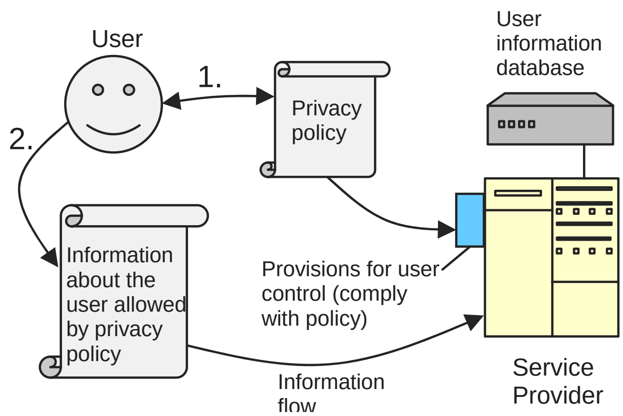
Privacy Policy Violations

Once the provider has agreed to comply with the user's privacy policy and is in possession of

Figure 1. Example privacy policy for e-learning

Header	*Policy Use:* e-learning *User:* Alice User *Valid:* unlimited
Privacy Rule	*Collector:* E-Learning Inc. *What:* name, address, tel *Purposes:* identification *Retention Time:* unlimited *Disclose-To*: none
Privacy Rule	*Collector:* E-Learning Inc. *What:* Course Marks *Purposes:* Records *Retention Time:* 2 years *Disclose-To*: none

Figure 2. Using a privacy policy to express user control over private information



the user's private data, the user's privacy is fully protected, provided there are no violations of the user's privacy policy. To define estimates of a provider's capability to protect privacy or to avoid privacy policy violations, it is necessary to look at where violations can occur. To see where violations can occur, the flow of the private information is traced, from the point where it leaves the service user to where the information is used and stored. This leads to the next definition.

DEFINITION 3: *The private information path (PIP) is the path drawn through all points traversed by the private information, from the point where it leaves the service user to the point(s) where it is stored, possibly traversing points in between where it is used. The PIP can traverse multiple providers where providers disclose the user's private information to other providers.*

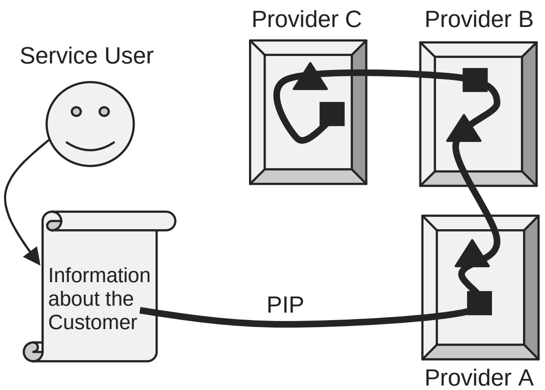
The PIP (see Figure 3) is an informal mechanism to help the security/privacy analyst visualize where attacks on the private information can happen.

Privacy violations can be classified as internal and external violations, as follows:

DEFINITION 4: *An internal violation (IV) (or inside attack) of privacy policy is one that is carried out by an insider of the provider organization (i.e., someone who has special data access privileges by virtue of an association with the organization [e.g., employee]), whose access and use of the private information does not comply with the user's privacy policy. An external violation (EV) (or outside attack) of privacy policy is one that is carried out by a noninsider of the provider organization, whose access and use of the private information does not comply with the user's privacy policy.*

An example of an internal violation is where an employee secretly copies private information and sells it to the provider's competitors, in violation of privacy policy. Other examples of internal violations are from the category of unintentional or accidental violations, where the user's private information could be leaked (e.g., private data inadvertently sent in an e-mail) or misplaced (e.g., recently, a prominent bank was discovered to be faxing confidential customer records to a farmer) due to poor business processes. Thus, an internal violation may be very difficult to detect since on the surface the employees may appear to comply with the user's privacy policy. An example of an external violation is where an attacker unknown to the provider plants a Trojan horse inside the

Figure 3. Example PIP; squares indicate storage points, triangles indicate use points



provider's computer system to steal confidential private data.

Estimating a service provider's capability to protect privacy involves looking at what provisions are in place to prevent IV and EV, that is, prevent attacks that violate the user's privacy policy.

Provisions against IV would have to cover both violations due to poor information management (e.g., lack of suitable tracking mechanisms) and violations that are intentional and malicious. These violations correspond to information management vulnerabilities within the provider organization that allow the violations to occur. Examples of such vulnerabilities are:

- Lack of anyone in the provider organization who is accountable for the private information in the organization's possession.
- Poor business processes that lack mechanisms to track which data are used where or used for what purpose; for example, employees casually providing the names of clients for a survey.
- Poor education and enforcement of company policies regarding the proper care and handling of personal information; for example, employees bringing work home that contains private information in the clear.
- Divulging personal information unwittingly to an attacker who uses social engineering.
- Lack of adequate security provisions to protect private information; for example, private data stored in the clear in data bases.
- Poor working conditions that give rise to employees feeling unfairly treated by management (can lead to employees seeking revenge through IV).

The following provisions aim to prevent IV or lessen the probability of it occurring:

- Educating employees and effectively enforcing company policies regarding the proper care and handling of personal information.
- Training employees on how to recognize and resist social engineering that targets the divulgence of personal information.
- Use of a privacy policy compliance system (PPCS) (Yee & Korba, 2004; Lategan & Olivier, n.d.) that automatically ensures that the user's privacy policy is not violated.
- Use of a monitoring system to monitor how insiders make use of the private data; the monitoring can be done in real time or offline (users' sessions recorded).
- Use of cryptographically secure logs (these logs can be later inspected to check for policy violations) to record each transaction involving private data on all servers.
- Use of reputation mechanisms to record and indicate the past performance of the provider organization in terms of integrity (e.g., Better Business Bureau).
- Use of seals of approval that attest to the fact that the provider organization has undergone and passed rigorous inspections of its processes; for example, ISO 9001: 2000 (International Organization for Standardization, n.d.).

This list is of course not exhaustive. A provider may employ none, one, or more than one of these provisions.

In the case of provisions against EV, the question to ask is: "What are possible EV violations of a privacy policy?" These violations are carried out by attackers who have not been granted access to the targeted private information. These attackers target a range of security vulnerabilities, from software systems that can be breached to access the private information to simple theft of laptops and other devices used to store private information.

Our estimates of the capability to protect privacy will depend on the provisions that have been put in place against both IV and EV vulnerabilities.

There are situations in which multiple service providers may be involved in a single service. In these situations, a provider may share private information with other providers. For example, an online book store (e.g., Amazon.com) may make use of an online payment service (e.g., Paypal.com) and a shipping service (e.g., fedex.com) in order to sell the consumer a book. For the sake of exposition, the *first* provider is the one with which the user chooses to interact. *Second* providers are providers with which the first provider shares the user's private data in order to complete its purpose (such as selling a book in the example above). *Third* providers are ones with which the second provider shares the original user's private data in order to use the third providers' services. Similarly, it is possible to define *fourth* providers, *fifth* providers, and so on. For convenience, label second, third, fourth, and so forth providers as *chained* providers. In order to evaluate the first provider for its capability to protect privacy, it is necessary to carry out the same evaluation for all chained providers that are linked to the first provider in terms of service usage, as just described, and that receive the original user's private information due to this linkage. In other words, IV and EV would need to be examined not just for the first provider but also for each chained provider in turn. Second providers would be identified in the original user's privacy policy under "Disclose-To" (see Figure 1). Similarly, third providers would be identified in the first provider's privacy policy, fourth providers would be identified in the second provider's privacy policy, and so on. Of course, all second providers have to agree to comply with the first provider's privacy policy (the first provider is the "user" here), all third providers have to agree to comply with the second provider's privacy policy, and so on. Further, the first provider would incorporate into its privacy policy the portions of the original user's privacy policy that relate to the private information to be shared with the second providers, each second provider would incorporate into its privacy policy the portions of the original user's privacy policy that relate to the private information to be shared with the third providers, and so on.

Past Violations

Intuitively, a service provider's record of past privacy violations should impact its future capability of privacy protection. As for a sex offender, one could say that the fact that a service provider has violated privacy in the past means that it is more likely to lack the capability of privacy protection in the future. However, the comparison may not be so clear cut. A service provider may be intensely motivated by profit and public image. In fact, one could argue that a company that has violated privacy in the past is more likely to put measures in place to avoid violating privacy in the future in order to protect its public image and profit, especially if knowledge of the past violation is in the public domain. In this case, the capability of privacy protection in the future is *increased* if it has violated privacy in the past, not decreased. The influence of a provider's past privacy violations on its future capability of protecting privacy can be postulated as depending on at least the following factors:

- The type of service provider, for example, profit-oriented, nonprofit-oriented: A profit-oriented provider will probably want to put measures in place to avoid future privacy violations to protect its profit, increasing its future capability of protecting privacy;
- The severity of the past privacy violation: The type of information violated and the number of people affected by the violation contribute to this severity, for example, a disclosure of credit card numbers affecting 10,000 people would generally be regarded as more severe than a disclosure of personal e-mail addresses affecting 100 people; probably the more severe the violation, the more the service provider is motivated to avoid violations in the future, likely increasing its future capability of protecting privacy;
- The time when new measures against future violations were installed: If the installation occurred after the past violations, this could indicate that the service provider was serious about avoiding future violations, likely increasing its future capability of protecting privacy. Of course, the installation may have been carried out for other reasons (e.g., window dressing to prop up the company's shares); it is difficult to be sure, but in general perhaps the provider can be given the benefit of the doubt.

Given the above discussion, it is suggested that past violation information not be integrated into estimates of future capability of protecting privacy, but rather that it is treated simply as *contextual information*, to be considered in conjunction with the estimates developed below as "support," that is, suggesting an increase in future privacy protection capability. This is primarily due to the imprecise nature of any conclusions that might be drawn from past violations, as discussed. Table 1 shows how the above factors from past violations can, to varying degrees, support the future capability of protecting privacy. To use Table 1, one would first calculate the estimates of future privacy protection capability as presented below. Then these estimates would be supported by the entry in Table 1 that corresponds to the choice of past violation influencing factors down the left side and across the top of the table. If the service provider has no past privacy violations, Table 1 does not apply, and the estimates are not affected by past violations.

Note that the above ideas on provider behavior in response to a privacy violation have not been verified. This is left for future work.

Definition of the Estimates

An estimate of a service provider's capability for protecting the service user's privacy may be defined as follows:

Table 1. Support for future privacy protection capability based on the service provider's response to past violations

	Nonprofit-Oriented Service Provider		Profit-Oriented Service provider	
Severity of Past Violations	No New Measures Installed Post Violations	Some New Measures Installed Post Violations	No New Measures Installed Post Violations	Some New Measures Installed Post Violations
Low	Very Low Support	Low Support	Low Support	Medium Support
Medium	Very Low – Low Support	Low-Medium Support	Low-Medium Support	Medium-High Support
High	Low Support	Medium Support	Medium Support	High Support

- **DEFINITION 5:** An *estimate* of a provider's capability for protecting user privacy is a numerical rating (e.g., percentage) that indicates the approximate degree to which the provider is capable of avoiding IV and EV.

In Definition 5, suppose for example that the estimate (rating) of a provider's capability for protecting user privacy is 70%. This means that the provider's installed provisions are *capable* of avoiding violations of user privacy approximately every 7 out of 10 times. This does NOT mean that the provider's installed provisions *actually* avoid the violations approximately every 7 out of 10 times. An estimate as described in Definition 5 only rates capability; rating actual privacy protection performance would be much more complex and is not needed to achieve the benefits claimed for the approach proposed here.

In Definition 5, the capability to avoid IV and EV depends on effective protective provisions (e.g., encrypting the private data, together with careful encryption key management) that the organization has in place to prevent violations. Let E denote an estimate of capability to protect privacy. By Definition 5, E will need to account for the provisions used against IV and EV.

To account for the provisions against IV, we propose that a special privacy impact assessment (PIA) (Treasury Board of Canada, n.d.), explained below—extended to identify vulnerabilities that can lead to malicious IV—be carried out to identify IV vulnerabilities. Suppose that such an assessment identified that m IV vulnerabilities and countermeasures (provisions against IV) are in place for p of these vulnerabilities. To account for provisions against EV, we propose that a special security threat analysis (Salter, Saydjari, Schneier, & Wallner, 1998), explained below, oriented towards discovering EV vulnerabilities be carried out. Suppose that this analysis identified that n security vulnerabilities and countermeasures (provisions against EV) are in place for q of these vulnerabilities. Then, one formulation of E is (See Box 1).

Another formulation of E is:

$E_2 = (e_i, e_e)$

where e_i accounts for the provisions used against IV and e_e accounts for the provisions used against EV, and (See Box 2).

E_1 has the advantage of providing a single number for ease of comparison between different providers. A threshold t for E_1 may be predetermined such that for E_1 above t, the provisions

Box 1.

$$E_1 = (p + q) / (m + n), \quad \text{if } m + n > 0, \quad \text{so that } 0 \le E \le 1$$
$$= 1, \quad \text{if } m + n = 0.$$

Box 2.

$$e_i = p/m, \quad \text{if } m > 0, \quad \text{so that } 0 \le e_i \le 1$$
$$= 1, \quad \text{if } m = 0$$

$$e_e = q/n, \quad \text{if } n > 0, \quad \text{so that } 0 \le e_e \le 1$$
$$= 1, \quad \text{if } n = 0.$$

installed by the provider against IV and EV are deemed to give it an adequate capability to protect privacy.

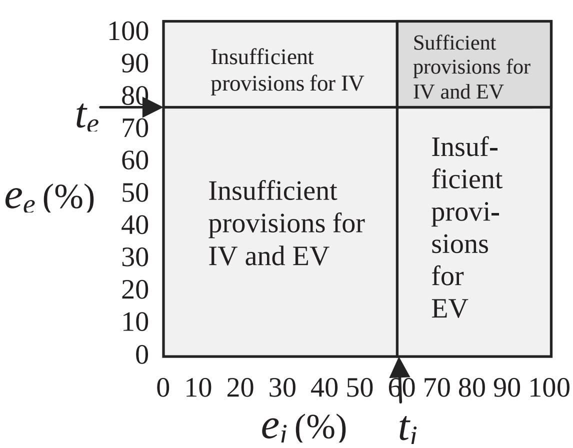
E_2 has the advantage of focusing in on where an organization stands in terms of its provisions against IV or EV separately. Thresholds t_i and t_e may be predetermined for e_i and e_e respectively, such that for e_i or e_e above its respective threshold, the corresponding installed provisions against IV or EV are deemed to give the provider an adequate capability to protect against IV or EV. In practice, e_i and e_e may be expressed as percentages that define a region in a 100 x 100 plane in which a provider's capability to avoid privacy policy violations is adequate (acceptable) (shaded region in Figure 4).

We will use both E_1 and E_2. The thresholds t, t_i, and t_e may be set by a privacy authority, such as a privacy commissioner, responsible for ensuring that the public's privacy is protected.

For chained provider situations (see above), the evaluation of E requires special treatment. The following rule is proposed for these situations.

- CHAINED RULE: *E evaluation for a service with chained providers:* A first provider passes E evaluation (i.e., each estimate is above or equal to its threshold) if and only if the first provider and each of its chained providers that receives the original user's private information all pass E evaluation.

Figure 4. Region (shaded) in which a service provider's capability to protect privacy is acceptable



The chained rule ensures that if at least one chained provider that receives the original user's private data fails E evaluation, the corresponding first provider is also regarded as having failed, even if it itself passes. This outcome seems to agree with personal wishes since the original user would not want private information abused by a chained provider after the user placed trust in the first provider.

CALCULATION OF THE ESTIMATES

Determination of *m* and *p* for IV

This determination requires a special PIA (Treasury Board of Canada, n.d.) in order to identify IV vulnerabilities and provisions against these vulnerabilities. A basic description of how a PIA is carried out, along with an extension to identify vulnerabilities that could lead to malicious IV follows.

A PIA is a comprehensive process designed to assist organizations to determine the impacts of program and service delivery initiatives on individual privacy. It has the following main stages:

a. Project Initiation: Define the scope for the PIA, allocate team resources, and adapt PIA tools according to the scope. The team may consist of privacy and security experts, legal experts, program managers, system managers, and so on.
b. Data Analysis: Describe proposed business process diagrams and identify clusters of personal information in business processes. Develop detailed data flow charts

c. Privacy Analysis: Complete privacy analysis questionnaires. Discuss answers to questions that require further detail. Identify and describe privacy issues and implications
d. Privacy Impact Assessment Report: Summarize the privacy risks and evaluate the degree of risk involved. Identify and discuss actions or options to mitigate the risks. End by taking any other considerations into account and describe the path forward.

We are primarily interested in the vulnerabilities identified by the privacy analysis portion of the PIA through a series of questionnaires. The latter are concerned with the management of private information in order to comply with privacy legislation. To identify vulnerabilities that could lead to malicious IV, we propose extending the privacy analysis with an additional questionnaire designed to discover risks that could result in malicious IV, as shown in Table 2, to be used in conjunction with the PIP.

In identifying vulnerabilities, the PIA team may weigh the vulnerabilities in terms of how likely they are to lead to violations, and eliminate the ones that are unlikely to be violated. The weighing process may consider such factors as risk to the violator that the violator could be caught as well as the violator's motivation for the violation.

The total number *m* of IV vulnerabilities is the sum of the number of vulnerabilities identified using this questionnaire and the number of vulnerabilities identified in the above PIA that are potential internal violations. The number of provisions already in place countering these vulnerabilities gives *p*. Since an organization may plan to install a certain number of such provisions

Table 2. Questionnaire to identify vulnerabilities leading to malicious IV

	Question	Rationale
1.	Is the private information of high value to outside agencies or a competitor?	The higher the value, the more a malicious attacker will be tempted to steal and sell the information.
2.	What are some possible ways for an unauthorized insider to gain access to the private information?	This question will identify security weaknesses.
3.	What are some possible ways for an authorized insider to violate the privacy policy?	This question will identify nonsecurity weaknesses (e.g., using the private information for a different purpose).
4.	Does the organization have an employee assistance program that includes counseling and help with financial difficulties?	Such a program may eliminate some financial motivation for a malicious IV.
5.	Does the organization have an ombudsman or other impartial agent to assist employees with their grievances?	Such an impartial agent may eliminate or reduce the motivation to seek revenge by committing a malicious IV.
6.	Does the organization have a history of perceived injustices to employees?	If the answer is "yes," employees may be motivated by revenge to commit a malicious IV.
7.	Does the organization conduct a stringent background and reliability check on a candidate for employment prior to hiring the candidate?	While a background and reliability check is not guaranteed to weed out potential inside attackers, it should eliminate those with criminal backgrounds.
8.	Does the organization require candidates for employment to disclose any potential conflicts of interest they may have with respect to their new employment and any outside interests prior to hire? Does the organization require ongoing disclosure of conflicts of interest after hire?	Eliminating conflicts of interest should reduce related motivations for malicious inside attacks. For example, an inside attacker may secretly compromise private information in favor of an outside interest, believing that the compromise is undetected.

in the future, it is possible to obtain *p*, reflecting both provisions in place and planned. However, the author's opinion is that *p* should count only provisions already in place, since something that is planned may never actually happen.

Determination of *n* and *q* for EV

This determination requires a threat analysis of security vulnerabilities in the organization's systems that could allow EV to happen. An overview of threat analysis follows.

Threat analysis or threat modeling is a method for systematically assessing and documenting the security risks associated with a system (Salter et al., 1998). The results can help development teams identify the strengths and weaknesses of the system and serve as a basis for investigations into vulnerabilities and required mitigation. Threat modeling involves understanding the adversary's goals in attacking the system based on the system's assets of interest. It is predicated on that fact that an adversary cannot attack a system without a way of supplying it with data or otherwise accessing it. In addition, an adversary will only attack a system if it has some assets of interest. The following threat modeling terminology is selected from Salter et al. (1998):

- *Attack path*: A sequence of conditions in a threat tree that must be met for an attack goal (threat) to be achieved. A valid attack path (one with no mitigated conditions) is a vulnerability.
- *Threat*: The adversary's goals, or what an adversary might try to do to a system. Threats to a system always exist, regardless of mitigation.
- *Threat Tree or Attack Tree*: An analysis tool that describes the attack paths for a particular threat. A threat tree is comprised of hierarchical conditions and allows the threat to be characterized. The root of the threat tree is the threat to which the tree corresponds.

The method of threat analysis given by Salter et al. (1998) was intended for external threats to systems. The steps in this method are:

1. Create attack trees for the system.
2. Apply weights to the leaves.
3. Prune the tree so that only exploitable leaves remain.
4. Generate corresponding countermeasures.
5. Optimize countermeasure options.

However, the above steps are oriented towards the development and implementation of systems. In this work, it is not necessary to optimize the countermeasures since we are not concerned with implementation. On the other hand, it is necessary to identify the threats before creating the attack trees. Thus, the above steps are modified to:

1. Identify threats on the user's private data.
2. Create attack trees for the provider's service.
3. Apply weights to the leaves.
4. Prune the tree so that only exploitable leaves remain. Count the number of such leaves or vulnerabilities (this gives the *n*).
5. Determine if countermeasures are in place for the vulnerabilities found in Step 4. Count the number of these vulnerabilities so mitigated (this gives the *q*).

A description of each step follows.

Step 1: Identify threats on the user's private data. This step requires experience and imagination and may involve confirming details with management or the developers of the service.

Examine the architecture and all available details of the service and enumerate possible outside threats on the user's private data. Represent the system pictorially to get the big picture. It is useful to identify the main or root threat which includes most if not all other threats, for then only one attack tree needs to be created. Disregard any existing provisions against outside threats; they will be accounted for in Step 5. For example, a possible outside threat for an online banking service is theft of private information from the bank's customer information database.

Step 2: Create attack trees for the provider's service. Corresponding to each treat identified in Step 1, systematically create an attack tree by putting yourself in the adversary's place in finding the weak points in the work processes or the service system and the paths which will lead to realizing the threat. This analysis terminates in a series of vulnerability leaves for each attack tree. (In this work, each attack tree is represented by hierarchical indented headings rather than pictorially, which can take up too much space and become unwieldy).

Step 3: Apply weights to the leaves. For each leaf, assign qualitative values (e.g., high, medium, low) for adversary risk, impediment to access, cost, and motivation (added for IV but applies to EV too). For example, an adversary sending an e-mail containing a virus attachment has low risk (probability of being identified is low), medium impediment to access (probability of the victim not opening the attachment and unleashing the virus is medium), low cost (cost to the adversary to create the virus e-mail is low), and high motivation (the adversary wants to cause as much disruption as possible). These values can be represented as a 4-tuple (L, M, L, H) where L, M, H stand for low, medium, high respectively, and the left most position in the tuple is risk, followed by impediment to access, followed by cost, and finally motivation. As another example, an adversary who is an insider with authorized access to private information and who wants to steal that information may be weighted as (L, L, L, H), that is, the adversary has low risk (probability of being caught is low), low impediment to access (adversary already has authorized access), low cost (cost to the adversary to make a copy of the information is low), and high motivation (the financial value of the information is very high). The provider with a vulnerability weighting of (L, L, L, H) has to think seriously about adding provisions to mitigate the vulnerability, as this weighting means that there is a very high probability that an attack using this vulnerability will occur.

Step 4: Prune the tree so that only exploitable leaves remain. Count the number of such leaves or vulnerabilities. Prune by deciding what levels (high, medium, low) of risk, impediment to access, cost, and motivation the provider is willing to have associated with the remaining vulnerabilities. These levels will determine which vulnerabilities are exploitable, and therefore add to the provider's cost to install countermeasures (provisions against privacy policy violations). Providers may choose to spend more or less on countermeasures by setting levels that result in less or more, respectively, leaves being pruned. For example, setting a level of (H, M, H, L) would prune all leaves with matching levels as well as all leaves that match (H, H, H, L) since (H, H, H, L) implies an even lower probability of attack than (H, M, H, L).

After pruning the tree, count the number *n* of exploitable leaves or vulnerabilities that remain.

Step 5: Determine if countermeasures are in place for the vulnerabilities found in Step 4. Count the number of these vulnerabilities so mitigated. Examine what countermeasures are in place for the vulnerabilities found in Step 4 and count the number of vulnerabilities *q* that have countermeasures. This step requires knowledge and experience of which countermeasures can be applied in a given situation.

Example Calculation of *n* and *q* Using Threat Analysis

Consider the automatic bank teller machine (ATM) that is ubiquitous in most shopping malls. There are many possible threats against an ATM but let us consider one threat: the threat of an adversary obtaining a bank customer's private account information for using an ATM. A possible attack tree for this threat is as follows:

1. Adversary obtaining a customer's account information for using an ATM
   - 1.1. Adversary holds up customer obtaining customer's access card and pin
   - 1.2. Adversary installs an ATM front end that secretly captures the customer's card info and pin
     - 1.2.1. The captured information is stored in the front end
     - 1.2.2. The captured information is transmitted to the adversary
   - 1.3. Adversary finds customer's access card and has a way of discovering the pin
     - 1.3.1. Adversary guesses the pin based on researching the customer
     - 1.3.2. Adversary uses a dictionary attack to discover the pin
     - 1.3.3. Adversary uses social engineering to obtain the pin

It can be seen that the root of this attack tree (1) contains the threat and the branches are attack paths that are alternative ways to realize the threat. For example, the path (1, 1.3, 1.3.3) is a legitimate attack path. The leaves of this tree are the nodes 1.1, 1.2.1, 1.2.2, 1.3.1, 1.3.2, and 1.3.3. As an example of applying weights to this tree, the leaf 1.1 has very high risk to the attacker, the impediments to access is low (customers with ATM cards are plentiful), the cost to the attacker to carry out the attack is low (e.g., fake gun), and the attacker's motivation is high (attacker is desperate for money). Because the risk to the attacker is very high, the path (1, 1.1) may be pruned from the tree. Similarly, the leaf 1.2.1 may be unlikely since the risk to the adversary is high (the front end could be discovered before the captured data could be retrieved). As well, the leaves 1.3.1 and 1.3.2 are infeasible due to the fact that the ATM captures the card and flags the account after a fixed number of pin entries. After pruning the unlikely attack paths from the tree, the tree that remains is as follows:

1. Adversary gaining access to customer's account information for using an ATM
   - 1.2. Adversary installs an ATM front end that secretly captures the customer's card info and pin
     - 1.2.2 The captured information is transmitted to the adversary
   - 1.3. Adversary finds customer's access card and has a way of discovering the pin
     - 1.3.3 Adversary uses social engineering to obtain the pin

Thus there are two likely attack paths: (1, 1.2, 1.2.2) and (1, 1.3, 1.3.3). A countermeasure for (1, 1.2, 1.2.2) would be to disallow any possibility of a front end being installed to an ATM, perhaps by physically redesigning the ATM or by frequent random inspections and monitoring. Suppose that this countermeasure is not yet in place. A countermeasure for (1, 1.3, 1.3.3) would be to strengthen procedures so that social engineering cannot succeed (e.g., no one is allowed to give out the customer's pin no matter what the circumstance). Suppose that this countermeasure is already in place. Then for this example, $e_e = q/n = 1/2$.

APPLICATION EXAMPLE

Consider a Web service, such as Easy123Drugs.com, that is an online drug store (e.g., Walgreens.com). Easy123Drugs is a multiprovider service that makes use of two business Web services:

an online payment service PayAsYouLikeIt.com (e.g., Paypal.com) and an accounting service AccountingAsNeeded.com (e.g., cbiz.com). Suppose Easy123Drugs, PayAsYouLikeIt, and AccountingAsNeeded (all fictitious names with no hits on Google) are all Web services that are based on the service-oriented architecture (SOA) (O'Neill et al., 2003), employing XML-based protocols (not necessarily the case for the real life examples cited here). Due to space limitations in this article, the details regarding UDDI lookup and service binding via SOAP and WSDL (O'Neill et al., 2003) will not be described here. It is assumed that these initialization steps occur as required. Figure 5 shows the network architecture of these services after service lookup and binding have occurred. The dashed lines in Figure 5 indicate logical communication channels.

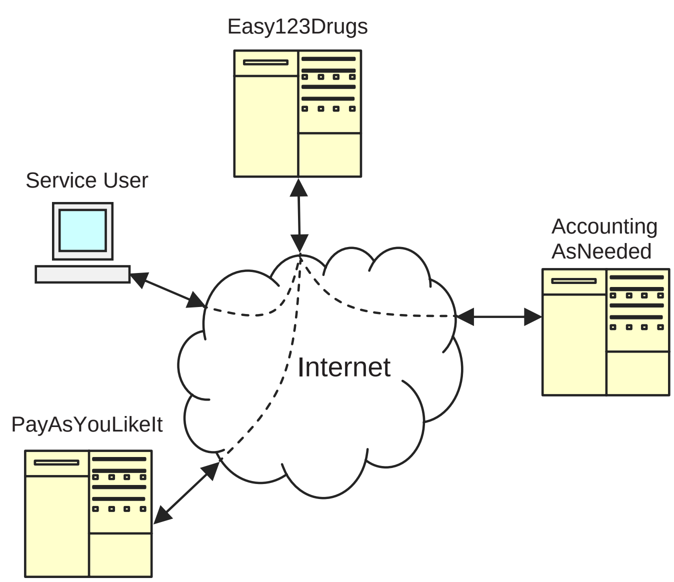
Table 3 shows the service user's private information required by each provider. The user provides required private information to Easy123Drugs once the user's privacy policy has been agreed to by Easy123Drugs. Easy123Drugs then discloses the user's private information to PayAsYouLikeIt and AccountingAsNeeded according to the user's privacy policy and after these second providers agreed with Easy123Drug's privacy policy (which contains the user's privacy preferences with regard to the information disclosed).

Easy123Drugs.com decides to hire a privacy auditor, certified to apply the above estimation methods, to estimate its capability of protecting privacy, with the intention of using the results in its advertising (assuming the results are good).

Calculation of *m* and *p*

To determine values for *m* and *p*, the auditor puts together a team to do a PIA for each service. Assume that each service stores the user's private data in a database and that the internal threats to the user's private data are about the same for each service. It is then possible to do one PIA that applies to all three services. However, the countermeasures in place are likely to be different for each service. The PIA team traces the flow of private information as shown by the PIP in Figure 6. In this figure, the customer's private

Figure 5. Network architecture of Easy123Drugs service



information (Table 3) arrives first at Easy123Drugs where it is received (first triangle), stored (first square), and processed, including forwarding some of the information to PayAsYouLikeIt and AccountingAsNeeded (second triangle). Similarly, selected information (Table 3) is received at PayAsYouLikeIt and AccountingAsNeeded where it is stored and processed.

The PIA team performs the PIA and uncovers IV vulnerabilities that fall under malicious attacks and unintentional disclosures, as follows:

Malicious attacks:

a. Attacker steals the data for sale to an outside interest.
b. Attacker uses the data for social engineering a personal goal.
c. Attacker passes the data to a friend as a favor.
d. Attacker passes the data to a competitor free of charge.

Unintentional disclosure:

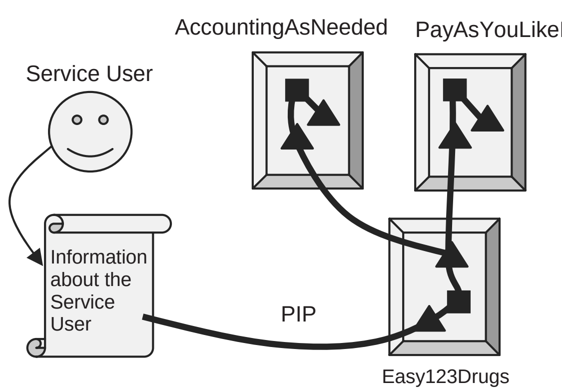
e. The data are inadvertently disclosed in an e-mail.
f. The data are inadvertently disclosed in a conversation.
g. A laptop containing the data is stolen.
h. Paper copies of the data are misplaced.
i. Paper copies of the data are left in a public area in plain view.
j. The data's retention time expires unknown to any staff.

For the malicious attacks, the PIA analysis considered the risks to the attacker and the attacker's motivation to carry out the attack. Due to the possibility that the source of the data could be traced, together with the fact that the data themselves are of relative low value (e.g., not military secrets), the risks to the attacker for a) and d) were considered very high and therefore these attacks are unlikely to occur. On the other hand, the risks to the attacker for b) and c) were considered low with the attacker's motivation high to medium since these attacks involve personal goals and relationships. Thus b) and c) were considered serious vulnerabilities that needed attention. Similarly, due to the fact that staff had undergone thorough training in safeguarding their laptops and paper copies, only vulnerabilities e), f), and j) were considered serious, requiring attention. Thus, the PIA identified five IV vulnerabilities that can be assigned to each provider as follows: Easy123Drugs gets the full $m=5$ vulnerabilities,

Table 3. Private information required

Web Service Provider	Private Information Required
Easy123Drugs	User's name, drug name and quantity, doctor's name and authorization, user's address
PayAsYouLikeIt	User's name, credit card details
AccountingAs Needed	User's name, drug name, doctor's name, quantity of drug sold, price paid by user, user's address

Figure 6. PIP for Easy123Drugs; squares indicate storage points, triangles indicate use



PayAsYouLikeIt gets *m=4* vulnerabilities since it has a retention time tracking mechanism, and AccountingAsNeeded gets *m=2* vulnerabilities since it has the retention time tracking and the remaining unintentional vulnerabilities do not apply to accountants because they are trained in safe data handling procedures.

Suppose the PIA found that Easy123Drugs and PayAsYouLikeIt have countermeasures in place against all vulnerabilities except for b) (the exact nature of the countermeasures is not important for this example). Suppose that AccountingAsNeeded has countermeasures in place against both of its vulnerabilities. Therefore, *p=4* for Easy123Drugs, *p=3* for PayAsYouLikeIt, and *p=2* for AccountingAsNeeded. Table 4 contains the values for *m* and *p.*

Calculation of *n* and *q*

The threat analysis described above is now applied to calculate *n* and *q* for each provider. Again, assume that each service stores the user's private data in a database. Assume also that the external threats to the user's private data are the same for each service. It is then possible to do one threat analysis that applies to all three services. However, the countermeasures in place are likely to be different for each service. Following the steps mentioned above,

Step 1: Using Figure 5 to visualize possible threats against the user's data, the main EV threat that includes most other EV threats is: "outside attacker compromises the user's private data."

Steps 2 and 3: The attack tree and weights are as follows.

1. Outside attacker compromises the user's private data.
   - 1.1. Attacker steals the user's private data.
     - 1.1.1. Attacker launches a man-in-the-middle attack on a communication channel to eavesdrop. (L, L, L, M)
     - 1.1.2. Attacker launches a Trojan horse attack on a provider's system. (L, L, L, M)
     - 1.1.3. Attacker launches a phishing attack on the user. (L, L, M, H)
     - 1.1.4. Attacker uses social engineering to deceive a provider staff member into giving out the user's data. (M, M, L, M)
     - 1.1.5. Attacker breaks into a provider's premises to steal the user's data. (H, H, M, H)
     - 1.1.6. Attacker mugs a provider employee and steals the employee's access card to enter a provider's premises and steal the user's data. (H, H, L, M)
   - 1.2. Attacker modifies the user's private data.
     - 1.2.1. Attacker launches a man-in-the-middle attack on a communication channel to modify the user's data. (L, L, L, M)

Table 4. Calculation of E_1 and E_2

Service	*m*	*p*	*n*	*q*	$E_1 = (p + q) / (m+ n)$	$e_i=p/m$	$e_e=q/n$	$E_2=(e_i, e_e)$ (%)
Easy123Drugs	5	4	7	6	.83	.80	.86	(80, 86)
PayAsYouLikeIt	4	3	6	5	.80	.75	.83	(75, 83)
AccountingAsNeeded	2	2	6	5	.88	1.0	.83	(100, 83)

1.2.2. Attacker launches a virus attack on a provider's system. (L, L, L, M)
1.2.3. Attacker uses social engineering to deceive a provider staff member into giving the attacker access to modify the user's data. (M, M, L, M)
1.2.4. Attacker breaks into a provider's premises to modify the user's data. (H, H, M, H)
1.2.5. Attacker mugs a provider employee and steals the employee's access card to enter a provider's premises and modify the user's data. (H, H, L, M)

Some of the reasoning behind the motivation weightings are: vulnerability 1.1.3 has motivation H as phishing is a quick way to obtain private data such as bank account information; vulnerabilities 1.1.5 and 1.2.4 have motivation H because breaking and entering is a serious crime and the attacker must be highly motivated before contemplating such an action.

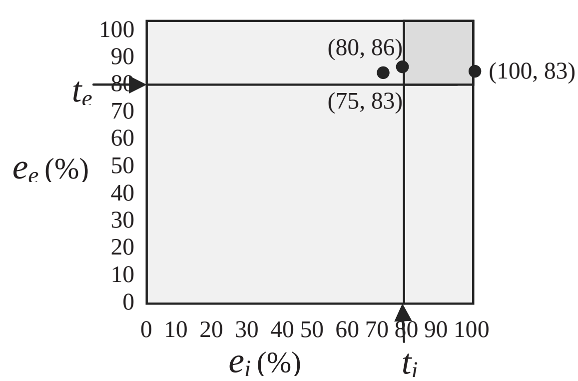
Step 4: The attack tree can be pruned by removing attack paths that are weighted with at least two Hs other than for motivation. Applying this criterion removes the attack paths (1, 1.1, 1.1.5), (1, 1.1, 1.1.6), (1, 1.2, 1.2.4), and (1, 1.2, 1.2.5). This leaves seven vulnerabilities that can be assigned to each provider as follows: Easy123Drugs gets the full $n=7$ vulnerabilities, PayAsYouLikeIt gets $n=6$ vulnerabilities since the phishing attack really only applies to Easy123Drugs, and AccountingAsNeeded gets $n=6$ vulnerabilities, again because the phishing attack does not apply to it. Note that the man-in-the-middle attack on a channel is double counted when it is considered a vulnerability for the provider at each end of the channel. However, this double counting is remedied by the countermeasure, which removes the vulnerability from both providers.

Step 5: Suppose that Easy123Drugs has countermeasures in place against all vulnerabilities except phishing (again, the exact nature of the countermeasures is not important here). Suppose also that PayAsYouLikeIt and AccountingAsNeeded have countermeasures in place against all vulnerabilities except social engineering. Therefore, $q=6$ for Easy123Drugs, $q=5$ for PayAsYouLikeIt, and $q=5$ for AccountingAsNeeded. Table 4 contains the values for n and q as well as the calculated results for E_1 and E_2.

Suppose the minimum acceptable threshold for E_1 is $t=85$. Then the results (Table 4) show that AccountingAsNeeded is the only provider that passes E_1 evaluation. The other providers need to add more provisions against IV or EV in order to pass. They may choose to add provisions that are easy to install or that are the least expensive. It can also be observed that AccountingAsNeeded is the most capable of protecting privacy whereas PayAsYouLikeIt is the least capable. However, comparing these providers to select which one to use based on E_1 is not feasible since their services are all different.

Plotting E_2 for minimum acceptable thresholds $t_i=80$ and $t_e=80$ according to Figure 4 gives Figure 7, which shows that each service passes E_2 evaluation except for PayAsYouLikeIt (with $E_2 = (75, 83)$).

Figure 7. Plots of E_2 for the example services



PayAsYouLikeIt is deficient in provisions for IV. For Easy123Drugs to pass E_2 evaluation as a first (i.e., multiprovider) service, PayAsYouLikeIt would have to add provisions against IV vulnerability b) above (see the chained rule above). Had thresholds t_i and t_e both been set to 90, no provider would pass. In this case, development would need to install a countermeasure against phishing, employees would need to be trained to resist social engineering, and provisions against IV would need to be added for providers that lacked these countermeasures. This shows that estimates of privacy protection capability can be used as a tool by development or management (e.g., to require social engineering training) in order to achieve predefined goals of privacy protection.

To consider the effects of past violations, suppose that Easy123Drugs had a disgruntled employee who 2 years ago passed private information to the news media that a certain famous actress was purchasing cancer drugs. Suppose that this made headlines and as a result, Easy123Drugs was sued by the actress and had to pay her 2 million dollars in damages. Easy123Drugs has since made policy changes to ensure that employees feel that they are treated fairly by the company, including the hiring of an ombudsman and a counselor. This past violation can thus be considered of high severity with some new measures installed post violation in order to avoid such violations in the future. Based on Easy123Drugs' response to this past violation, Table 1 shows *high support* (right-most column under profit-oriented service provider) for the provider's future privacy protection capability. In other words, Easy123Drugs will be highly motivated to make sure that it is capable of protecting privacy.

RELATED WORK

The literature appears empty on works dealing directly with estimates of a service provider's capability to protect privacy. Only works that are indirectly related were found. These refer to the economics of security or the economics of privacy (see http://www.cl.cam.ac.uk/~rja14/econsec.html - available as of May 6, 2006). These authors hold the view that the lack of security or the lack of privacy are not due to the lack of technological solutions but rather are due to other (perhaps perverse) considerations such as economics and profitability. This work differs from that view in that the proposed approach evaluates the capability to protect privacy by counting provisions against privacy violations that ARE in place, NOT WHY they may or may not be in place. Nevertheless, their view is valuable in understanding how to improve privacy protection, and is thus seen as complementary to this work. Another area that is indirectly related to this work concerns privacy audits (Enright, n.d.) and privacy risk (or impact) assessment (or analysis) (Treasury Board of Canada, n.d.). As explained above, the latter consists of methods or guidelines on how to identify vulnerabilities or risks in managing private information in order to comply with privacy legislation. We applied a privacy impact assessment to find IV vulnerabilities. In general, these and other methods in the privacy audit domain can be introduced into this work to better understand risks that lead to IV and EV. For example, they could help identify a new class of privacy vulnerabilities. Alternatively, this work could be used in a privacy audit to obtain more comprehensive results, but further research in this regard is needed. Of course, the area of threat analysis is indirectly related to this work (although we applied threat analysis, it is still only indirectly related as a means to an end). Threat analysis has been used for many years and several forms of it exist. The different forms can differ in how weighting is done (as this work too has introduced motivation into the weighting), how a threat is defined, how many people should be involved in carrying out a threat analysis, how much detail is recorded, and so on. Other authors who have recently written on threat analysis include Karger (2006) and

Rippon (2006). An older reference for basic threat analysis is Bauer (2002). Finally, other indirectly related work consists of the entire area of privacy enhancing technologies, some of which were mentioned above as provisions against IV. Also in this category is the work allowing Web sites to specify their privacy policies using platform for privacy preferences (P3P) (W3C, n.d.) to allow for automatic user agent interpretation of a privacy policy, and the machine readable privacy policy languages of P3P preference exchange language (APPEL) (W3C, 2002) and enterprise privacy authorization language (EPAL) (IBM, 2003).

EVALUATION OF APPROACH

Since, as far as can be determined, this work is new, an evaluation of the proposed approach by direct comparisons to other similar works is not possible. Therefore, this evaluation is conducted by considering how well the approach can accomplish its stated purpose by considering in turn the suitability of each component of the approach or how well a component can perform its function.

The goal of the proposed approach is to evaluate the privacy protection capability of a service provider. This is achieved by considering a provider's installed provisions against violations of privacy policy and the provider's history of past violations, if any. It was postulated that a past violation under the "right" circumstances (i.e., profit-oriented provider, sizable violation known to the public, additional provisions to prevent future violations installed after the past violation) would motivate the provider to improve its capability to protect privacy. The approach consists of three components: i) a model of privacy protection in terms of privacy policy violations, ii) definition of estimates of a provider's capability to protect privacy, and iii) methods for calculating the estimates.

The model of privacy protection is based on preventing internal and external violations against privacy policies. This model requires that personal privacy policies exist in a provider organization and that the provider has agreed to uphold them. This requirement is reasonable since most online organizations today have their own privacy policies. Today, personal privacy policies are in the minority when compared to organizational privacy policies, but personal privacy policies should increase with time, in line with the increasing demand for organizations to respect personal privacy preferences.

The definitions of the estimates appear to be straightforward, and follow the model of privacy protection in terms of counting provisions against violations. Moreover, they provide not only a useful single number (E_1) comparison between different providers of similar services, but also an easy graphical visualization (by plotting E_2) of where a provider stands in terms of its capability to protect against IV and EV separately (Figure 4). The advantage of straight forward estimates should not be underestimated, since they need to be understandable by the general public, and "the simpler, the better."

The proposed methods for calculating the estimates naturally have strengths and weaknesses. Strengths include: a) both PIA and threat analysis are well known and have been practiced for years; b) both are well documented in the literature; and c) PIA is an approach recommended by privacy authorities. Weaknesses include: a) PIA is a rather long and involved procedure requiring management support; b) the questionnaire for identifying vulnerabilities that could lead to malicious IV requires testing to confirm its validity; c) PIA results may depend on the skill and knowledge of the people involved with doing the PIA; d) the results from threat analysis may depend on the skill and knowledge of the threat analyst; and e) in the threat analysis, the subjective weighting of the leaves and the pruning criterion are not as exact as one would like (the pruning criterion depends partly on how much money the organization is willing to spend on countermeasures, and links back to the economics of security related work mentioned above). Despite the weaknesses of the threat analysis, it is an accepted procedure and has been applied in many situations to assess threats. The threat analysis procedure may be improved with further research, or the application of other methods (such as privacy audit methods mentioned above) can be incorporated. To help mitigate weaknesses a), c), d), and to some extent e), we suggest that the methods be applied by a separate, impartial firm specialized in performing PIA and threat analysis. To go further, guidelines on applying the methods could be standardized by a privacy authority, and only firms certified by the same authority may apply the methods. This would ensure that the calculation of the estimates of capability to protect privacy is done fairly and consistently for each provider. This approach of using certified third parties to evaluate something is not new; it has been applied successfully to determine conformance to the ISO 9000 series of standards (International Organization for Standardization, n.d.) as well as to evaluating ratings relative to the capability maturity model integration (CMMI) (Carnegie Mellon Software Engineering Institute, n.d.) for software producers.

The accuracy of the proposed estimates may be another point of contention. It may be argued that a simple count of provisions against IV or EV is not enough; actual protection of privacy depends on the *effectiveness* of these provisions. The author agrees that if effectiveness could be incorporated into the estimates, the results would be more accurate. However, effectiveness itself depends on many factors, such as the capability of the provision to do what it is supposed to do, how the provision is implemented, the environment in which the provision operates, and so on. The decision was made to avoid these complexities, opting instead for straightforward easy-to-calculate estimates. The author believes that this is reasonable, trusting that providers would not throw money away on ineffective provisions.

The idea that under the “right” circumstances (see second paragraph in this section), past violations would motivate a provider to protect privacy in the future is untested, as noted above under “Past Violations.” Nevertheless, the idea seems intuitively reasonable. Testing this idea is left for future work.

Finally, it should be pointed out that a provider that is estimated to have the capability to protect privacy may still not do so in reality. For instance, this may happen if key people responsible for maintaining privacy enhancing technologies leave the provider. Another reason may be that the provider perceives goals other than privacy protection as having more immediate higher priority, such as spending money on new production capabilities rather than hiring people needed to maintain privacy enhancing technologies. Still a third reason may be that the provider has undergone a restructuring (e.g., acquired by another provider) and the new people in charge have different priorities. Incorporating such almost always unforeseeable changes into a method for likelihood estimation of privacy policy compliance would be very complicated and difficult. Nevertheless, it may be assumed that in general, providers that have the capability to protect privacy will do so, especially if failure to protect privacy means stiff legal penalties. Stiff legal penalties for failure to protect privacy have already been implemented (U.S. Government, n.d.).

In summary, the author believes that the approach proposed in this work is a reasonable first attempt at evaluating the capability of a provider to protect user privacy. Key challenges regarding the potentially subjective nature of the threat analysis may be reduced with further research.

CONCLUSION AND FUTURE RESEARCH

This work has proposed estimates for evaluating a provider’s capability to protect privacy and illustrated the calculation of the estimates using an example of a multiprovider service. The estimates serve at least four important functions: 1) they make it possible for providers to be challenged if their capability for protecting privacy is perceived to be inadequate; 2) they allow for enforceable privacy protection legislation requiring providers to ensure that they can meet privacy policy compliance requirements; 3) they allow customers to compare providers in terms of their capability to protect privacy when deciding which provider to use for a particular service; and 4) they enable the providers themselves to improve their services by showing them i) where they stand in terms of privacy protection capability against IV and EV and ii) what provisions against IV or EV they need to add in order to improve their standing.

It is envisioned that providers will want to advertise their estimates to show that they exceed standard privacy protection capability thresholds (which could be standardized by an international body) in the same way that they advertise conformance to ISO 9000. This could encourage providers to achieve higher levels of privacy protection, which in turn could lead to greater public trust in Web service providers resulting in increased commerce.

Future research includes improving the methods for calculating the estimates, such as increasing the effectiveness of the procedure for threat analysis by automating it and making it more foolproof, as well as investigating other possible estimates of capability to protect privacy.

ACKNOWLEDGMENT

The author gratefully acknowledges the support of the National Research Council Canada for this work.

REFERENCES

Adams, C., & Barbieri, K. (2006). Privacy enforcement in e-services environments. In G. Yee (Ed.), *Privacy protection for e-services*. Hershey, PA: Idea Group, Inc.

Bauer, M. (2002). Practical threat analysis and risk management. *Linux Journal*, *2002*(93), 9. Retrieved November 7, 2008, from http://www.linuxjournal.com/article/5567

Carnegie Mellon Software Engineering Institute. (n.d.). *Welcome to the CMMI Website*. Retrieved November 9, 2006, from http://www.sei.cmu.edu/cmmi/cmmi.html

Enright, K. P. (n.d.). *Privacy audit checklist*. Retrieved May 6, 2006, from http://cyber.law.harvard.edu/clinical/privacyaudit.html

Goldberg, I., Wagner, D., & Brewer, E. (1997). Privacy-enhancing technologies for the Internet. In *Proceedings of the IEEE COMPCON'97* (pp. 103-109).

IBM. (2003, June 12). *The enterprise privacy authorization language (EPAL 1.1)*. Retrieved June 2, 2007, http://www.zurich.ibm.com/security/enterprise-privacy/epal/

International Organization for Standardization. (n.d.). *Selection and use of the ISO 9000:2000 family of standards*. Retrieved January 28, 2006, from http://www.iso.org/iso/en/iso9000-14000/understand/selection_use/selection_use.html

Iyengar, V. S. (2002). Transforming data to satisfy privacy constraints. In *Proceedings of the SIGKDD'02*, Edmonton, Alberta (pp. 279-288).

Karger, P. A. (2006, July). Privacy and security threat analysis of the federal employee personal identity verification (PIV) program. In *Proceedings of the Second Symposium on Usable Privacy and Security*, Pittsburgh, Pennsylvania (pp. 114-121).

Kenny, S., & Korba, L. (2002, November). Adapting digital rights management to privacy rights management. *Computers & Security*, *21*(7), 648-664.

Kobsa, A., & Schreck, J. (2003, May). Privacy through pseudonymity in user-adaptive systems. *ACM Transactions in Internet Technology*, *3*(2), 149-183.

Lategan, F., & Olivier, M. (n.d.). *PrivGuard: A model to protect private information based on its usage*. Retrieved December 14, 2005, from http://mo.co.za/open/privgrd.pdf

O'Neill, M., Hallam-Baker, P., MacCann, S., Shema, M., Simon, E., Watters, P. A., et al. (2003). *Web services security*. McGraw-Hill/Osborne.

Rippon, W. J. (2006, April). Threat assessment of IP based voice systems. In *Proceedings of the 1st IEEE Workshop on VoIP Management and Security 2006*, Vancouver, B.C., Canada (pp. 19-28).

Salter, C., Saydjari, O. S., Schneier, B., & Wallner, J. (1998, September). Towards a secure system engineering methodology. In *Proceedings of New Security Paradigms Workshop* (pp. 2-10).

Song, R., Korba, L., & Yee, G. (2006). Pseudonym technology for e-services. In G. Yee (Ed.), *Privacy protection for e-services*. Hershey, PA: Idea Group, Inc.

Treasury Board of Canada. (n.d.). *The privacy impact assessment guidelines: A framework to manage privacy risk*. Retrieved May 6, 2006, from http://www.tbs-sct.gc.ca/pgol-pged/piatp-pfefvp/course1/mod2/mod2-5_e.asp

U.S. Government. (n.d.). *General overview of standards for privacy of individually identifiable health information*. Retrieved October 19, 2006, from http://www.hhs.gov/ocr/hipaa/guidelines/overview.pdf

W3C. (n.d.). *Platform for privacy preferences (P3P) project*. Retrieved November 9, 2006, from http://www.w3.org/P3P/

W3C. (2002, April 15). *A P3P preference exchange language 1.0 (APPEL1.0)* (W3C Working Draft). Retrieved November 9, 2006, from http://www.w3.org/TR/P3P-preferences/

Yee, G. (2006, September 18-22). Measuring privacy protection in Web services. In *Proceedings of the 2006 IEEE International Conference on Web Services (ICWS 2006)*, Chicago (pp. 647-654).

Yee, G., & Korba, L. (2003a, May 18-21). *The negotiation of privacy policies in distance education*. Paper presented at the 14th IRMA International Conference, Philadelphia.

Yee, G., & Korba, L. (2003b, January 27-31). Bilateral e-services negotiation under uncertainty. In *Proceedings of the 2003 International Symposium on Applications and the Internet (SAINT2003)*, Orlando, Florida (pp. 352-355).

Yee, G., & Korba, L. (2004, July 6-9). Privacy policy compliance for Web services. In *Proceedings of the 2004 IEEE International Conference on Web Services (ICWS 2004)*, San Diego (pp. 158-165).

Yee, G., & Korba, L. (2005). Semi-automatic derivation and use of personal privacy policies in e-business. *International Journal of E-Business Research*, *1*(1), 54-69. IGI Global, Inc.

ENDNOTE

[1] NRC Paper Number: NRC 50725. This article is a significant extension of Yee (2006).

This work was previously published in the International Journal of Web Services Research, Vol. 6, Issue 2, edited by L. Zhang, pp. 20-41, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 7.7
Privacy Concerns for Web Logging Data

Kirstie Hawkey
University of British Columbia, Canada

ABSTRACT

This chapter examines two aspects of privacy concerns that must be considered when conducting studies that include the collection of Web logging data. After providing background about privacy concerns, we first address the standard privacy issues when dealing with participant data. These include privacy implications of releasing data, methods of safeguarding data, and issues encountered with re-use of data. Second, the impact of data collection techniques on a researcher's ability to capture natural user behaviors is discussed. Key recommendations are offered about how to enhance participant privacy when collecting Web logging data so as to encourage these natural behaviors. The author hopes that understanding the privacy issues associated with the logging of user actions on the Web will assist researchers as they evaluate the tradeoffs inherent between the type of logging conducted, the richness of the data gathered, and the naturalness of captured user behavior.

INTRODUCTION

Privacy is an important consideration when conducting research that utilizes Web logs for the capture and analysis of user behaviors. Two aspects of privacy will be discussed in this chapter. First, it is important that governmental regulations, such as the Personal Information Protection and Electronic Documents Act (PIPEDA) in Canada, or organizational regulations, such as a university's local research ethics board (REB) policies, are met. These regulations will dictate requirements for the storage and safeguarding of participant data as well as the use, re-use, and transfer of that data. Secondly, researchers may also find that providing privacy enhancing mechanisms for participants can impact the success of a study. Privacy assurances can ease study recruitment and encourage natural Web browsing behaviors. This is particularly important when capturing rich behavioral data beyond that which is ordinarily recorded in server transaction logs, as is generally the case for client-side logging. It is this second aspect of privacy that will be the primary focus of this chapter.

DOI: 10.4018/978-1-59904-974-8.ch005

There are privacy concerns associated with viewing and releasing Web browsing data. Web browsers are typically used for a wide variety of tasks, both personal and work related (Hawkey & Inkpen, 2006a). The potentially sensitive information that may be visible within Web browsers and in data logs is tightly integrated with a person's actions within the Web browser (Lederer, Hong, Dey, & Landay, 2004). Increasingly the Internet has become a mechanism by which people can engage in activities to support their emotional needs such as surfing the Web, visiting personal support forums, blogging, and investigating health concerns (Westin, 2003). Content captured within Web browsers or on server logs may therefore include such sensitive items as socially inappropriate activities, confidential business items, and personal activities conducted on company time, as well as more neutral items such as situation-appropriate content (e.g., weather information). Visual privacy issues have been investigated with respect to traces of prior Web browsing activity visible within Web browsers during co-located collaboration (Hawkey, 2007; Hawkey & Inkpen, 2006b). Dispositional variables, such as age, computer experience, and inherent privacy concerns, combine with situational variables, such as device and location, to create contextual privacy concerns. Within each location, the social norms and Web usage policies, role of the person, and potential viewers of the display and users of the device impact both the Web browsing behaviors and privacy comfort levels in a given situation. The impacted Web browsing behaviors include both the Web sites visited, as well as convenience feature usage such as history settings and auto completes. Furthermore, most participants reported taking actions to further limit which traces are potentially visible if given advanced warning of collaboration.

Recently the sensitivity of search terms has been a topic in the mainstream news. In August 2006, AOL released the search terms used by 658,000 anonymous users over a three month period (McCullagh, 2006). These search terms revealed a great deal about the interests of AOL's users, and their release was considered to be a privacy violation. Even though only a few of the users were able to be identified by combining information found within the search terms they used, AOL soon removed the data from public access. This data highlighted the breadth of search terms with respect to content sensitivity as well as how much the terms could reveal about the users in terms of their concerns and personal activities.

In addition to taking actions to guard visual privacy within Web browsers, users may also take steps to guard the transmission of their personal information online. When concerned about privacy as they interact on the Web, users may opt to mask their identities by using a proxy server or other anonymizing (Cranor, 1999). The Platform for Privacy Preferences Project (www.w3.org/P3P/) has developed standards that facilitate user awareness of the privacy policies that govern the use of their personal information at participating websites. Research into online privacy generally examines issues concerning the transfer of personal data to business or governmental entities; the relationships are between consumers and corporations. This may be quite different from the privacy concerns associated with others viewing traces of previous Web browsing activity, as in the case of logged Web browsing data in a research context. Although in both cases personal information may be viewed, there are differences in the nature of the relationship to the viewer of the information. When the viewers of the captured information are not anonymous but are known to the user, privacy concerns may be heightened (Lederer, Mankoff, & Dey, 2003).

Field research theoretically allows the study of actual behaviors in a realistic environment. However, the act of observing or recording participants' personal interactions may cause them to alter those behaviors (McGrath, 1995). This is often referred to as the Hawthorne Effect. For example, behaviors deemed to be socially inappropriate (Fisher,

1993) may be avoided during the period of the study. As well, participants may be unwilling to have logging software installed that may record personal interactions, particularly if that software logs data across applications. Software (e.g., a keystroke logger, or custom web browser) that has the potential of capturing user names and passwords may cause additional concerns (Weinreich, Obendorf, Herder, & Mayer, 2006). Privacy preserving mechanisms can help encourage participants to engage in their natural Web browsing behaviors and activities while allowing researchers to study the behaviors of interest. Appropriate methods of mitigating participants' privacy concerns depend on the research questions and the experimental logging environment in use.

The objectives of this chapter are to provide researchers with an understanding of the privacy issues associated with the logging of Web activity. Background will be provided in the areas of privacy theory in general and privacy concerns for Web browsing data in particular. It is important that privacy concerns are understood so that observational effects on behavior can be reduced during studies. Furthermore, the tradeoffs between participants' privacy and the collection of rich, yet natural data for various logging techniques will be discussed. Finally, guidelines for mitigating participants' privacy concerns during studies investigating Web behaviors will be presented.

BACKGROUND

General Privacy Theory

Westin (2003) defines individual privacy as "the claim of an individual to determine what information about himself or herself should be known to others." Over the past forty years, Westin has primarily dealt with consumer privacy rights, such as when personal information can be collected and how others can make use of the information. Westin also discusses how individuals seek a balance between maintaining privacy and fulfilling a need for communication and disclosure. How an individual manages this tradeoff depends on their personal situation including their family life, education, social class, and psychological composition. Furthermore, Westin states that an individual's privacy needs are highly contextual and continually shift depending on situational events.

This contextual nature of privacy is well established in the literature. Goffman (1959) first introduced the need to project different personas or faces during social interactions. The face presented in any given situation depends not only on the current audience but also on the current conditions. The combination of audience and situation determines how much and what information will be disclosed. Furthermore, as discussed by Palen and Dourish (2003), people can have many roles between which they fluidly move and can act in multiple capacities, often simultaneously. For example, one may act as an individual, a family member, and a representative of an organization. A person's role can influence their sense as to whether their behaviors would be considered socially acceptable. If information is conveyed that is out of character for the person's current role, the boundaries that have been maintained can collapse creating opportunities for social, bodily, emotional, and financial harm (Phillips, 2002). Lederer et al. (2003) discuss how activities convey the essence of a persona. Knowledge of an individual's prior activities is more sensitive when their identity is known as the activities can reveal hidden personae.

Privacy Concerns for Web Browsing Data

Web users conduct a wide range of activities within their Web browsers, resulting in visited web pages with a variety of content sensitivity (Hawkey & Inkpen, 2006a). Teltzrow and Kobsa (2004) summarized thirty published consumer surveys and studies investigating Internet privacy. Results consistently revealed that the majority of Internet users are concerned about the security of personal information as well as concerned about being tracked on the Internet, with a lesser amount being concerned that someone might know what websites they visited. Two field studies have specifically examined visual privacy concerns for visited Web pages. For the first study, conducted in 2004 (Hawkey & Inkpen, 2005), 42% of visited pages were classified as public (suitable for anybody to view), 25% as semi-public (suitable for a subset of viewers), 15% as private (suitable perhaps only for a close confident), and 18% as don't save (either irrelevant or extremely private). Similar results were found in the second study, conducted in 2005 (Hawkey & Inkpen, 2006a): 40% public, 20% semi-public, 25% private, 15% don't save. It must be noted that participants in both studies exhibited a great deal of individual variability in their privacy classifications with some participants having greater privacy concerns than others. This variability is both as a result of participants having differing privacy concerns for similar content and as a result of them having conducted browsing activities of differing sensitivity.

Studies have found that privacy concerns are highly nuanced and individual (Ackerman, Cranor, & Reagle, 1999; Hawkey & Inkpen, 2006a). Recent information sharing research has investigated privacy concerns for various types of information and recipients of that information. For example, one study investigated privacy comfort for participants when sharing information with a recipient (Olson, Grudin, & Horvitz, 2005). Privacy concerns differed depending on the person's relationship to the receiver of the information as well as on the type of information being shared. Their results suggest that some of the types of information that may be revealed in Web logs, such as personal activities like viewing non-work related websites and transgressions like viewing erotic material, are considered more sensitive than information such as contact and availability information. The amount of control that the individual retains over the disclosure of information may also impact their level of comfort (Palen & Dourish, 2003).

A person's demographics such as age and gender may affect their privacy disposition (Hawkey, 2007). However, a person's disposition to privacy, that is, their inherent privacy concern, is also grounded in their life experience. For example, their technical level or computer experience may impact their inherent privacy concerns. Additionally, dispositional variables may moderate the effect of situational variables. Someone with strong inherent privacy concerns may always be very private, someone with weak concerns may be less private, others may be more pragmatic and may more often modify their privacy comfort and browsing activities in response to the state of the environment (Hawkey, 2007; P&AB, 2003).

While inherent privacy concerns indicate someone's overall privacy preferences, the situational context will determine which information a person feels is appropriate to reveal (Joinson, Paine, Reips, & Buchanan, 2006; Westin, 2003). For example, in a study examining online disclosure of information, independent pathways were found for the dispositional variable of participant's general privacy concerns as well as the situational variables of perceived privacy (in terms of anonymity and confidentiality) and participants' trust in the receiver of the information (Joinson, Paine, Reips, & Buchanan, 2006). Similarly Malhotra et al. (2004) developed a causal model of online consumers' information privacy concerns. Their model considered the effect that Internet users' information privacy concerns have on trusting beliefs, risk beliefs, and their behavioral intention to reveal personal information. Furthermore, they incorporated the sensitivity of the information requested by marketers as a contextual variable and considered covariates such as sex, age, education, Internet experience, identity misrepresentation, past experiences with privacy invasion, and media exposure. They developed measures for new factors of privacy concerns including control (i.e., whether the user has control over the data) and awareness (i.e., whether the user is adequately informed as to use of the data) to augment existing scales for this domain which consider collection of information such as whether the exchange of personal information is equitable.

Privacy comfort for the viewing of Web browsing activity has also been found to depend not only on a person's disposition to privacy, but also on the situational context when the activity is revealed (Hawkey, 2007). Situational variables for privacy concerns associated with traces of activity in Web browsers include the computing device used and the location of use. Furthermore, within each location there may be other variables such as the current role of the user, social norms for the location, rules for personal Web browsing activities, and different types of viewers of the display and users of the device. These variables may constrain or shape both the browsing activities and the subsequent privacy concerns. For example, someone with Web access on both a home and a work computer may refrain from conducting many personal activities while at work, while someone with only access at work may conduct a broader range of activities in the workplace. A laptop user may perform the majority of their browsing activities on their laptop, but their viewing concerns may change as they move between different locations with different social norms. One's browser settings and preventative actions taken may also change depending on the usage environment. Beyond which traces are potentially visible as a result of these changes, the perceived sensitivity of the traces may also change as a result of the viewing situation. The cost and benefit of disclosure depends on the specifics of each situation (Joinson, Paine, Reips, & Buchanan, 2006).

Marx (2003) identified several privacy enhancing methods that people use when under surveillance, with *self-regulating*, *blocking*, *masking*, *switching*, and *refusal* activities being particularly applicable to mitigating privacy concerns associated with Web browsing data. For example, Web browsing activities may be *self-regulated* in the workplace to avoid surveillance by an employer, with more personal activities being conducted solely at home (Hawkey & Inkpen, 2006b). A person's attitudes and perceptions about privacy, trust, and social relationships or norms (e.g., workplace rules) will influence his behavior in a situation (Liu, Marchewka, Lu, & Yu, 2004). A common privacy preserving strategy employed within Web browsers is to *block* the recording of visited sites by turning off the convenience features such as history files and auto complete data (Hawkey, 2007). One downside to this approach is that a complete lack of visited sites within the browser's history files may be viewed as an indicator that there is an activity worth hiding. A more subtle approach would be to *mask* the activity rather than to block it completely (Marx, 2003). For example, to mask browsing activities in their personal bookmarks, users can rename stored sites to conceal the nature of the page (Hawkey, 2007). In order to guard privacy at the server level, users may opt to anonymize their browsing, thereby masking their identity (Cranor, 1999). Internet users in the studies surveyed by Teltzrow and Kobsa (2004) have taken steps such as *refusing* to give personal information to a Website and supplying false information to a Website when asked to register. S*witching* computers or browser applications to avoid logging software is a privacy enhancing mechanism that can impact the breadth of data recorded during studies (Kellar, Hawkey, Inkpen, & Watters, 2008). Finally, *refusing* to take part in studies altogether may also occur if the privacy concerns are too high (Tang, Liu, Muller, Lin, & Drews, 2006).

PRIVACY CHALLENGES ASSOCIATED WITH THE LOGGING OF WEB DATA

Much of the privacy background just presented was focused on the privacy concerns associated with the types of data that may be captured in Web logs. In this section, two facets of privacy challenges associated with Web logging data are discussed. The first are standard privacy concerns with respect to the capture, storage, transfer, and re-use of data. These are largely dictated by governmental and organizational regulations. The second are privacy concerns that participants may have about their activities being recorded. These concerns may affect their natural Web browsing behaviors during the study period and can be challenging to address.

Governmental and Organizational Regulations

The first concern when designing a study with Web log analysis is ensuring that governmental regulations (e.g., PIPEDA in Canada) or organizational regulations, such as a university's Research Ethics Board (REB) policies, with respect to privacy are met. These regulations will specify requirements for data collection including the storage and safeguarding of participant data as well as the use, re-use, and transfer of that data. As these regulations are specific to the country and institution where the research is located, they will not be extensively described here. However, some general areas for consideration will be presented. It is up to individual researchers to ensure that they are in compliance with the policies that govern their research.

Many REB and governmental policies address the period of time that data may be kept and the storage requirements for that data. In addition, data re-use may be limited to the purposes identified in the study materials and agreed to by participants. While it may be tempting to provide very broad potential use cases, more narrow usage possibilities may assuage participant concerns about the capture of what can be potentially sensitive data (Teltzrow & Kobsa, 2004).

Governmental regulations may even dictate which data logging software is used. For example, in Nova Scotia, Canada, the Personal Information International Disclosure Protection Act has recently been approved by the Nova Scotia provincial government (Dalhousie Research Services, 2006). This legislation deals with protection, storage, and management of personal information of Nova Scotians, and the issue of data transfer outside Canada is prominent. Special approval is required to use software, hardware, or services that store personal information of Nova Scotians outside of Canada, and permission must be granted to transfer data containing personal information to researchers outside of Canada.

Governmental regulations will likely apply to the storage and use of the data, although requirements may be lessened if the data is anonymized. Data collection itself may be anonymous (i.e., collected with no associated identifying information) or the data set may be anonymized through removal of any links between the data and identifying information. To be considered anonymized, there must be no way for an investigator to connect the data with a specific participant. This can be difficult with small data sets due to the potential triangulation of the data to a specific individual in the study population. Care must be taken that the data does not include potentially identifying information such as highly detailed demographic information or IP addresses (Dalhousie Research Services, 2006).

Depending on where the log data is captured (i.e., server-side, client-side) and the frequency with which the data needs to be transferred between the participant and the researchers, different security mechanisms are required to safeguard the data and ensure that participants' privacy is not inadvertently compromised. While discussion of security mechanisms is outside the scope of this chapter, there are several resources that may be useful (Garfinkel & Spafford, 2001; Huseby, 2004; Meier et al., 2003). The discussion in this section is limited to the tradeoffs inherent with different approaches.

When possible, researchers should take advantage of opportunities of anonymizing or otherwise transforming the data before receiving it. For example, with client-side storage of data, a data collection script can remove any identifying information such as IP addresses that may be stored in the data logs and assign a random user ID number that is not tied to recruitment or screening data. Furthermore, potentially identifying or sensitive information can be transformed into higher level data. For example, if a study would like to record where laptop users accessed the internet, a data collection script could take as input IP addresses and location labels and replace the personally identifying IP addresses in the data records with a general location field (whether home, work, or school) (Hawkey, 2007).

One question that arises during research that makes use of Web logs is where to store the data, and when and how to transfer the data between participant and researcher (Kellar, Hawkey, Inkpen, & Watters, 2008). When data is logged during a laboratory experiment, or with proxy or server-side logging applications, it is typically stored directly on a research computer. With proxy logging, researchers should provide a secure connection to the proxy server. Additional complexities arise for client-side data logging as data transfer and storage issues must be determined. Storing the data locally on the participant's machine for the duration of the study (and removing it physically during an uninstall session) may simplify the participants' duties and minimize privacy risks associated with the transfer of data; however, researchers run the risk of data loss if the participant's machine crashes. If data is transferred more frequently, the participant may be inconvenienced and there is a need to provide secure methods of transmission.

Impact of Privacy Concerns for Data Collection on Natural Web Browsing Behaviors

The remainder of this chapter will deal with the impact of privacy concerns on the ability of studies to capture natural Web browsing behaviors. It is important to consider that the act of recording visited sites may impact participants' normal Web browsing activity (McGrath, 1995). As previously discussed, there are privacy concerns associated with others viewing visited websites (Hawkey & Inkpen, 2006b; Olson, Grudin, & Horvitz, 2005; Teltzrow & Kobsa, 2004). Self-regulation of activity is one mechanism used to preserve privacy when under surveillance (Marx, 2003). However, for most studies involving the logging of Web data, it is important that participants conduct their Web-related activities as they normally would, regardless of the social desirability of the content (Fisher, 1993) or the personal information that may be captured.

Which traces of prior activity may be disclosed depend on the type of data logging being done. There are several challenges and tradeoffs when trying to capture rich contextual data (Kellar, Hawkey, Inkpen, & Watters, 2008). One key factor in determining an appropriate data logging strategy is the tradeoff between the amount of control the researcher retains and the amount of intrusiveness for the participant (McGrath, 1995). This chapter considers this tradeoff, extending the discussion of the impact on privacy concerns and the ability to capture natural Web browsing behavior for various Web logging strategies.

One approach is to use trace measures or archival records (McGrath, 1995). Archival records are records of user behavior that are collected for other purposes and may either be private or public knowledge. Examples of archival records include blogs or stored bookmarks in the Web browser. Trace measures are records of behavior inadvertently left by participants, such as Web server data logs created through server-side logging. If data is gathered after the fact, there will be no behavioral changes due to observation. However, there are several drawbacks to this approach, and the available data may not be appropriate depending on the research questions of interest.

Cockburn and McKenzie (2001) used archival data to conduct an empirical analysis of Web page revisitation. They analyzed the history records from academic user accounts captured on server backups. History records are stored within a Web browser to enable revisitation of previously visited sites. An advantage to their approach was that there were no behavioral changes due to observation as the participants were unaware at the time of web browsing that their browsing activities would be examined as part of a study. However, the data available was not complete. The history files only included the most recent timestamp for accessing a URL, so some visits were not captured temporally. Data collected in this fashion is generally limited in contextual information about the activities underway. Additionally, it may be difficult to get permission to use archival data if the participant is unclear about which sensitive activities may have been conducted during the study period.

The use of trace measures such as server logs will similarly remove behavioral changes due to observation. However, server-side logging generally limits the breadth of the data collected, either capturing only the access to a single website or access through a specific Web portal (Yun, Ford, Hawkins, Pingree, & McTavish, 2006). The data is also usually limited to the IP address of users, a time stamp, and the URL requested. Web server logs may be incomplete records of an activity, since page requests may not be received and recorded at the server if the page has been cached by the browser or a proxy server (Fenstermacher & Ginsburg, 2003). There is ongoing tension between Web users' privacy needs and a website's requirement for information about its users (Cooley, Mobasher, & Srivastava, 1999). While the use of cookies can alleviate problems of identifying returning individuals that are associated with dynamic IP addresses (Anick, 2003), users may turn off cookies in order to protect their privacy (Teltzrow & Kobsa, 2004). Users may also attempt to enforce privacy through obscurity, controlling release of personal information by using an anonymization service such as a proxy server (Sackmann, Struker, & Accorsi, 2006). A proxy server may assign many users to the same IP address and can make user identification difficult.

Observations consist of records of behavior intentionally collected by a researcher or their software; observations may or may not be visible to the participant (McGrath, 1995). For example, a researcher watching a person interacting with an application would be visible to the participant, while the application logs capturing user interactions would not. One of the main concerns with observational data is that natural behaviors will often be adjusted if the participant is aware of the observations. Software that captures observational data can be proxy-based or client-side.

If Web activity is captured through proxy logging, the user must login at the beginning of each session. Advantages to this approach are that it is easier to capture data across websites, and there are fewer participant identification issues than with server-side logging due to the use of a participant account. However, users may bypass the proxy server if concerned about the sensitivity of their browsing or if they are forgetful. This may limit the breadth of data collected. Another advantage to proxy logging is that participants can work within their normal Web browser environment. However, with traditional proxy logging, browser interactions cannot be captured; and there are still caching issues if pages are cached at the browser level (Barford, Bestavros, Bradley, & Crovella, 1999). One emerging method of data logging is to embed Javascript into delivered web pages through the proxy server (Atterer, Wnuk, & Schmidt, 2006). This method can be used to capture additional data including mouse movement, scroll bar use, and key presses. Proxy servers have also been found to be less reliable and accurate than client-side logging tools for temporal measurements of Web activity (Kelly & Belkin, 2004).

One advantage of field research over laboratory experiments is that participants have access to their usual Web tools, browsers, and physical environments (Kellar, Hawkey, Inkpen, & Watters, 2008). However, with client-side logging, there is a danger of altering the participants' Web browsing environment when attempting to capture natural Web browsing behavior that is also rich in detail. The Web browsing environment includes many factors such as the user's physical location and their usual browser application, including all its normal settings. One of the main reasons for selecting field studies as a methodology is to capture natural user behavior which can be important for studies which are investigating patterns of activity. It is therefore important that the experimental software not interrupt the flow of participants' Web browsing (Chatterjee, Hoffman, & Novak, 2003).

The choice of a client-side logging tool can help mitigate concerns about changing the Web browsing environment of the user. For example, a browser helper object (BHO) can be ideal for this purpose as participants can continue using Internet Explorer with their normal settings intact, including their Favorites, History, and Google toolbar (Kellar, Hawkey, Inkpen, & Watters, 2008). The automatic loading of the BHO means that participants do not have to remember to use the study instrument. However, a BHO can only record limited types of data (i.e., interactions at the Web document level). In order to record richer interactions with the Web browser itself, a custom Web browser must be used. Developing a custom Web browser that fully mimics the appearance and functionality of participants' commercial browser applications, including all installed features (e.g., user-installed toolbars) is challenging. In some instances, researchers may have access to the source code of a commercial browser. Adapting open source software (e.g., Mozilla Firefox) is a popular choice for researchers wanting to augment browser functionality to include logging (Weinreich, Obendorf, Herder, & Mayer, 2006); however, this can limit the user population or result in participants using a different Web browsing environment as the most common browser in use is still Internet Explorer.

There are additional privacy challenges if trying to capture participants' Web activities across all contexts of use with client-side logging. It can be difficult to install the software on all computers and devices in use, particularly if custom logging software is not robust and well-tested (Kellar, Hawkey, Inkpen, & Watters, 2008). If a computer in use is not owned by the participant (i.e., one located in the workplace), it may be difficult to receive corporate permission to record data (Tang, Liu, Muller, Lin, & Drews, 2006). Self-regulation of browsing activities that are not work-related may occur if the participant believes there is a chance that the employer may have access to the logs or be able to discern their identity in subsequent analysis. In such a case, it is very important to provide privacy preserving mechanisms to help alleviate concerns of both the participant and the employer.

As summarized in Table 1, while server-side logging has relatively few privacy concerns due to the difficulty of linking the data to specific users and their personal information, it suffers from a reduced amount of information that can be gathered. Data is primarily limited to navigation with a website and data entered at that site; it will not include navigation to cached pages or websites located on other servers. Client-side logging can provide richer data, but the data collection is more invasive from a privacy perspective. As the software must be installed on client computers, the participants (and their personal information) are usually known to the researchers. Depending on the logging software, a great deal more information may be logged, including interactions with the Web browser and key strokes. Proxy logging, particularly if making use of Javascript to capture some of the user interactions, may be

Table 1. Summary of tradeoffs by type of logging for richness of data, completeness of data, ability to discern individual participants, and naturalness of their Web browsing environment.

	Server-side Logging	Proxy Logging	Client-side Logging
Richness of data	Limited to navigation, data entry on site	Limited to navigation, some form data, - improved with scripting	Rich data including navigation, key strokes, browser interaction, but BHO more limited
Completeness of data	Caching issues, site specific	Caching issues, can be bypassed	Can be bypassed by using other browser or other computer
Ability to discern individual users participants	Can be difficult (anonymization services)	Good (must log in)	Good
Naturalness of participants' browsing environment	Completely natural	Participants aware of the logging /but browser environment unchanged	Participants aware at install. Environment depends on software (BHO generally transparent, but custom browsers may not have usual functionality and settings)

a viable compromise depending on the research objectives.

ENHANCING PRIVACY DURING OBSERVATIONAL DATA COLLECTION

Collection of observational data, particularly through client-side logging applications can provide researchers with rich data about Web browsing activities and behaviors, including interactions with the Web browser. However, the intrusiveness of this type of data collection may cause participants to alter their natural Web browsing behaviors, avoid using study software, or refuse to take part in the study altogether. It is important for researchers to provide mechanisms for participants to preserve their privacy. Recommendations for such privacy enhancing mechanisms are presented next.

Lederer et al. (2004) discuss how users should be able to maintain personal privacy through understanding and action. Understanding is required so that users are aware of potential privacy violations. Opportunities for action are required so that users can appropriately manage their privacy when necessary. Following this lead, the recommendations for providing privacy preserving mechanism in this chapter will be presented with two thrusts. The first is to increase participants' understanding of the data logging and its privacy implications and to also increase their trust in the researchers' ability to maintain their privacy. In addition to educating participants, trust can be increased by limiting the recorded data to that necessary to answer the research questions and providing opportunities for participants to inspect the recorded data. Second, recommendations will be given for privacy-enhancing actions that may be afforded to participants building on the methods identified by Marx (2003) for maintaining privacy in case of surveillance. These actions include the ability to pause recording as well as the ability to mask or delete sensitive records.

Recommendations for Increasing Understanding and Trust

Lederer et al. (2004) make the point that unless users can readily determine the nature and extent of potential information disclosure, they will not be able to fully understand the privacy implications as a result of system use. For participants to be comfortable enough with the logging software to engage in their usual Web browsing activities, it is important that they understand the data being captured. The issue of trust is also an important facet of privacy concerns. Internet users' willingness to share information with a website may depend on their level of trust towards the owner of the website (Teltzrow & Kobsa, 2004).

Recommendation 1: Educate Participants

The Platform for Privacy Preferences (P3P) Initiative provides mechanisms for Web users to understand the privacy policies of websites with which they interact. Privacy in this sense is based on transparency through policies; users can inspect an organization's privacy policies and must rely on their trust in an organization to follow the stated polices (Sackmann, Struker, & Accorsi, 2006). Similarly, transparency in the process can be used to educate participants in studies involving data logging.

Consent forms should explicitly describe data collection and use so that participants have a clear understanding of what data will be collected, who will be able to see it, how the data will be used, and how it will be reported. By explicitly providing this information to participants, as well as detailing any privacy preserving mechanisms in place, researchers should be able to assuage any general privacy concerns that may prevent potential participants from taking part in the study as well as address potential privacy violations specifically. Interestingly, participants may not always take advantage of the privacy preserving mechanisms provided (Kellar, Hawkey, Inkpen, & Watters, 2008); however, the very existence of these mechanisms can give potential participants a sense of control over the privacy of their Web browsing activities which may encourage them to take part in an intrusive field study (Obendorf, personal communication, January 2008).

Recommendation 2: Only Record / Receive as Much Information as Needed

In the E-Commerce domain, it is suggested that websites gathering personal information for the purposes of personalization only gather that information that is required for the immediate service (Teltzrow & Kobsa, 2004). Limiting data collection can also increase users' willingness to disclose the information. A similar policy should help with data logging for research purposes. While it is tempting to gather as much information as possible, privacy concerns may be minimized by only recording that data which is necessary to answer the research questions. By limiting the data collected (and providing details to participants about how it will be used), participants should feel more secure that their data is being respected and being used to further research in the area of interest.

Furthermore, there may be times when very detailed raw data will be collected, but the measures of interest are aggregate scores or temporal patterns. In such cases, it may be possible to collect and process the data on the client's machine, only receiving the processed data (Hawkey, 2007). For example, if the data of interest is revisitation patterns, the URL may be necessary to identify unique pages, but otherwise irrelevant to the research questions. A script could process the data, assigning a unique ID to each URL. This would preserve the data necessary for calculations while obscuring the actual sites visited which should alleviate privacy concerns. One disadvantage to this approach is that the researcher must be well-prepared and be sure of all data analysis that will be required.

Hawkey and Inkpen (2005) were interested in investigating overall privacy concerns and temporal privacy patterns associated with the later viewing of visited Web pages. The page title and URL of visited pages were collected in order to allow participants to annotate their browsing with a privacy level in an electronic diary. In order to provide participants with as much privacy as possible, the page title and URL were stripped from the records after annotation, so that only a browser window ID, date/time stamp, and privacy level were sent to the researchers. These data were sufficient to investigate the preliminary research questions, and it was hoped that this reduction in information would encourage participants to engage in their regular Web browsing activities regardless of the sensitivity of visited pages. After an informal survey of privacy concerns associated with their longitudinal field study of Web browsing behavior, Weinreich, Obendorf, Herder & Mayer (2006) opted to use a capturing system that did not record user names and passwords entered in the browser and that ignored activity on secure connections.

Recommendation 3: Provide Opportunities for Inspection of Data

Recent research discussions have suggested providing evidence creation as a way to increase transparency and allow auditing of the data collected (Sackmann, Struker, & Accorsi, 2006). Privacy evidence is created by interpreting the collected logged data about an individual through the lens of the policies applicable to that data to illustrate compliance. Providing opportunities for participants to inspect the data being sent to researchers is a method of increasing this transparency and reassuring them that only the agreed upon data is being transferred.

This was a technique used by Hawkey and Inkpen (2005; 2006a). After using an electronic diary to annotate their visited Web pages with a privacy level, participants generated a report to email to the researchers. This report allowed participants to inspect (but not change) the data, which served as confirmation of precisely which aspects of their Web browsing activity were being transferred to the researchers. Weinreich et al. (2006) also took this approach, allowing participants to view their logged data prior to transmitting it to researchers; as discussed later, they did allow participants to take actions on that data.

Affording Privacy Preservation through Action

Three of Lederer et al.'s (2004) pitfalls relate to privacy preserving actions. The authors state that users should not have to extensively configure a system a priori in order to maintain privacy, but rather should be able to manage privacy within their normal interaction with the system. Additionally, their normal interaction with the system should not be hampered by the actions they must take to preserve privacy, nor should their normal mechanisms of preserving privacy, such as taking advantage of plausible deniability, be hampered by the technology. Furthermore, users should be able to quickly stop the release of information (i.e., have mechanisms of coarse-grained control) so that they can respond to unanticipated or quickly changing situations of use. One difficulty with providing real-time privacy enhancing mechanisms for participants is that this feedback may impact the natural flow of their Web browsing activities and make them more conscious of being observed.

The amount of control a person has over what information is recorded in Web data logs must be balanced with the need for that data for the research purposes. While not all of these recommendations may be appropriate for a given study, providing participants with some level of control over their data should help alleviate privacy concerns (Teltzrow & Kobsa, 2004). The intent of these recommendations is to provide participants with similar privacy-preserving mechanisms to those that they might use in their normal Web interactions when trying to limit the data collected by Web-servers (Cranor, 1999) or when under surveillance (Marx, 2003).

Recommendation 4: Provide the Ability to Pause Recording

Client-side logging software can be developed to automatically log all Web browsing actions or to be manually started by participants on a periodic basis. As previously stated, browser helper objects automatically load when Internet Explorer is loaded; a similar method is the Cross Platform Component Object Model (XPCOM) for Mozilla's Firefox browser. Participants may still bypass the collection of data, however, by using a different Web browser. Custom Web browsers generally must be manually started by the participant. This provides participants with an opportunity to only log those browsing activities that they wish to share. This may be suitable for research investigating episodes of targeted activity such as information seeking tasks as in Kellar et al. (2007). In that case, participants were asked to use a custom web browser periodically to perform information seeking tasks. Their participants could opt not to use the custom browser when conducting sensitive browsing activities.

If periodic recording of data is suitable for the research question, custom logging software such as Web browsers or toolbars associated with browser helper objects or other browser plug-ins should include a recording button that can be toggled on and off. This will allow participants to pause recording of their browsing when engaging in sensitive activities such as visiting socially inappropriate websites or engaging in confidential transactions. This can be very important when recording keystroke data that may include passwords. Alternatively, data logging software could be developed to avoid collecting password data or form field data if this data is not pertinent to the research question (Weinreich, Obendorf, Herder, & Mayer, 2006).

Recommendation 5: Provide the Ability for Participants to Mask Data

Another way to provide privacy for participants is to allow them to mask sensitive data. This may be more appropriate for studies which would like to capture all of a participant's browsing activities. Depending on the research questions, one or more fields in a data log may be candidates for masking.

Kellar et al. (2007) used masking in their field study investigating information seeking tasks and their impact on the use of Web browser navigation mechanisms. Participants could remove details about specific visited pages deemed to be sensitive. Masking was also an approach taken by Hawkey and Inkpen (2006a) in a field study investigating participants' visual privacy concerns for traces of their Web browsing activity. In this study, the researchers wanted to investigate the impact of context (location, visited page) on privacy concerns. They therefore needed to not only collect the URL and page title for annotation by participants within the electronic diary (as in their 2005 study), but to also receive that information as part of the generated report. As they did not want receipt of this additional information to impact participants' willingness to visit sensitive sites, they provided participants with the ability to selectively blind any sensitive data contained in the URL and page title. The electronic diary in Hawkey and Inkpen's (2005) study was modified to allow participants to mask entries in the diary by removing the page title and URL after applying a privacy level to a visited web page. When masking an entry, participants were asked to give a general reason for the sanitized browsing such as "looking for medical information"; the default label was "no reason given." An inspection of the visited pages revealed that the proportion of participants in the field study with instances of adult content was comparable to frequency reports of erotica viewing as reported by participants in a related anonymous survey (Hawkey & Inkpen, 2006b).

This may indicate that participants' normal Web usage, including those activities not considered to be socially desirable (Fisher, 1993), was recorded during the study.

Recommendation 6: Provide Participants with the Ability to Delete Data

Deletion of records may be feasible for some research questions. Deletion is similar to pausing of the recording but is done after the fact. Research questions that may be answered by investigating specific episodes of Web browsing would be candidates for this approach. In order to preserve the integrity of the data, researchers may want to limit how the data can be handled, perhaps providing a data viewer that allows deletion at the record level, but no modifications of individual fields. Alternatively, deletion could be offered at the session level by providing participants with the opportunity to consent to the session being included in the study data upon exiting the data collection software. Weinreich, Obendorf, Herder & Mayer (2006) allowed their participants to view the data logs (as text files) before transmitting them to researchers. The text files were editable, so participants could potentially modify the data at will, either through masking or deletion of specific records or entire files; however, no participants are believed to have actually modified their data (Obendorf, personal communication, January 2008).

FUTURE TRENDS

The previous sections presented current challenges for researchers attempting to capture observational data and provided several recommendations for enhancing participant privacy in an effort to encourage users to engage in their normal Web browsing behaviors (summarized in Table 2). Privacy concerns of participants can be expected to increase as researchers gather more contextual information during studies, including their users' activities, goals, attitudes, and processes, to augment logged data (Kellar, Hawkey, Inkpen, & Watters, 2008). Contextual information plays an important role in how we understand and interpret people's everyday behavior. Information that provides additional details about people, such as their location or task, can help us better understand and interpret their actions. In a Web environment, contextual information can be used to determine the activities in which a user is engaging, their motivations for engaging in those activities, as well as perceptions about the current tool or the information being viewed. Participant annotation of log data is one emerging method of gaining additional context (Kellar, Hawkey, Inkpen, & Watters, 2008). Another method is to retrospectively discuss portions of the data logs with participants using critical incident techniques (Choo, Detlor, & Turnbull, 2000).

There is also an increasing need to capture Web activity across usage contexts. It is important during studies of natural browsing behaviors that we record specific aspects of context that may be influencing behaviors at the time and capture those behaviors across all normal usage contexts. Web usage can vary across different locations (e.g., home, work) and devices (laptop, desktop) (Hawkey, 2007). Additionally, different Web browsers or Web browser settings may be used in these environments, and browsing may be conducted for different purposes (e.g., personal, work-related). There will be many research challenges to ensure that participant privacy is considered across contexts of use, as well as the privacy of any companies or organizations involved. As the boundaries between personal time and work time decrease, more and more participants may be multi-tasking across contexts (Olson-Buchanan & Boswell, 2006).

Loggers that capture data across applications are becoming more common as researchers investigate behaviors at the level of the activity or

Table 2. Summary of recommendations for enhancing participants' privacy and thereby encouraging natural Web browsing behaviors.

Recommendations for Enhancing Privacy when Logging Web Browsing Activity
Increase Privacy and Trust
1. Educate participants about what information is being collected
2. Only record/receive as much information as is needed for the research questions
3. Provide opportunities for participants to inspect the data collected
Afford Privacy Preservation Through Action
4. Provide the ability for participants to pause recording of the data
5. Provide the ability for participants to mask particularly sensitive data
6. Provide participants with the ability to delete data

are gathering more contextual information about multi-tasking. Such logging applications increase privacy concerns of participants, whether they are keystroke loggers or screen capture applications. Screen capture software gives context by revealing what the user sees while interacting with their Web browser including applications outside of the Web browser. If such applications are used, participants are essentially agreeing to have all of their computer activity logged. It can be very difficult to recruit users to take part in such studies, and there may be privacy concerns not only for the participants, but for those with whom they communicate (i.e., email correspondence) (Tang, Liu, Muller, Lin, & Drews, 2006). Research ethics boards may require informed consent from all collaborators before their data is recorded.

As more contextual data is captured and more logging is conducted across applications, it will be increasingly important for researchers to consider participants' privacy concerns (Kellar, Hawkey, Inkpen, & Watters, 2008). Providing privacy enhancing methods such as those suggested in the recommendations should help alleviate privacy concerns which may impact recruitment efforts and encourage participants to engage in their usual activities. Researchers will need to be innovative in their methodological techniques as they balance the participants' desires for privacy with researchers' need for rich data to answer questions of interest.

Researchers must also keep abreast of changing privacy regulations at the governmental and organizational level. Given current political climates, it is expected that more rigid protections of data and their re-use will be legislated. Keeping informed of current practices is particularly important if conducting research across borders, as regulations vary widely.

CONCLUSION

This chapter first presented relevant privacy literature including general privacy theories and privacy concerns specific to Web browsing activities. This background provided the necessary grounding for the subsequent discussions of privacy issues with respect to the collection of log data for analysis. The main privacy issues presented were 1) ensuring that governmental and organizational regulations with respect to the safeguarding of participant data are met and 2) providing privacy preserving mechanisms for participants in order to encourage natural Web browsing behaviors. Privacy concerns will depend on the type of data logging. Several tradeoffs were discussed according to the location of the data logging (see Table 1 for a summary). While server-side data is less intrusive for participants and allows them to engage in their normal privacy preserving mechanisms, the data collected is limited and often unreliable. Client-side logging can provide richer data including Web browser interactions; however, data collection is more intrusive.

Several key recommendations for mechanisms to enhance participants' privacy were suggested (see Table 2 for a summary). These include ways to increase participants' understanding and trust of the data logging for the study as well as methods to allow them to control the capture of particularly sensitive data through masking, blocking, or deleting it. The author hopes that these recommendations will prove to be useful for researchers designing research methodologies that include the capture of observational data.

REFERENCES

Ackerman, M., Cranor, L., & Reagle, J. (1999). Privacy in E-Commerce: Examining User Scenarios and Privacy Preferences. In *1st ACM conference on Electronic commerce* (pp. 1-8). Denver, CO: ACM.

Anick, P. (2003). Using Terminological Feedback for Web Search Refinement - A Log-based Study. In *26th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 88-95). Toronto, Canada: ACM.

Atterer, R., Wnuk, M., & Schmidt, A. (2006). Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In *15th International Conference on World Wide Web* (pp. 203-212). Edinburgh, Scotland: ACM.

Barford, P., Bestavros, A., Bradley, A., & Crovella, M. (1999). Changes in Web client access patterns: Characteristics and Caching Implications. *World Wide Web (Bussum)*, *2*(1-2), 15–28. doi:10.1023/A:1019236319752

Chatterjee, P., Hoffman, D. L., & Novak, T. P. (2003). Modeling the Clickstream: Implications for Web-Based Advertising Efforts. *Marketing Science*, *22*(4), 520–541. doi:10.1287/mksc.22.4.520.24906

Choo, C. W., Detlor, B., & Turnbull, D. (2000). Information Seeking on the Web: An Integrated Model of Browsing and Searching. *First Monday, 5*(2), Retrieved August 3, 2004, from http://firstmonday.org/issues/issue2005_2002/choo/index.html.

Cockburn, A., & McKenzie, B. (2001). What do web users do? An empirical analysis of web use. *International Journal of Human-Computer Studies*, *54*(6), 903–922. doi:10.1006/ijhc.2001.0459

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data Preparation for Mining World Wide Web Browsing Patterns. *Knowledge and Information Systems*, *1*(1), 5–32.

Cranor, L. F. (1999). Internet Privacy. *Communications of the ACM*, *42*(2), 28–31. doi:10.1145/293411.293440

Dalhousie Research Services. (2006). *Directives to Researchers regarding Compliance with the University Policy for the Protection of Personal Information from Access Outside Canada*. Retrieved July 29, 2007 from http://researchservices.dal.ca/files/Personal_Information_Protection_Guide.pdf.

Fenstermacher, K., & Ginsburg, M. (2003). Client-Side Monitoring for Web Mining. *Journal of the American Society for Information Science and Technology*, *54*(7), 625–637. doi:10.1002/asi.10255

Fisher, R. J. (1993). Social Desirability Bias and the Validity of Indirect Questioning. *The Journal of Consumer Research*, *20*(2), 303–315. doi:10.1086/209351

Garfinkel, S., & Spafford, G. (2001). *Web Security, Privacy & Commerce, 2nd Edition*: O'Reilly.

Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Garden City, New York: Doubleday Anchor Books.

Hawkey, K. (2007). *Managing the visual privacy of incidental information in web browsers.* Unpublished PhD Dissertation, Dalhousie University, Halifax, Nova Scotia.

Hawkey, K., & Inkpen, K. (2005). Privacy Gradients: Exploring ways to manage incidental information during co-located collaboration. In *CHI '05 Extended Abstracts of Human Factors in Computing Systems* (pp. 1431-1434). Portland, Oregon: ACM.

Hawkey, K., & Inkpen, K. M. (2006a). Examining the Content and Privacy of Web Browsing Incidental Information. In *15th International Conference on World Wide Web* (pp. 123-132). Edinburgh, Scotland: ACM.

Hawkey, K., & Inkpen, K. M. (2006b). Keeping Up Appearances: Understanding the Dimensions of Incidental Information Privacy. In *SIGCHI Conference on Human Factors in Computer Systems* (pp. 821-830). Montreal, Quebec: ACM.

Huseby, S. H. (2004). *Innocent Code: A Security Wake-Up Call for Web Programmers*. UK: John Wiley & Sons Ltd.

Joinson, A. N., Paine, C., Reips, U.-D., & Buchanan, T. (2006). Privacy and Trust: The role of situational and dispositional variables in online disclosure. In *Privacy, Trust, and Identity Issues for Ambient Intelligence Workshop, Pervasive 2006* (pp. 1-6). Dublin, Ireland.

Kellar, M., Hawkey, K., Inkpen, K. M., & Watters, C. (2008). Challenges of Capturing Natural Web-based User Behaviours. *International Journal of Human-Computer Interaction*, *24*(4), 385–409. doi:10.1080/10447310801973739

Kellar, M., Watters, C., & Shepherd, M. (2007). A Field Study Characterizing Web-based Information Seeking Tasks. *Journal of the American Society for Information Science and Technology*, *58*(7), 999–1018. doi:10.1002/asi.20590

Kelly, D., & Belkin, N. (2004). Display Time as Implicit Feedback: Understanding Task Effects. In *27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 377-384). Sheffield, UK: ACM.

Lederer, S., Hong, J. I., Dey, A. K., & Landay, J. A. (2004). Personal privacy through understanding and action: five pitfalls for designers. *Personal and Ubiquitous Computing*, *8*(6), 440–454. doi:10.1007/s00779-004-0304-9

Lederer, S., Mankoff, J., & Dey, A. K. (2003). Towards a Deconstruction of the Privacy Space. *Workshop on Ubicomp Communities: Privacy as Boundary Negotiation, UBICOMP 2003*: Retrieved August 12, 2005 from http://guir.berkeley.edu/pubs/ubicomp2003/privacyworkshop/papers/lederer-privacyspace.pdf

Liu, C., Marchewka, J. T., Lu, J., & Yu, C.-S. (2004). Beyond concern: a privacy-trust-behavioral intention model of electronic commerce. *Information & Management*, *42*(1), 127–142.

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research*, *15*(4), 336–355. doi:10.1287/isre.1040.0032

Marx, G. T. (2003). A Tack in the Shoe: Neutralizing and Resisting the New Surveillance. *The Journal of Social Issues*, *59*(2), 369–390. doi:10.1111/1540-4560.00069

McCullagh, D. (2006). *AOL's disturbing glimpse into user's lives*. CNET News.com, Online at: http://news.com.com/2100-1030_3-6103098.html. Retrieved October 5, 2006.

McGrath, J. E. (1995). Methodology matters: doing research in the behavioral and social sciences. In J. G. R. Baeker, W. Buxton, and S. Greenberg (Ed.), *Human-Computer Interaction: Toward the Year 2000* (pp. 152-169).

Meier, J. D., Mackman, A., Dunner, M., Vasireddy, S., Escamilla, R., & Murakan, A. (2003). *Improving Web Application Security: Threats and Countermeasures*: Microsoft Press.

Olson, J. S., Grudin, J., & Horvitz, E. (2005). A Study of Preferences for Sharing and Privacy, *in CHI '05 Extended Abstracts of Human Factors in Computing Systems* (pp. 1985-1988). Portland, Oregon: ACM. P&AB. (2003). Consumer Privacy Attitudes: A Major Shift Since 2000 and Why. *Privacy & American Business Newsletter, 10*(6), 1,3-5.

Olson-Buchanan, J. B., & Boswell, W. R. (2006). Blurring boundaries: Correlates of integration and segmentation between work and nonwork. *Journal of Vocational Behavior*, *68*(3), 432–445. doi:10.1016/j.jvb.2005.10.006

Palen, L., & Dourish, P. (2003). Unpacking "Privacy" for a Networked World. In *SIGCHI Conference on Human Factors in Computing Systems* (pp. 129-136). Ft. Lauderdale, FL: ACM.

Phillips, D. J. (2002). Context, identity, and privacy in ubiquitous computing environments. In *Workshop on socially-informed design of privacy-enhancing solutions, Ubicomp 2002*. Goteborg, Sweden.

Sackmann, S., Struker, J., & Accorsi, R. (2006). Personalization in Privacy-Aware Highly Dynamic Systems. *Communications of the ACM*, *49*(9), 32–38. doi:10.1145/1151030.1151052

Tang, J. C., Liu, S. B., Muller, M., Lin, J., & Drews, C. (2006). Unobtrusive but invasive: using screen recording to collect field data on computer-mediated interaction. In *20th anniversary conference on Computer supported cooperative work* (pp. 479-482). Banff, Alberta: ACM.

Teltzrow, M., & Kobsa, A. (2004). Impacts of user preferences on personalized systems: a comparative study. In *Designing personalized user experiences in eCommerce* (pp. 315-332). Norwell, MA, USA: Kluwer Academic Publishers.

Weinreich, H., Obendorf, H., Herder, E., & Mayer, M. (2006). Off the Beaten tracks: Exploring Three Aspects of Web Navigation. In *15th International Conference on World Wide Web* (pp. 133-142). Edinburgh, Scotland: ACM.

Westin, A. F. (2003). Social and Political Dimensions of Privacy. *The Journal of Social Issues*, *59*(2), 431–453. doi:10.1111/1540-4560.00072

Yun, G. W., Ford, J., Hawkins, R. P., Pingree, S., & McTavish, F. (2006). On the validity of client-side vs server-side web log data analysis. *Internet Research*, *16*(5), 537–552. doi:10.1108/10662240610711003

KEY TERMS AND DEFINITIONS

Anonymized Data: Data that has been collected with identifying information, but has had subsequent removal of any links between the data and identifying information so that the researcher can no longer discern the specific owner of the data.

Anonymous Data: Data that is collected without any associated identifying information.

Client-Side Logging: Software that records Web browsing behavior at the user's computer. This is generally achieved either through a custom web browser or through browser plug-ins such as tool bars or browser helper objects.

Contextual Privacy Concerns: Privacy concerns vary in any given instance according to the inherent privacy concerns of the user and the situational factors at play. These include the viewer of the information, level of control retained over the information, and the type of information. Furthermore, these factors can vary according to the device in use and the location.

Inherent Privacy Concerns: An individual's general privacy concerns; their disposition to privacy. Factors which may impact a person's disposition to privacy include their age and computer experience.

Privacy: "The claim of an individual to determine what information about himself or herself should be known to others." (Westin, 2003).

Proxy Logging: Software that serves as an intermediary between the user's web browser and the web site servers. Users generally have to log-in to the proxy and the proxy server can be used to augment retrieved web pages.

Server-Side Logging: Software that records Web browsing behavior at the server. Data collection is generally limited to navigation information.

Web Browsing Behaviors: User behaviors on the Web including their browsing activities and Web browser interactions. Privacy concerns have been found to impact Web browsing behaviours.

Web Browsing Environment: The context within which Web browsing occurs. For studies of Web usage this includes the Web browser and its associated tools (e.g., history, specialized toolbars), the task, and the motivation for conducting the browsing.

This work was previously published in the Handbook of Research on Web Log Analysis, edited by B. Jansen, A. Spink, and I. Taksa, pp. 80-98, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 7.8
A Model-Based Approach for Diagnosing Fault in Web Service Processes

Yuhong Yan
Concordia University, Canada

Philippe Dague
University Paris-Sud 11, France

Yannick Pencolé
LAAS-CNRS, France

Marie-Odile Cordier
IRISA, France

ABSTRACT

Web service orchestration languages are defined to describe business processes composed of Web services. A business process can fail for many reasons, such as faulty Web services or mismatching messages. It is important to find out which Web services are responsible for a failed business process because we could penalize these Web services and exclude them from the business process in the future. In this paper, we propose a model-based approach to diagnose the faults in a Web service-composed business process. We convert a Web service orchestration language, BPEL4WS, into synchronized automata, so that we have a formal description of the topology and variable dependency of the business process. After an exception is thrown, the diagnoser can calculate the business process execution trajectory based on the formal model and the observed evolution of the business process. The faulty Web services are deduced from the variable dependency on the execution trajectory.

INTRODUCTION

Web services not only function as middleware for application invocation and integration, but also function as a modeling and management tool for business processes. In a Service Oriented Architecture paradigm, a business process can be composed of Web services distributed over the Internet. This kind of business processes can be flexible and optimal by using the best services from multiple companies. Various Web service process description languages are designed by standard bodies and companies. Among them, Business Process Execution Language for Web Service (BPEL4WS, denoted as BPEL after) (Andrews, Curbera, Dholakia, Goland, el al., 2003) is the de facto standard used to describe an executable Web service process. In this article, we study the behavior of a business process described in BPEL.

As any other systems, a business process can fail. For a Web service process, the symptom of a failure is that exceptions are thrown and the process halts. As the process is composed of multiple Web services, it is important to find out which Web services are responsible for the failure. If we could diagnose the faulty Web services, we could penalize these Web services and exclude them from the business process in the future. The current throw-and-catch mechanism is very preliminary for diagnosing faults. It relies on the developer associating the faults with exceptions at design time. When an exception is thrown, we say certain faults occur. But this mechanism does not guarantee the soundness and the completeness of diagnosis.

In this article, we propose a model-based approach to diagnose faults in Web service processes. We convert the basic BPEL activities and constructs into synchronized automata whose *states* are presented by the values of the *variables.* The process changes from one state to another by executing an *action*, for example, assigning variables, receiving or emitting messages in BPEL. The emitting messages can be a triggering *event* for another service to take an action. The diagnosing mechanism is triggered when exceptions are thrown. Using the formal model and the runtime observations from the execution of the process, we can reconstruct the unobservable trajectories of the Web service process. Then the faulty Web services are deduced based on the variable dependency on the trajectories. Studying the fault diagnosis in Web service processes serves the ultimate goal of building self-manageable and self-healing business processes.

This article is organized as follows: the Advanced Fault Management for Web Service Processes section analyzes the fault management tasks in Web service processes and motivates the use of Model-based Diagnosis (MBD) for Web services monitoring and diagnosis; The Principle of Model-based Diagnosis for Discrete Event Systems section presents the principles for MBD; the Modeling Web Service Processes with Discrete-Event Systems section formally defines the way to generate an automaton model from a BPEL description; the Model-based Diagnosis for Web Service Processes section extends the existing MBD techniques for Web service monitoring and diagnosis; the Related Work and Discussion section is the related work, and lastly is the Conclusion section.

ADVANCED FAULT MANAGEMENT FOR WEB SERVICE PROCESSES

A Web service process can run down for many reasons. For example, a composed Web service may be faulty, an incoming message mismatches the interface, or the Internet is down. The *symptom*[1] of a failed Web service process is that *exceptions* are thrown and the process is halted. The current fault handling mechanism is throw-and-catch, similar to programming languages. The *exceptions* are thrown at the places where the process cannot be executed. The *catch* clauses process the exceptions, normally to recover the failure effects by executing predefined actions.

The throw-and-catch mechanism is very preliminary for fault diagnosis. The exception reports where it happened and returns some fault information. The exceptions can be regarded as associated with certain faults. When an exception is thrown, we deduce that its associated fault occurred. Customized exceptions are especially defined for this purpose. This kind of association relations rely on the empirical knowledge of the developer. It may not be a real cause of the exceptions. In addition, there may exist multiple causes of an exception which are unknown to the developer. Therefore, the current throw-and-catch mechanism does not provide sound and complete diagnosis. For example, when a Web service throws an exception about a value in a customer order, not only the one that throws the exception may be faulty, but the one that generates these data may also be faulty. But a Web service exception can only report the Web service where the exception happens with no way to know who generated these data. In addition, all the services that modified the data should be also suspected. Not all of this kind of reasoning is included in the current fault handling mechanism. *A systematic diagnosis mechanism which is based on the model of the Web service process and a solid theoretical foundation needs to be developed.* This is the objective of this article.

The diagnosis task is to determine the Web services responsible for the exceptions. These Web services will be diagnosed faulty. During the execution of a BPEL process, the exceptions come from the BPEL engine or the infrastructure below, for example. Apache Tomcat, and Internet. We classify the exceptions into ***time-out*** exceptions and ***business logic*** exceptions.

The *time-out* exceptions are due to either a disrupted network or unavailable Web services. If there is a lack of response, we cannot distinguish whether the fault is in the network or at the remote Web service, except if information is transmitted by the network fault management in the first case. Since we cannot diagnose which kind of faults prevent a Web service from responding, we can do little with *time-o*ut exceptions. Indeed what can be done is more statistics at the level of process classes (and not process instances) that will be used by experts to improve the QoS.

The *business logic* exceptions occur while invoking an external Web service and executing BPEL internal activities. For example, mismatching messages (including the type of parameters and the number of parameters mismatching) cause the exceptions to be thrown when the parameters are passed to the remote method. BPEL can throw exceptions indicating the input data is wrong. During execution, the remote service may stop if it cannot process the request. The most common scenarios are the invalid format of the parameters, for example, the data is not in a valid format, and the data is out of the range. The causes of the exceptions are various and cannot be enumerated. The common thread is that a business logic exception brings back information on the variables that cause the problem. In this article, our major effort is on diagnosing business logic-related exceptions at the process instances level.

The advanced fault management mechanism serves the ultimate goal to build self-manageable Web service processes. Fault management mechanisms can be among other self-manageable functions. Some functions related to fault management are:

- **Monitoring** the execution of Web service process, and record necessary and sufficient information for online/offline diagnosis. Insufficient information cannot produce correct diagnosis. In Web service processes, we need to keep a chronological record for some of the variables.
- **Detecting** faulty behavior. In other physical tasks, detecting needs to compare the observations with the predictions from the system description to discover the discrepancies. For Web service processes, this task is a trivial one to observe exceptions. But we can imagine building new detectors in order

to detect symptoms earlier and "closer" to the causes.

- **Diagnosing** the causes of exceptions. This is the major focus of this article. See Model-based Diagnosis for Web Service Processes section for detail.
- **Recovering** from the failure effects. BPEL uses predefined compensation handlers and fault handlers to eliminate failure effects. As failure effects cannot be revealed by the empirical diagnosis mechanism in BPEL, the predefined compensation actions may not be sufficient. A more advanced recovery mechanism has to be defined, based on the model-based diagnosis developed in this article, although it is not covered in this article.

THE PRINCIPLE OF MODEL-BASED DIAGNOSIS FOR DISCRETE EVENT SYSTEMS

MBD is used to monitor and diagnose both static and dynamic systems. It is an active topic in both Artificial Intelligence (AI) and Control Theory communities. Automated diagnosis has been applied to all kinds of systems, such as communication systems, plant processes and automobiles. The early results in MBD are collected in (Hamscher, Console, & de Kleer, 1992). Let us briefly recall the terminology and notations adopted by the model-based reasoning community.

- *SD*: system description. In the AI-rooted diagnostic techniques, SD is symbolically modeled, for example, in first-order logic sentences, and in DES as used in this article.
- *COMPS*: a finite set of constants to represent the components in a system.
- *System*: a pair (*SD*, *COMPS*).
- *D*: a mode assignment to each component in the system. An assignment to a component is a unary predicate. For example, for a component $c_i \in COMPS$, $\neg ab(c_i)$ means c_i working properly, and $ab(c_i)$ means c_i is in an abnormal mode. Obviously a component has different behavior for different modes.
- *Observables*: the variables that can be observed/measured. For a physical system, the observables are the variables measured by sensors, or events reported by alarms, and so forth.
- *OBS*: a set of observations. They are the values of the *Observables*. They can be a finite set of first-order sentences, for example, value assignments to some variables.
- *Observed system*: (*SD*, *COMPS*, *OBS*).

Diagnosis is a procedure to determine which components are correct and which components are faulty in order to be consistent with the observations and the system description. Therefore, logically, a consistency-based diagnosis is:

Definition 1 *D is a consistency-based* ***diagnosis*** *for the observed system* $\langle SD, COMPS, OBS\rangle$, *if and only if it is a mode assignment and* $SD \cup D \cup OBS \not\models \perp$.

From Definition 1, diagnosis is a mode assignment *D* that makes the union of *SD*, *D* and *OBS* logically consistent. *D* can be partitioned into two parts:

- D_{ok} which is the set of components which are assigned to the $\neg ab$ mode;
- D_f which is the set of components which are assigned to the *ab* mode.

Usually we are interested in those diagnoses which involve a minimal set of faults, that is, the diagnoses for which D_f is minimal for set inclusion.

Definition 2 *A diagnosis D is* ***minimal*** *if and only if there is no other diagnosis D' for* $\langle SD, COMPS, OBS\rangle$ *such that* $D_f' \subset D_f$.

The dual concept of a diagnosis is a conflict.

Definition 3 *A set $CO \subseteq COMPS$ is a conflict for $\langle SD, COMPS, OBS\rangle$, if and only if $SD \cup OBS \cup \{\neg ab(C) | C \in CO\} \models \bot$.*

Similarly a minimal conflict is a conflict that is minimal for set inclusion. In (Reiter, 1987), Reiter introduces the hitting set algorithm for computing minimal diagnoses using the set of conflicts.

Definition 4 *((Reiter, 1987)) Let C be a collection of sets. A hitting set for C is a set $H \subseteq \bigcup_{S \in C} S$ such that $H \cap S \neq \emptyset$ for each $S \in C$. A hitting set is minimal if no proper subset of it is a hitting set.*

Theorem 1 *((Reiter, 1987)) A set $D \subseteq COMPS$ is a minimal diagnosis for $\langle SD, COMPS, OBS\rangle$ if and only if D is a minimal hitting set for the collection of conflicts (or equivalently for the collection of minimal conflicts).*

When the system description is in first order logic, the computation of all diagnoses is more generally rooted in automated reasoning, relying on prime implicates of $SD \cup OBS$ in the form of disjuncts of *ab*-literals, and on their prime implicants in the form of conjuncts of *ab*-literals (Hamscher et al., 1992).

When applying MBD, a formal system description is needed. Therefore, we need to study the proper formal model for Web service processes. As the interactions between Web services are driven by message passing, and message passing can be seen as discrete events, we consider the Discrete Event Systems (DES) suitable to model Web service processes. Many discrete event models, such as Petri nets, process algebras and automata, can be used for Web service process modeling. These models were invented for different purposes, but now they share many common techniques, such as symbolic representation (in addition to graph representation in some models) and similar symbolic operations. In this article, we present a method to represent Web service processes described in BPEL as automata in the Modeling Web Service Processes with Discrete-Event Systems section. Here we introduce MBD techniques for automata. A classic definition of deterministic automaton is as follows:

Definition 5 *An **automaton** Γ is a tuple $\Gamma = \langle X, \Sigma, T, I, F\rangle$ where:*

- X is a finite set of states;
- *Σ is a finite set of events;*
- *$T \subseteq X \times \Sigma \rightarrow X$ is a finite set of transitions;*
- *$I \subseteq X$ is a finite set of initial states;*
- *$F \subseteq X$ is a finite set of final states.*

Definitions 6, 7 and 8 are some basic concepts and operations about automata.

Definition 6 ***Synchronization** between two automata $\Gamma_1 = \langle X_1, \Sigma_1, T_1, I_1, F_1\rangle$ and $\Gamma_2 = \langle X_2, \Sigma_2, T_2, I_2, F_2\rangle$, with $\Sigma_1 \cap \Sigma_2 \neq \emptyset$, produces an automaton $\Gamma = \Gamma_1 \| \Gamma_2$, where $\Gamma = \langle X_1 \times X_2, \Sigma_1 \cup \Sigma_2, T, I_1 \times I_2, F_1 \times F_2\rangle$, with:*

$T((x_1, x_2), e) = (T_1(x_1, e), T_2(x_2, e))$, if $e \in \Sigma_1 \cap \Sigma_2$,

$T((x_1, x_2), e) = (T_1(x_1, e), x_2)$, if $e \in \Sigma_1 \backslash \Sigma_2$

$T((x_1, x_2), e) = (x_1, T_2(x_2, e))$, if $e \in \Sigma_2 \backslash \Sigma_1$

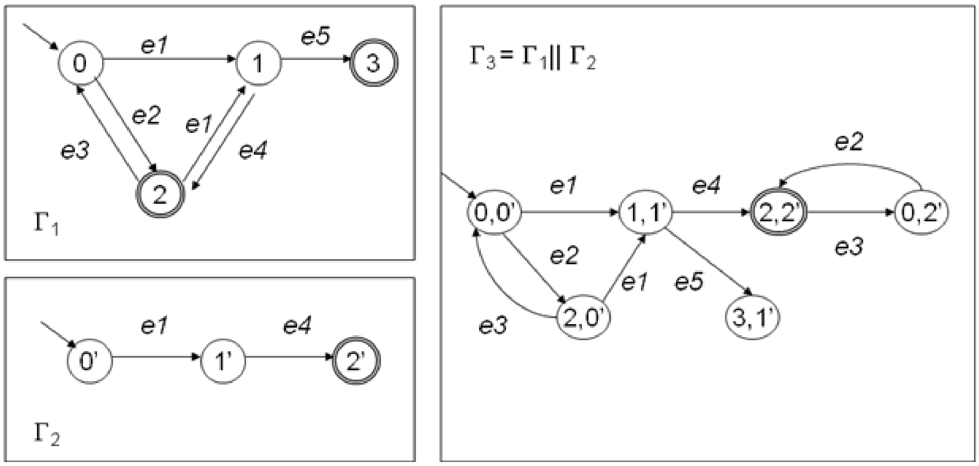
Assume $s = \Sigma_1 \cap \Sigma_2$ is the joint event set of Γ_1 and Γ_2, Γ can also be written as $\Gamma = \Gamma_1 \|_s \Gamma_2$.

Example 1 *In Figure 1, Γ_1 and Γ_2 are two automata. The third one Γ_3 is produced by synchronizing Γ_1 and Γ_2.*

Definition 7 *A **trajectory** of an automaton is a path of contiguous states and transitions in the automaton that begins at an initial state and ends at a final state of the automaton.*

Example 2 *The trajectories in the automaton Γ_3 in Figure 1 can be represented as these two*

Figure 1. An example of synchronization



formulas, in which [] means the content in [] repeated 0 or more times:*

$$[(0,0') \xrightarrow{e2} (2,0') \xrightarrow{e3}]^*[(0,0') \xrightarrow{e1} (1,1') \xrightarrow{e4} (2,2')][\xrightarrow{e3} (0,2') \xrightarrow{e2} (2,2')]^*,$$

$$[(0,0') \xrightarrow{e2} (2,0') \xrightarrow{e3}]^*[(0,0') \xrightarrow{e2} (2,0') \xrightarrow{e1} (1,1') \xrightarrow{e4} (2,2')][\xrightarrow{e3} (0,2') \xrightarrow{e2} (2,2')]^*.$$

Definition 8 ***Concatenation*** *between two automata* $\Gamma_1 = \langle X_1, \Sigma_1, T_1, I_1, F_1\rangle$ *and* $\Gamma 2 = \langle X_2, \Sigma_2, T_2, I_2, F_2\rangle$*, with* $\Sigma_1 \cap \Sigma_2 = \varnothing$ *and* $F_1 \cap I_2 \neq \varnothing$*, produces an automaton* $\Gamma=\Gamma_1 \circ \Gamma_2$*, where* $\Gamma = \langle X_1 \cup X_2, \Sigma_1 \cup \Sigma_2, T_1 \cup T_2, I_1, F_2 \cup (F_1 \backslash I_2)\rangle$.

The principle of diagnosis using DES models was founded by (Sampath, Sengupta, Lafortune, Sinnamohideen, & Teneketzis, 1995) and (Cordier & Thiébaux, 1994). System description *SD* models both correct and faulty behavior of a system. Assume system description *SD* is an automaton Γ, and observed events in chronological order are represented as another automaton *OBS*. Assume the joint event set of Γ and *OBS* is s. In this context, we call Diagnosis the automaton produced by synchronizing Γ and *OBS*:

$$\text{Diagnosis} = \Gamma \|_s OBS \qquad (1)$$

From the definition of synchronization, it is easy to prove that each trajectory in Diagnosis explains the sequence of observations in the sense that observable events in the trajectory occur in the identical chronological order as in *OBS*, that is:

$$\text{Diagnosis} \models OBS \qquad (2)$$

Therefore, Diagnosis for DES is what is called an *abductive diagnosis* in MBD theory.

Example 3 *In Figure 1,* Γ_1 *is a system description in which* e_2 *and* e_3 *represent occurrences of faults which are not observable directly (otherwise, the diagnosis would be trivial).* Γ_2 *is an observation in which two events* e_1 *and* e_4 *are observed sequentially. The* Diagnosis *is* Γ_3.

It is not so easy to compute the trajectories of Diagnosis because there are several possibilities for trajectory expansion that can arise from partial observations. We need to get all the possible trajectories. For trajectory expansion, people basically use search algorithms. Other algorithms, rooted from search algorithms, can also be used. For example, planning tools and model checking tools are used for trajectory expansion. Of course, these tools have to be modified in order to get complete trajectories.

Diagnostic process is almost achieved when Diagnosis is obtained, because Diagnosis explains the observations based on SD (as an automaton Γ). If we want to obtain diagnoses {D} as mode assignments as in the consistency-based framework, we need a mapping function f: Diagnosis $\rightarrow \{D\}$. Each trajectory t in Diagnosis is mapped into a D, that is, $t \rightarrow D$. As domain knowledge, a faulty event e_f is known to be associated with a fault mode $F(c_i)$ of some component c_i, that is, $e_f \leftrightarrow F(c_i)$[2]. If e_f is included in a trajectory t, we deduce that the correspondent fault $F(c_i)$ occurs. Formally,

Proposition 1 *Assume t is a trajectory in* Diagnosis*, then $t \rightarrow D$ where mode assignment D is defined by $D_f = \{c_j \mid e_f \leftrightarrow F(c_j) \text{ and } e_f \in t\}$ (and thus $D_{ok} = \{c_j \mid c_j \in COMPS \backslash D_f\}$).*

As each fault event maps to a fault, practically we need only to know the set of faulty events in a trajectory:

$$t \rightarrow \{e_f \mid e_f \in t\} \quad (3)$$

From (3), if we know $\{e_f\}$, we can easily get D_f and thus D. In the following, we use $\{e_f\}$ to represent a D_f. As there are often multiple trajectories $\{t^i\}$ in Diagnosis, the diagnoses $\{D^i\}$ are also multiple:

Proposition 2 *Assume $\{t_i\}$ is the set of all trajectories in* Diagnosis*, then $\{t^i\} \rightarrow \{D^i\}$, where $D_f^i = \{c_j \mid e_f^i \leftrightarrow F(c_j) \text{ and } e_f^i \in t^i\} \subseteq D^i$.*

In general, we are interested only in minimal diagnoses, that is, in Proposition 2 we keep only those D_f^i which are minimal.

Example 4 *From Diagnosis Γ_3 in Figure 1, we get 2 kinds of possible sequences of faulty events:*

{[*e*2, *e*3]*, [*e*3, *e*2]*}, {[*e*2, *e*3]*, *e*2, [*e*3, *e*2]*}.
From the above sequences, we can get three diagnoses: {}, {*e*2}, {*e*2, *e*3}. *The minimal diagnosis is* {}, *which means no fault.*

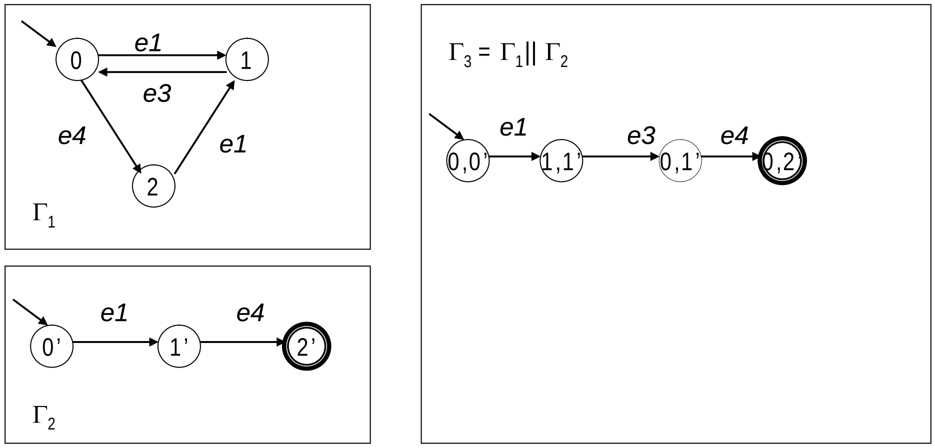
In Example 4, different trajectories give us different diagnoses. It can be no faults, or *e*2 (mapped to its fault), or both *e*2 and *e*3. They are all sound. Adding more observables is a way to clarify the ambiguity. To determine the observables for diagnosing a certain fault is the problem of diagnosability which is not covered in this article. Another example without ambiguity:

Example 5 *In Figure 2, Γ_1 is SD and Γ_2 is OBS. Γ_3 is* Diagnosis*. Since e3 is within the only trajectory, we can deduce that a fault represented by e3 occurred.*

We need to point out that the existing diagnosis methods for physical systems modeled as DES are not in general suitable for Web service processes. First, we cannot enumerate faults in Web service environments because we do not know how a Web service can be faulty if it belongs to another company. Second, it is relatively easy to keep a record for how the software is executed by recording any selected variables. In contrast, it is more difficult to insert a sensor in a physical system. Therefore it is very difficult to reconstruct the trajectories for a physical system, but it is not a key issue for diagnosing a Web service process. We will discuss the diagnosis of Web services in the Model-based Diagnosis for Web Service Processes section.

Several advances have recently been made: the decentralized diagnoser approach (Pencolé & Cordier, 2005) (a diagnosis system based on several interacting DESs); the incremental diagnosis approach (Grastien, Cordier, & Largouët, 2005) (a monitoring system that online updates

Figure 2. An example of Diagnosis



diagnosis over time given new observations); active system approaches (Baroni, Lamperti, Pogliano, & Zanella, 1999) (approaches that deal with hierarchical and asynchronized DESs); and diagnosis on reconfigurable systems (Grastien, Cordier, & Largouët, 2004). The existing techniques, such as the diagnoser approach (Pencolé, Cordier, & Rozé, 2002) or the silent closure (Baroni et al., 1999), reconstruct the unobservable behavior of the system that are required to compute diagnoses.

MODELING WEB SERVICE PROCESSES WITH DISCRETE-EVENT SYSTEMS

Description of the Web Service Processes

BPEL is an XML-based orchestration language developed by IBM and recognized by OASIS (Andrews et al., 2003). BPEL is a so-called executable language because it defines the internal behavior of a Web service process, as compared to choreography languages that define only the interactions among the Web services and are not executable.

BPEL defines fifteen activity types. Some of them are *basic activities*; the others are *structured activities*. Among the basic activities, the most important are the following:

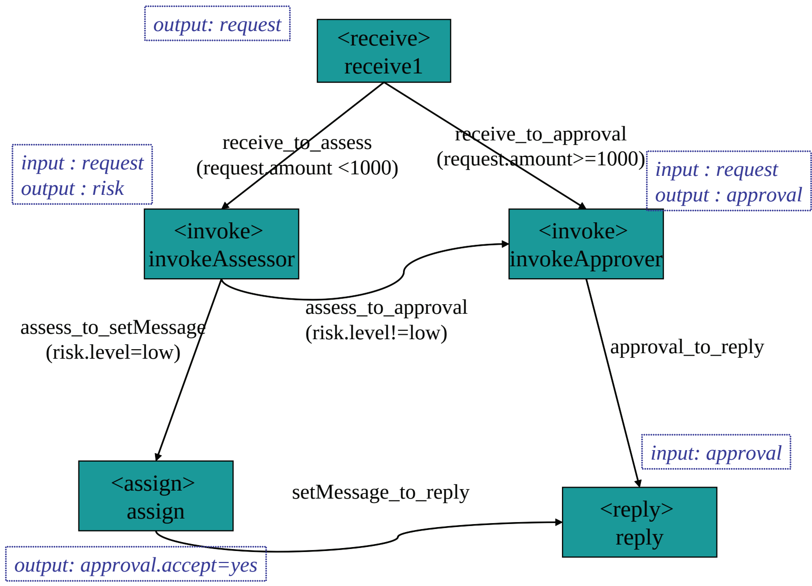
1. the ⟨receive⟩ activity is for accepting the triggering message from another Web service;
2. the ⟨reply⟩ activity is for returning the response to its requestor;
3. the ⟨invoke⟩ activity is for invoking another Web service.

The structured activities define the execution orders of the activities inside their scopes. For example:

- Ordinary sequential control between activities is provided by ⟨sequence⟩.
- Concurrency and synchronization between activities is provided by ⟨flow⟩.
- Loop is provided by ⟨while⟩.
- Nondeterministic choice based on external events is provided by ⟨pick⟩ and ⟨switch⟩.

Execution orders are also modified by defining the synchronization links between two activities (cf. Synchronization Links of Activities section).

Figure 3. A loan approval process. Activities are represented in shaded boxes. The inVar and outVar are respectively the input and output variables of an activity.



Normally, BPEL has one entry point to start the process and one point to exit, though multiple entry points are allowed. The variables in BPEL are actually the Simple Object Access Protocol (SOAP) messages defined in Web Service Description Language (WSDL). Therefore, the variables in BPEL are objects that have several attributes (called "parts" in WSDL).

An Example: the Loan Approval Process

Example 6 *The loan approval process is an example described in the BPEL Specification 1.1 (Andrews et al., 2003). It is diagrammed in Figure 3.*

This process contains five activities (big shaded blocks). An activity involves a set of input and output variables (dotted box besides each activity). All the variables are of composite type. The edges show the execution order of the activities. When two edges are issued from the same activity, only one edge that satisfies a triggering condition (shown on the edge) will be activated. In this example, the process is triggered when a ⟨receive⟩ activity named *receive1* receives a message of a predefined type. First, *receive1* initializes a variable *request*. Then, *receive1* dispatches the request to two ⟨invoke⟩ activities, *invokeAssessor* and *invokeApprover*, depending on the amount of the loan. In the case where the amount is large (*request.amount* >= 1000), *invokeApprover* is called for a decision. In the case where the amount is small (*request.amount* < 1000), *invokeAssessor* is called for risk assessment. If *invokeAssessor* returns with an assessment that the risk level is low (*risk.level = low*), a reply is prepared by an ⟨assign⟩ activity and later sent out by a ⟨reply⟩ activity. If the risk level is not low, *invokeApprover* is invoked for a final decision. The result from *invokeApprover* is sent to the client by the ⟨reply⟩ activity.

Modeling Web Services Process with Discrete-Event Systems

A Web service process defined in BPEL is a composition of activities. We are going to model a BPEL activity as an automaton. A BPEL code has a finite set of variables and a BPEL state is associated with an assignment of these variables. A BPEL activity is triggered when its initial state satisfies a finite set of triggering conditions which is a certain assignment of variables. After an activity is executed, the values of the state variables are changed. We need to extend the classic automaton definition to include the operations on state variables.

Assume a BPEL process has a finite set of variables $V = \{v_1, ..., v_n\}$, and the domain $D = \{D_1, ..., D_n\}$ for V is real values $\mathbf{R}$ or arbitrary strings. $C = \{c_1, ..., c_m\}$ is a finite set of constraints. A constraint c_j of some arity k is defined as a subset of the cartesian product over variables $\{v_{ji}, ..., v_{jk}\} \subseteq V$, that is, $c_j \subseteq D_{j1} \times \cdots \times D_{jk}$, or a first order formula over $\{v_{ji}, ..., v_{jk}\}$. A constraint restricts the possible values of the k variables.

A BPEL state s is defined as an assignment of variables. A BPEL transition t is an operation on the state s_i, that is,, $(s_j, post(V_2)) = t(s_i, e, pre(V_1))$, where $V_1 \subseteq V$, $V_2 \subseteq V$, $pre(V_1) \subseteq C$ is a set of preconditions that s_i has to satisfy and $post(V_2) \subseteq C$ is a set of post-conditions that the successor state s_j will satisfy. In another word, the transition t is triggered only when the starting state satisfies the preconditions, and the operation of this transition results in a state that satisfies the post-conditions. If a state s satisfies a constraint c, we annotate as $c \wedge s$. Then, the semantics of transition t is also represented as:

$$t : (s_i \wedge pre(V_1)) \xrightarrow{e} (s_j \wedge post(V_2)).$$

Definition 9 *A **BPEL activity** is an automaton* $\langle X, \Sigma, T, I, F, C\rangle$*, where C is a constraint set that defines states X and T*: $X \times \Sigma \times 2^C \rightarrow X \times 2^C$.

Modeling Basic Activities

In the following, we enumerate the model for each basic activity.

Activity ⟨receive⟩: $\langle\{s_o, s_f\}, \{received\}, \{t\}, \{s_o\}, \{s_f\}, C\rangle$ with
$t : (s_o \wedge SoapMsg.type = MsgType) \xrightarrow{received} (s_f \wedge RecMsg = SoapMsg)$, where *MsgType* is a predefined message type. If the incoming message *SoapMsg* has the predefined type, *RecMsg* is initialized as *SoapMsg*.

Activity ⟨reply⟩: $\langle\{s_o, s_f\}, \{replied\}, \{t\}, \{s_o\}, \{s_f\}, C\rangle$ with
$t : (s_o \wedge exists(RepMsg)) \xrightarrow{replied} (SoapMsg = RepMsg \wedge s_f)$, where $exists(RepMsg)$ is the predicate checking that the replay message *RepMsg* is initialized. *SoapMsg* is the message on the wire.

Activity ⟨invoke⟩
Synchronous invocation: $\langle\{s_o, wait, s_f\}, \{invoked, received\}, \{t_1, t_2\}, \{s_o\}, \{s_f\}, C\rangle$with
$t_1 : (s_o \wedge exists(InVar)) \xrightarrow{invoked} (wait)$, and
$t_2 : (wait \wedge SoapMsg.type = MsgType) \xrightarrow{received} (s_f \wedge exists(OutVar))$ where *InVar* and *OutVar* are the input and output variables.
Asynchronous invocation: $\langle\{s_o, s_f\}, \{invoked\}, \{t\}, \{s_o\}, \{s_f\}, C\rangle$ with
$t : (s_o \wedge exists(InVar)) \xrightarrow{invoked} (s_f)$,
asynchronous invocation does not wait for a return message.

Activity ⟨assign⟩: $\langle\{s_o, s_f\}, \{assigned\}, \{t\}, \{s_o\}, \{s_f\}, C\rangle$ with
$t : (s_o \wedge exists(InVar)) \xrightarrow{assigned} (s_f \wedge OutVar = InVar)$

Activity ⟨throw⟩: $\langle\{s_o, s_f\}, \{thrown\}, \{t\}, \{s_o\}, \{s_f\}, C\rangle$ with
$t : (s_o \wedge Fault.mode = Off) \xrightarrow{thrown} (s_f \wedge Fault.mode = On)$

Activity ⟨wait⟩: ⟨$\{s_o, wait, s_f\}$, $\{waiting, waited\}$, $\{t_1, t_2\}$, $\{s_o\}$, $\{s_f\}$, C⟩ with

$t_1 : (s_o \wedge Wait\_mode = Off) \xrightarrow{waiting} (wait \wedge Wait\_mode = On)$

$t_2 : (wait \wedge Wait\_mode = On) \xrightarrow{waited} (s_f \wedge Wait\_mode = Off)$

This model is not temporal. We do not consider time, so the notion of delay is not considered in this activity.

Activity ⟨empty⟩: ⟨$\{s_o, s_f\}$, $\{empty\}$, $\{t\}$, $\{s_o\}$, $\{s_f\}$, C⟩

$t : (s_o) \xrightarrow{empty} (s_f)$

Modeling Structured Activities

Structured activities nest other activities. We can model the structured activities as automata. Note that any automaton modeling a basic activity or a structured activity has only one initial state and one final state. In the following are the automata for the structured activities.

Sequence

A ⟨sequence⟩ can nest n activities ⟨A_i⟩ in its scope. These activities are executed in sequential order. Assume ⟨A_i⟩ : ⟨S_{Ai}, Σ_{Ai}, T_{Ai}, $\{s_{Aio}\}$, $\{s_{Aif}\}$, C_{Ai}⟩, $i \in \{1, ..., n\}$.

Activity ⟨sequence⟩: ⟨$\{s_o, s_f\} \cup \bigcup S_{Ai}$, $\{end\} \cup \bigcup \{callA_i\} \cup \bigcup \Sigma_{Ai}$, $\{t_i\} \cup \bigcup T_{Ai}$, $\{s_o\}$, $\{s_f\}$, $\bigcup C_{Ai}$⟩ with

$t_0 : (s_o) \xrightarrow{callA_1} (s_{A1o})$

$t_i : (s_{Aif}) \xrightarrow{callA_{i+1}} (s_{Ai+1o})$

$t_n : (s_{Anf}) \xrightarrow{end} (s_f)$

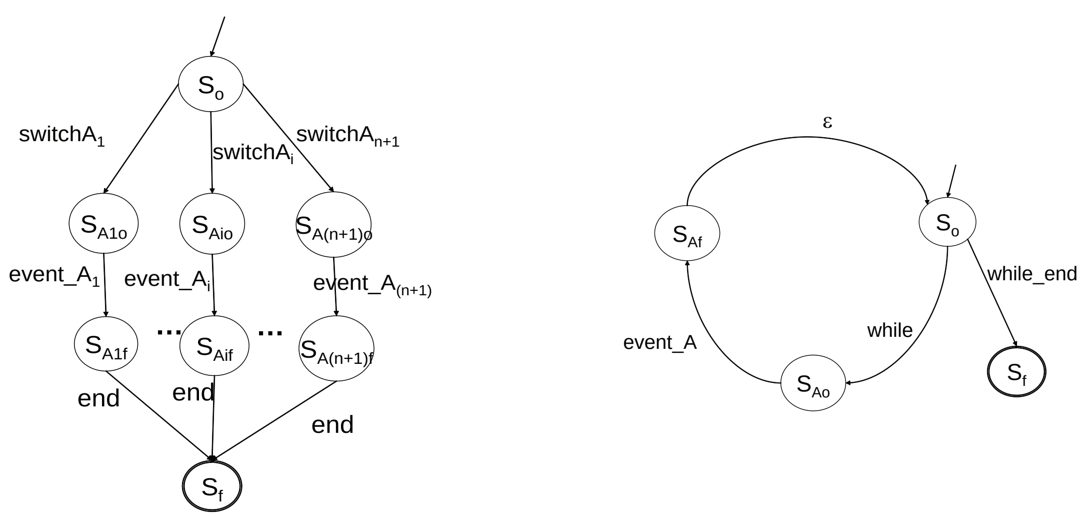
If assume $s_o = s_{A1o}$, $s_f = s_{Anf}$, and $s_{Aif} = s_{Ai+1o}$, for $i = [1, ..., n\text{-}1]$, a short representation of ⟨sequence⟩ is the concatenation of the nested activities A1◦A2 ⋯ ◦An.

Switch

Assume a ⟨switch⟩ has n ⟨case⟩ branches and one ⟨otherwise⟩ branch (see Figure 4(a)). Assume ⟨A_i⟩ : ⟨S_{Ai}, Σ_{Ai}, T_{Ai}, $\{s_{Aio}\}$, $\{s_{Aif}\}$, C_{Ai}⟩, i $\in \{1, ..., n+1\}$.

Activity ⟨switch⟩: ⟨$\{s_o, s_f\} \cup \bigcup S_{Ai}$, $\{end\} \cup \bigcup \{switchA_i\} \cup \bigcup \Sigma_{Ai}$, $\bigcup \{t_{io}\} \cup \bigcup \{t_{if}\} \cup \bigcup T_{Ai}$, $\{s_o\}$, $\{s_f\}$, $\bigcup C_{Ai}$, $\cup \bigcup pre(V_i)$⟩.

Figure 4. The automata for ⟨switch⟩ and ⟨while⟩



(a)The automaton for ⟨switch⟩ *(b)The automaton for ⟨while⟩*

Assume V_1, ..., V_n are variable sets on n ⟨case⟩ branches, $pre(V_1), \ldots, pre(V_n)$ are the constraints defined by the attributes condition in ⟨case⟩. The transitions are defined as:

$$t_{io} : (s_o \wedge \neg pre(V_1) \wedge \cdots \wedge pre(V_i) \cdots \wedge \neg pre(V_n)) \xrightarrow{switchA_i} (s_{Aio}), \forall i \in \{1, ..., n\}$$
$$t_{(n+1)o} : (s_o \wedge \neg pre(V_1) \wedge \cdots \wedge \neg pre(V_i) \cdots \wedge \neg pre(V_n)) \xrightarrow{switchA_{n+1}} (s_{A(n+1)o})$$
$$t_{if} : (s_{Aif}) \xrightarrow{end} (s_f), \forall i \in \{1,..., n+1\}$$

While Assume ⟨while⟩ nests an activity ⟨A⟩: ⟨S_A, Σ_A, T_A, $\{s_{Ao}\}$, $\{s_{Af}\}$, C⟩ (see Figure 4(b)).
Activity ⟨while⟩: $\{s_o, s_f\} \cup S_A$, $\{while, while\_end\} \cup \Sigma_A$, $\{t_o, t_f, t\} \cup T_A$, $\{s_o\}$, $\{s_f\}$, $C \cup pre(W)$⟩.

Assume W is a variable set, and $pre(W)$ is the constraint defined by the attribute condition in ⟨while⟩.

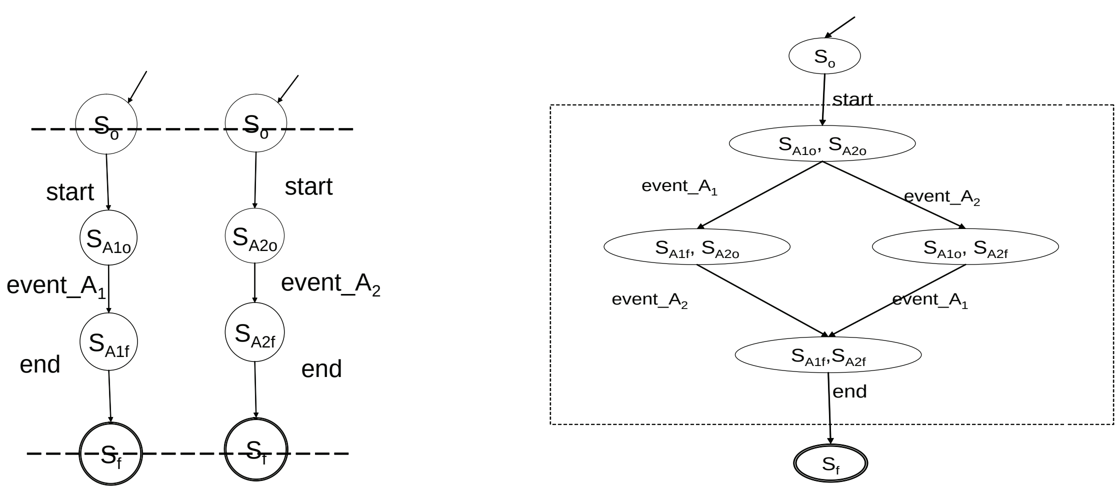
$$t_o : (s_o \wedge pre(W)) \xrightarrow{while} (s_{Ao})$$
$$t_f : (s_o \wedge \neg pre(W)) \xrightarrow{while\_end} (s_f)$$
$$t : (s_{Af}) \xrightarrow{\varepsilon} (s_o)$$

Flow
A ⟨flow⟩ can nest n activities ⟨A_i⟩ in its scope. These activities are executed concurrently. Assume ⟨A_i⟩ : ⟨S_{Ai}, Σ_{Ai}, T_{Ai}, $\{s_{Aio}\}$, $\{s_{Aif}\}$, C_{Ai}⟩, i $\in\{1, ..., n\}$.
Activity ⟨flow⟩: ⟨$\{s_o, s_f\} \cup \bigcup S_{Ai}$, $\{start, end\} \cup \bigcup \Sigma_{Ai}$, $\bigcup \{t_{io}, t_{if}\} \cup \bigcup T_{Ai}$, $\{s_o\}$, $\{s_f\}$, $\bigcup C_{Ai}$⟩ with

$$t_{io} : (s_o) \xrightarrow{start} (s_{Aio})$$
$$t_{if} : (s_{Aif}) \xrightarrow{end} (s_f)$$

Notice that the semantic of automata cannot model concurrency. We actually model the n-paralleled branches into n automata and define synchronization events to build their connections. The principle is illustrated in Figure 5. At the left, each branch is modeled as an individual automaton. The entry state s_o and the end state s_f are duplicated in each branch. Events *start* and *end* are the synchronization events. At the right is the automaton resulted by synchronization. More complicated case in joining the paralleled branches is discussed in subsection Synchronization Links of Activities. The key point in reasoning about decentralized automata is to postpone the synchronization until a synthesis result is needed,

Figure 5. Build concurrency as synchronized DES pieces



(a) Concurrency branches for DES pieces

(b) The joint DES model

in order to avoid the state explosion problem (Pencolé et al., 2002; Pencolé & Cordier, 2005). In Web service diagnosis, it is the situation (cf. subsection Multiple Exceptions).

Pick

Assume a ⟨Pick⟩ has n ⟨onMessage⟩ and one ⟨onAlarm⟩ branches. ⟨onMessage⟩ branches are triggered by predefined events. Assume activities $\{A_1,..., A_n\}$ are corresponding to the n branches respectively. ⟨onAlarm⟩ branch is triggered by a time-out event produced by a timer. Assume activity A_{n+1} is corresponding to ⟨onAlarm⟩ branch. Exactly one branch will be selected based on the occurrence of the event associated with before any others. Assume $\langle A_i \rangle : \langle S_{Ai}, \Sigma_{Ai}, T_{Ai}, \{s_{Aio}\}, \{s_{Aif}\}, C_{Ai}\rangle$, i $\in \{1, ..., n+1\}$.

Activity ⟨pick⟩: $\langle \{s_o, s_f\} \cup \bigcup S_{Ai}, \bigcup\{start_{Ai}\} \cup \{end\} \cup \bigcup \Sigma_{Ai}, \bigcup\{t_{io}, t_{if}\} \cup \bigcup T_{Ai}, \{s_o\}, \{s_f\}, \bigcup C_{Ai} \cup \bigcup exists(\ event_{Ai})\rangle$ with

$t_{io} : (s_o \wedge exists(event_{Ai})) \xrightarrow{start_{Ai}} (s_{Aio})$

$t_{if} : (s_{Aif}) \xrightarrow{end} (s_f)$

Synchronization Links of Activities

Each BPEL activity can optionally nest the standard elements ⟨source⟩ and ⟨target⟩. The XML grammar is defined as:

<*source linkName* =“*ncname*” *transitionCondition* = “*bool − expr*”?/>

<*target linkName* =“*ncname*”/>

A pair of ⟨source⟩ and ⟨target⟩ defines a link which connects two activities. The target activity must wait until the source activity finishes. Therefore, links define the sequential orders of activities. When one ⟨flow⟩ contains two parallel activities which are connected by a link, the two activities become sequentially ordered. An activity may have multiple ⟨source⟩ or ⟨target⟩ elements. Links can express richer logics, but they make the processes more difficult to analyse.

⟨source⟩ can be modeled similarly like an ⟨activity⟩, with “*transitionCondition*” as the triggering condition.

Activity ⟨source⟩: $\langle\{s_o, s_f\}, \{\varepsilon\}, \{t\}, \{s_o\}, \{s_f\},$ *transitionCondition*⟩ with

$t : (s_o \wedge transitionCondition) \xrightarrow{\varepsilon} (s_f)$,

When an activity is the ⟨target⟩ of multiple links, a join condition is used to specify how these links can join. The join condition is defined within the activity. BPEL specification defines standard attributes for this activity:

<*activityName* =“*ncname*”, *joinCondition* =“*bool − expr*”, *suppressJoinF ailure* =“*yes|no*”/>

where *joinCondition* is the logical OR of the liveness status of all links that are targeted at this activity. If the condition is not satisfied, the activity is bypassed, and a fault is thrown if *suppressJoinFailure* is no.

In this case, the synchronization event end as in Figure 5(a) is removed. If the ending state of ⟨flow⟩ is the starting state s_o’ of the next activity, the precondition of s_o’ is the *joinCondition*. For example, either of the endings of the two branches can trigger the next activity can be represented as: $s_o' \wedge (exists(s_{A1f}) \vee exists(s_{A2f}))$.

Modeling the Loan Approval Process

In this section, we present the complete DES model for the process in Example 6.

Example 7 *The loan approval process in Example 6 contains five activities:* ⟨receive1⟩, ⟨invokeAssessor⟩, ⟨invokeApprover⟩, ⟨assign⟩, ⟨reply⟩. *The five activities are contained in a* ⟨flow⟩. *Six links*, ⟨receive_to_assess⟩, ⟨receive_to_approval⟩, ⟨assess_to_setMessage⟩,

$\langle$assess_to_approval$\rangle$, $\langle$approval_to_reply$\rangle$, *and* $\langle$setMessage_to_reply$\rangle$, *connect the activities and change the concurrent orders to sequential orders between the activities. In this special case, there are actually no concurrent activities. Therefore, for clarity, the event caused by* $\langle$flow$\rangle$ *is not shown. Assume the approver may return an error message due to an unknown error. Here is the formal representation of the process (also reference to Figure 6).*

$\langle$receive1$\rangle$ = $\langle\{x_0, x_1\}, \{received\}, \{t_1\}, \{x_0\}, \{x_1\}, C\rangle$, with

$t_1 : (x_0 \wedge SoapMsg.type = MsgType) \xrightarrow{received} (x_1 \wedge request = SoapMsg)$, where

MsgType is a predefined message type. If the incoming message *SoapMsg* has the predefined type, request is initialized as *SoapMsg.*

$\langle$receive_to_assess$\rangle$ = $\langle\{x_1, x_2\}, \{\varepsilon\}, \{t_2\}, \{x_1\}, \{x_2\}, C\rangle$, with

$t_2 : (x_1 \wedge request.amount < 1000) \xrightarrow{\varepsilon} (x_2)$.

$\langle$receive_to_approval$\rangle$ = $\langle\{x_1, x_3\}, \{\varepsilon\}, \{t_3\}, \{x_1\}, \{x_3\}, C\rangle$, with

$t_3 : (x_1 \wedge request.amount \geq 1000) \xrightarrow{\varepsilon} (x_3)$.

$\langle$invokeAssessor$\rangle$ = $\langle\{x_2, x_4, x_5\}, \{invoked\_assessor, received\_risk\}, \{t_4, t_5\}, \{x_2\}, \{x_5\}, C\rangle$ with

$t_4 : (x_2 \wedge InVar = request) \xrightarrow{invoked\_assessor} (x_4)$, and

$t_5 : (x_4) \xrightarrow{received\_risk} (x_5 \wedge OutVar = risk)$ where *InVar* and *OutVar* are the input and output variables.

$\langle$assess_to_setMessage$\rangle$ = $\langle\{x_5, x_6\}, \{\varepsilon\}, \{t_6\}, \{x_5\}, \{x_6\}, C\rangle$, with

$t_6 : (x_5 \wedge risk.level = low) \xrightarrow{\varepsilon} (x_6)$.

$\langle$assess_to_approval$\rangle$ = $\langle\{x_5, x_3\}, \{\varepsilon\}, \{t_7\}, \{x_5\}, \{x_3\}, C\rangle$, with

$t_7 :(x_5 \wedge risk.level = high) \xrightarrow{\varepsilon} (x_3)$.

$\langle$invokeApprover$\rangle$ = $\langle\{x_3, x_7, x_8\}, \{invoked\_approver, received\_approval, received\_aplError\}, \{t_8, t_9, t_e\}, \{x_3\}, \{x_8\}, C\rangle$ with

$t_8 : (x_3 \wedge InVar = request) \xrightarrow{invoke\_approver} (x_7)$, and

$t_9 : (x_7) \xrightarrow{received\_approval} (x_8 \wedge OutVar = approval)$, and

$t_e : (x_7) \xrightarrow{received\_aplError} (x_8 \wedge OutVar = errorMessage)$ where

InVar and *OutVar* are the input and output variables.

$\langle$assign$\rangle$: $\langle\{x_6, x_9\}, \{assigned\}, \{t_{10}\}, \{x_6\}, \{x_9\}, C\rangle$ with

$t_{10} : (x_6) \xrightarrow{assigned} (x_9 \wedge approval.accept = yes)$

$\langle$setMessage_to_reply$\rangle$ = $\langle\{x_9, x_{10}\}, \{\varepsilon\}, \{t_{11}\}, \{x_9\}, \{x_{10}\}, C\rangle$, with

$t_{11} : (x_9) \xrightarrow{\varepsilon} (x_{10})$.

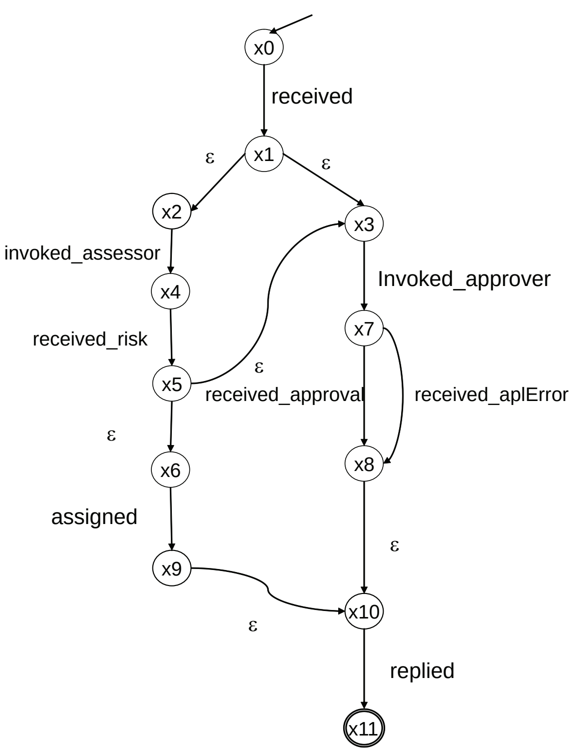
$\langle$approval_to_reply$\rangle$ = $\langle\{x_8, x_{10}\}, \{\varepsilon\}, \{t_{12}\}, \{x_8\}, \{x_{10}\}, C\rangle$, with

$t_{12} : (x_8) \xrightarrow{\varepsilon} (x_{10})$.

$\langle$reply$\rangle$: $\langle\{x_{10}, x_{11}\}, \{replied\}, \{t_{13}\}, \{x_{10}\}, \{x_{11}\}, C\rangle$ with

$t_{13} :(x_{10} \wedge exists(approval)) \xrightarrow{replied} (x_{11} \wedge SoapMsg = approval)$, where *SoapMsg* is the message on the wire.

Figure 6. Automaton modeling loan approval process



MODEL-BASED DIAGNOSIS FOR WEB SERVICE PROCESSES

A Web service process can run down for many reasons. For example, a composed Web service may be faulty, an incoming message mismatches the interface, or the Internet is down. The diagnosis task is to determine the Web services responsible for the exceptions. These Web services will be diagnosed faulty. In this article, our major effort is on diagnosing business logic-related exceptions.

In our framework, *COMPS* is made up of all the basic activities of the Web service process considered, and *OBS* is made up of the exceptions thrown and the events of the executed activities. These events can be obtained by the monitoring function of a BPEL engine. A typical correct model for an activity $\langle A \rangle$ is thus:

$$\neg ab(A) \land \neg ab(A.input) \Rightarrow \neg \mathrm{ab(A.output)} \quad (4)$$

For facilitating diagnosis, the BPEL engine has to be extended for the following tasks: 1) record the events emitted by executed activities; 2) record the input and output SOAP messages; and 3) record the exceptions and trigger the diagnosis function when the first exception is received. ***Diagnosing*** is triggered on the first occurred exception[3]. The MBD approach developed relies on the following three steps with the techniques we introduced in the content above.

1. A prior process modeling and variable dependency analysis

All the variables in BPEL are global variables, that is, they are accessible by all the activities. An activity can be regarded as a function that takes input variables and produces output variables. An activity has two kinds of relation to its input and output variables: defining and utilizing. We use *Def*(A, V) and *Util*(A, V) to present the relation that activity A defines variable V or utilizes V. An activity is normally a utilizer of its input variables, and is a definer of its output variables. This is similar to the viewpoint of programming slicing, a technique in software engineering for software debugging (cf. subsection A Brief Comparison to Program Slicing). But BPEL can violate this relation by applying some business logic. For example, some variables, such as order ID and customer address, are not changeable after they are initialized in a business process. Therefore, a BPEL activity may be a utilizer of its output variables. In BPEL, it is defined in *correlation sets.* In this case, we use *Util*(A, ($V1$, $V2$)) to express that output $V2$ is correlated to input $V1$. In this case, Formula 4 can be simplified as:

$$\neg ab(A.input) \Rightarrow \neg ab(A.output), \text{ if } Util(A, (A.input, A.output)) \quad (5)$$

In Example 8, we give a table to summarize the variable dependency for the load approval process. This table can be obtained automatically from BPEL. The approach is not presented due to lack of space.

Example 8 *The variable dependency analysis for the loan approval process is in Table 1.*

2. Trajectories reconstruction from observations after exceptions are detected

As mentioned earlier, the observations are the events and exceptions when a BPEL process is executed. The events can be recovered from the log file in a BPEL engine. The observations are formed in an automaton. The possible trajectories of the process are calculated by synchronizing the automaton of the observations with the automaton of the system description:

$$\text{trajectories} = \textit{trajectories of } SD||OBS \quad (6)$$

Table 1. The variable dependency analysis for the loan approval process

Variables	Parts	Definer	Utilizer
request	firstname lastname amount	receive1 receive1 receive1	invokeAssessor, invokeApprover invokeAssessor, invokeApprover invokeAssessor, invokeApprover
risk	level	invokeAssessor	
approval	accept	assign, invokeApprover	reply
error	errorCode	invokeApprover	

We do not require recording each event during the execution, but just enough to be able to identify the real trajectory of the process. This is very useful when some events are not observable and when there are too many events to record. Reference to the subsection Without Full Observability for more discussion.

Example 9 *In the loan approval example, assume that OBS={received, invoked_assessor, received_risk, invoked_approver, received_aplErr} (as in Figure 7(a)). Received_aplErr is an exception showing that there is a type mismatch in received parameters. We can build the trajectory of evolution as shown, also shown in Figure 7(b).*

$$(x_0) \xrightarrow{received} (x_1) \xrightarrow{\varepsilon} (x_2) \xrightarrow{invoked\_assessor} (x_4) \xrightarrow{received\_risk} (x_5) \xrightarrow{\varepsilon} (x_3) \xrightarrow{invoked\_approver} (x_7) \xrightarrow{received\_aplErr} (x_8)$$

3. Accountability analysis for mode assignment

Not all the activities in a trajectory are responsible for the exception. As a software system, the activities connect to each other by exchanging variables. Only the activities which change the attributes within a variable can be responsible for the exception.

Assume that activity A generates exception e_f, and t is a trajectory ending at A. The responsibility propagation rules are (direct consequences of the contraposition of Formula 4 and 5):

$$e_f \in \Sigma_A \ |\text{-}\ ab(A) \vee \vee\{ab(A.InVar.part) | A.InVar.part \in A.InVar\} \quad (7)$$

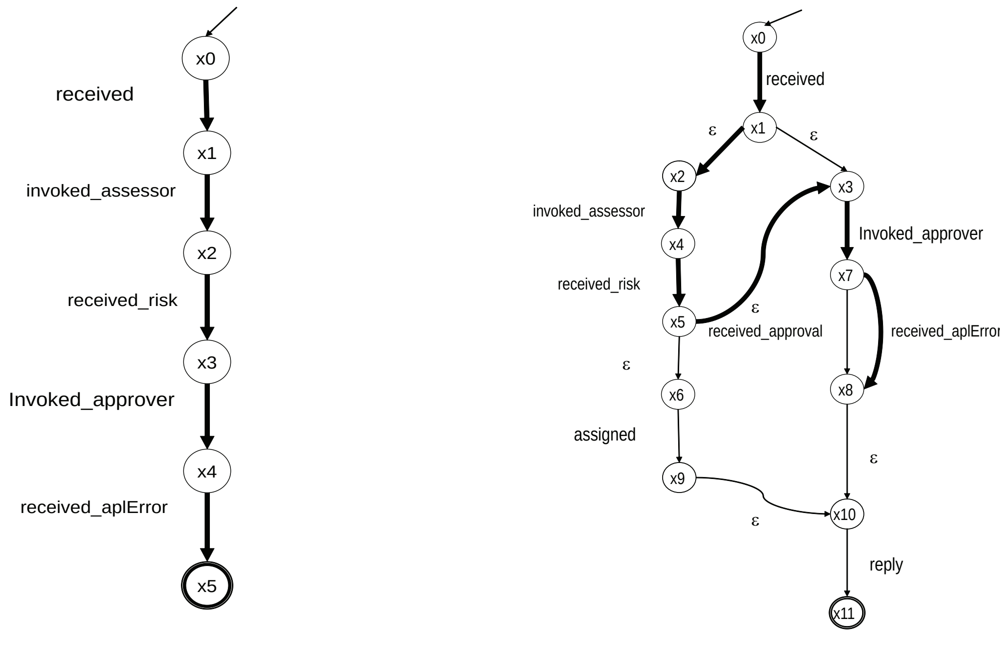
$\forall A_i, A_j \in t, A_j \neq A_i$, A_j is the only activity between A_j and A_i such that $Def(A_j, A_i.InVar.part)$,

$$ab(Ai.InVar.part) \ |\text{-}\ ab(A_j) \vee \vee\{ab(A_j.InVar.part) | A_j.InVar.part \in A_j.InVar\} \quad (8)$$

The first rule in (7) states that if an activity A generates an exception e_f, it is possible that activity A itself is faulty, or any part in its *A.InVar* is abnormal. Notice a variable is a SOAP message which has several parts. *A.InVar.part* is a part in *A.InVar*[4]. The second rule in (8) propagates the responsibility backwards in the trajectory. It states that an activity $A_j \in t$ that defines a part of A_i.*InVar* which is known as faulty could be faulty; and its inputs could also be faulty. If there are several activities that define a part of A_i.*InVar*, only the last one counts, because it overrides the changes made by the other activities, that is, A_j is the last activity "between" A_j and A_i that defines A_i.*InVar*, as stated in (8). After responsibility propagation, we obtain a *responsible set* of activities $RS = \{A_i\} \subseteq t$.

The set $CO = \{A\} \cup \{A_i | A_i \in RS\}$ is a conflict set, because if all the components in *CO* are correct, there should be no exceptions. Then a diagnosis is any of A or A_i in the responsible set is faulty:

Figure 7. (a) the observations; (b) the loan approval process evolution trajectory up to the exception



(a) *(b)*

$$\{D_f\} = \{\{A\}\} \cup \{\{A_i\} | A_i \in RS\} \qquad (9)$$

Each D_f is a single fault diagnosis and the result is the disjunct of the D_f. The algorithm is as following. Lines 1-2 apply rule (7). Lines 3-8 apply rule (8). This algorithm checks each activity in *t*. Therefore the complexity of this algorithm is $O(|t|)$.

Example 10 *For the loan approval example, we have the trajectory as in Example 9. We do the responsibility propagation. As invokeApprover generates the exception, according to Formula (7), invokeApprover is possibly faulty. Then its input request is possibly faulty. Among all the activities {receive1, invokeAssessor, invokeApprover} in the trajectory, receive1 defines request, invokeAssessor and invokeApprover utilize request. Therefore, receive1 is possibly faulty, according to Formula (8). receive1 is the first activity in the trajectory. The propagation stops. The diagnosis is:*

$$\{D_f\} = \{\{receive1\}, \{invokeApprover\}\}$$

Example 10 has two single faults {*receive1*} and {*invokeApprover*} for the exception *received_aplErr*, which means either the activity ⟨receive1⟩ or ⟨invokeApprover⟩ is faulty. In an empirical way, an engineer may associate only one fault for an exception. But our approach can find all possibilities. Second, if we want to further identify which activity is indeed responsible for the exception, we can do a further test on the data. For example, if the problem is wrong data format, we can verify the data format against some specification, and then identify which activity is faulty.

Multiple Exceptions

There are two scenarios where multiple exceptions can happen. The first scenario is the chained exceptions when one exception causes the others to happen. Normally the software reports this chained relation. We need to diagnose only the first occurred exception, because the causal relations for other exceptions are obvious from the chain.

The second scenario is the case when exceptions occur independently, for example, two paralleled branches report exceptions. As the exceptions are independent, we diagnose each exception independently, the synthesis diagnoses are the union of all the diagnoses. Assume the minimal diagnoses for exception 1 are $\{D_i^1\}$, where $i \in [1, ..., n]$, and the minimal diagnoses for exception 2 are $\{D_j^2\}$, where $j \in [1, ..., m]$, the synthesis diagnoses are any combinations of D_i^1 and D_j^2: $\{D_i^1 \cup D_j^2 | i \in [1, ..., n], j \in [1, ..., m]\}$.

What interests us most is the synthesis of the minimal diagnoses. So, we remove the $D_i^1 \cup D_j^2$ that are supersets of other ones. This happens only if at least one activity is common to $\{D_i^1\}$ and $\{D_j^2\}$, giving rise to a single fault that can be responsible for both exceptions. Such activities are thus most likely to be faulty (single faults being preferred to double faults).

Without Full Observability

Equation 6 can recover trajectories from *OBS*. Actually if we can record all the events in a model, trajectories are equal to *OBS*. It is a trivial case. The problem occurs when we do not have full observability. For example, a third party BEPL engine does not allow us to record all the events crucial for diagnosis, or the process is too large to record every event. Equation 6 gets all the possible trajectories satisfying *OBS*. Therefore, this method can deal with missing events. At the meantime, if there are multiple trajectories satisfying *OBS*, the diagnoses are the union of the diagnoses obtained from all the trajectories. This can result in a larger number of diagnoses, that is, diagnosis is not precise.

It is a trade off between observability and diagnosability. Increasing observability, that is, observing more events, can result in more precise diagnosis, while increasing the observing cost. It is our future work to study the minimal observables for diagnosing a fault.

Offline Diagnosability Analysis

Diagnosability analysis is normally conducted offline without executing the processes. We do not touch diagnosability analysis problems in this article. But diagnosability is related to the method of diagnosis. Assuming an exception at a place in a BPEL process, diagnosability analysis of this exception involves calculating all the trajectories from the process entry point to the assumed exception and find diagnoses on each trajectory. The method is similar as the three steps in the Model-based Diagnosis for Web Service Processes section, just the second step is replaced by a graph traverse algorithm to compute all the trajectories between two nodes on the graph formed by the automaton model.

Multiple Trajectories

Lack of full observability and offline diagnosability analysis can cause multiple trajectories. Assume trajectories $\{t_1, ..., t_n\}$. Using our diagnosis algorithm, each trajectory t_i has conflict set CO_i. But as the trajectories are the possible execution paths, they do not occur at the same time, the conflict sets are not all contradictory at the same time. Indeed only one of these trajectories, even if which one is unknown, really happened. In this case, we do not have to use hitting set algorithm to compute diagnoses. We define simply the synthesis diagnoses as the disjunction of all the diagnoses, $\vee\{D_i\}$, which means diagnoses are in any of $\{D_i\}$.

Obtaining the Dependency Table

The variable dependency table can be automatically constructed from BPEL. Regard a BPEL activity $\langle A \rangle$ as a function $OutVar = f_A(InVar)$. Then $\langle A \rangle$ is the utilizer of *InVar* and definer of *OutVar*.

Before, we have defined $\langle A \rangle$ as an automaton. Then *InVar* is the variables used in so and *Outvar* is the variables used in s_f.

Due to some business logic, some variables, such as order ID and customer address, are not changeable after they are initialized in a business process. BPEL uses correlation set to define that two variables are identical in values. The *correlation set* is referenced by an activity. When an activity has a correlation set within its scope, the correlation indicates if this activity initiates the variables by setting the attribute initiate. If initiate is "yes", this activity is the definer for both of the variables; otherwise, this activity is the utilizer for both of the variables.

Implementation

There are many BPEL engines in the market. We extended ActiveBPEL (Active Endpoint, 2007), an open source from Active Endpoints, to implement our diagnosis mechanism. ActiveBPEL allows us to record every executed activity and messages in the execution. These activities and messages are the observations during execution and they correspond to a subset of the events and states in our formal model. Therefore, from the synchronization of the observations and the formal model result the execution trajectories. The diagnosis function is a java package that is invoked when an exception is caught. It takes the observations and the dependency table as inputs, calculates the trajectories and uses Algorithm 1 to calculate diagnoses.

RELATED WORK AND DISCUSSION

A Brief Comparison to Program Slicing

Program slicing is a well known technique in software engineering for software debugging (Weise, 1984). If we have a specific program Π, a location within this program #*n* (*n* is a number given to a line), and a variable x, then a slice is itself a program that is obtained from the original program by removing all statements that have no influence on the given variable at the specified position. Since slices are usually smaller than the original program they focus the user's attention on relevant parts of the program during

Algorithm 1. Calculate diagnosis for a faulty Web service process

INPUT: A_0 - the activity generating the exception.
t - a list of activities in a reserved trajectory ending at A_0, taken in reverse order with A_0 excluded.
Variables: *V* - a list of faulty variable parts, initialized as {}.
OUTPUT: *D* - the list of all possible faulty activities, initialized as $\{A_0\}$.
Notes about the algorithm: 1) list.next() returns the first element of a list; list.add(element) adds an element at the end of the list; list.remove(element) removes an element from the list. 2) Activity *A* has a list of input variables *A.InVars* and output variables *A.OutVars*. 3) a variable *var* has a list of parts *var.Parts*.

```
for each variable var in A0.InVars do
    V.add(var.Parts)
while A = t.next()! = null do
    if ∃p ∈ V, Def(A, p) then
        D.add(A)
        V.remove(v)
        for each variable var in A.InVars do
            V.add(var.Parts)
return D
```

debugging. Slices can be computed from Program Dependence Graph (PDG) (Ottenstein & Ottenstein, 1984) as a graph reachability problem. A PDG G_Π for a program Π is a direct graph. The vertices of G_Π represent assignment statements and control predicates that occur in program Π. In addition G_Π includes the distinguished *entry* vertex. The edges of the graph represent either control or data dependencies. Given a criterion $\langle n, x\rangle$, the slice is computed in two steps. First, the vertex *v* representing the last program position before n where variable x is defined must be localized. Second, the algorithm collects all vertices that can reach v via a control or flow dependency edge. The statements represented by the collected vertices (including vertex *v*) are equal to the program slice for Π.

Wotawa has discussed the relationship between MBD based debugging and program slicing (Wotawa, 2002). In his work, each statement in a program is viewed as a component with a mode, inputs and outputs. The logic representation of a statement #*n* is $\neg ab(n) \rightarrow out(n)= f(in(n))$, that is, if #*n* is not faulty, the output *out*(*n*) is a function of the input in(n) according to the syntax of the program. He observed that the strongly connected components in the PDG have an influence one each other. Only if all the components are not faulty, the super component composed by these components is not faulty. He defined a dependency model whose nodes are the strongly connected components in the PDG and added a logic rule to describe the relation between the super component and the components within it. Assume $\{s_1,s_2, ..., s_n\}$are strongly connected and the name of the super component is *SC*, then the rule is $\neg ab(s_1) \wedge ... \wedge\neg ab(s_n) \rightarrow \neg ab(SC)$. With the additional rule, logic deduction can more precisely identify the faulty components. Under this kind of modeling, slices of a single variable are equivalent to conflicts in MBD. And MBD and program slicing should draw equivalent conclusions on which statements are faulty.

We consider that diagnosing Web service processes is not equivalent to program debugging. First, we are interested in the faults due to the unknown behavior of the external Web services and due to the interaction between Web services. We assume that the Web service processes are described correctly in BPEL or a Web service process description language. This implicitly excludes the structured activities to be faulty. This is equivalent to consider only data dependency in program slice. Second, though Web service process description languages are like programs, they are simpler than programs. For example, they do not use pointers or other complicated data structures as in programs, and they do not use *Goto* and its other restricted forms as in unstructured program. This makes it possible that diagnosing Web service processes can be simpler than diagnosing programs.

The diagnosis method developed in this article can be compared to dynamic slicing introduced in (Korel & Laski, 1988). Similar to our method, dynamic slicing considers the bugs should be within the statements that actually affect the value of a variable at a program point for a particular execution of the program. Their solution, following after Weiser's static slicing algorithm, solves the problem using data-flow equations, which is also similar to the variable dependency analysis presented in this article, but not the same. An external Web service can be seen as a procedure in a program, with unknown behavior. For a procedure, we normally consider the outputs brought back by a procedure are generated according to the inputs. Therefore, in slicing, the outputs are considered in the definition set (the set of the variables modified by the statement). For Web services, we can know some parts in SOAP response back from a Web service should be unchanged, for example, the name and the address of a client. This relation is defined as correlation set. We should point out that the variable dependency analysis in this article is different from slicing. As a consequence, the diagnosis obtained from MBD approach in this article can be different from slicing, and actually more precise.

As MBD approach can integrate more business logic into its model, it is less rigid than slicing. In this sense, MBD is more business oriented, not program oriented, which makes it more suitable for diagnosing Web service processes than slicing.

MBD in Diagnosing Component-Based Software

Besides Wotawa's work mentioned above, some other people have applied MBD on diagnosing component-based software systems. We found that when diagnosing such systems, the modeling is rather at the component level than translating lines of statements into logic representations. Grosclaude in (Grosclaude, 2004) used a formalism based on Petri nets to model the behaviors of component-based systems. It is assumed that only some of the events are monitored. The history of execution is reconstructed from the monitored events by connecting pieces of activities into possible trajectories. Console's group is working towards the same goal of monitoring and diagnosing Web services like us. In their article (Ardissono, Console, Goy, Petrone, Picardi, & Segnan, 2005), a monitoring and diagnosing method for choreographed Web service processes is developed. Unlike BPEL in our article, choreographed Web service processes have no central model and central monitoring mechanism. (Ardissono et al., 2005) adopted grey-box models for individual Web services, in which individual Web services expose the dependency relationships between their input and output parameters to public. The dependency relationships are used by the diagnosers to determine the responsibility for exceptions. This abstract view could be not sufficient when dealing with highly interacting components. More specifically, if most of the Web services claim too coarsely that their outputs are dependent on their inputs, which is correct, the method in (Ardissono et al., 2005) could diagnose almost all the Web services as faulty. Yan et al. (Yan, Pencolé, Cordier, & Grastien, 2005) is our preliminary work to the present one, focusing on Web service modeling using transition systems. The major work in this article is to complete the monitoring and diagnosis methods and present the diagnosis algorithm. The syntax of modeling in this article is improved from (Yan et al., 2005) with simplified representation of states and explicit definition of constraints. As a result, the model for a process can be more readable and a slightly fewer states. This article is also self-contained with MBD background and discussions on fault management tasks for Web service processes.

Related Work in Web Service Monitoring, Modeling and Composition

Several groups of researchers work on Web service monitoring frameworks. (Baresi, Ghezzi, & Guinea, 2006) proposes BPEL[2] which is the original BPEL with monitoring rules. Monitoring rules define how the user wants to oversee the execution of BPEL. But (Baresi et al., 2006) does not specify the monitoring tasks. (Mahbub & Spanoudakis, 2004) proposes a framework for monitoring requirements of BPEL-based service compositions. Their approach uses event calculus for specifying the requirements that must be monitored. Requirements can be behavioral properties of the coordination process or assumptions about the atomic or joint behavior of the deployed services. Events, produced by the normal execution of the process, are stored in a database and the runtime checking is done by an algorithm based on integrity constraint checking in temporal deductive databases. These frameworks can be used for recording the events and messages used for diagnosis.

In addition to automata used in this article, Petri nets and process algebra are also used as formal models for Web service processes. For example, (Ferrara, 2004; Salaün, Bordeaux, &

Schaerf, 2004; Viroli, 2004) map BPEL into different Process Algebra; (Ouyang, Aalst, Breutel, Dumas, Hofstede, & Verbeek, 2005; Schmidt & Stahl, 2004) present different semantics of BEPL in Petri nets; (Fisteus, Fern´andez, & Kloos, 2004; Foster, Uchitel, Magee, & Kramer, 2003; Fu, Bultan, & Su, 2004) use automata to model BPEL for verification. These models have similar expression power and similar reasoning or computing techniques.

Web service composition techniques are relevant to this article because they generate new Web service processes. AI planning methods are the most commonly used techniques for Web service composition. (Narayanan & McIlraith, 2002) starts from DAML-S descriptions and automatically transforms them into Petri nets. Other works, as (Berardi, Calvanese, de Giacomo, Lenzerini, & Mecella, 2003; Lazovik, Aiello, & Papazoglou, 2003; Pistore, Traverso, Bertoli, & Marconi, 2005), rely on transition rules systems. (Rao & Su, 2004) is a survey article on automated Web service composition methods. Re-planning is relevant to this article because it can be used to modify the Web service processes for fault recovery. (Canfora, Penta, Esposito, & Willani, 2005) presents a re-planning technique based on slicing techniques. When the estimated QoS metrics are not satisfied, the re-planning selects other Web services to replace the ones in the process.

CONCLUSION

Web services are the emergent technology for business process integration. A business process can be composed of distributed Web services. The interactions among the Web services are based on message passing. To identify the Web services that are responsible for a failed business process is important for e-business applications. Existing throw-and-catch fault handling mechanism is an empirical mechanism that does not provide sound and complete diagnosis. In this article, we developed a monitoring and diagnosis mechanism based on solid theories in MBD. Automata are used to give a formal modeling of Web service processes described in BPEL. We adapted the existing MBD techniques for DES to diagnose Web service processes. Web service processes have all the features of software systems and do not appear to function abnormally until an exception is thrown and they are stopped, which makes the diagnosis principle different from diagnosing physical systems. The approach developed here reconstructs execution trajectories based on the model of the process and the observations from the execution. The variable dependency relations are utilized to deduce the actual Web services within a trajectory responsible for the thrown exceptions. The approach is sound and complete in the context of modeled behavior. A BPEL engine can be extended for the monitoring and diagnosis approach developed in this article.

REFERENCES

Active Endpoint. (2007). Retrieved from, *www.active-endpoints.com/active-bpel-engineoverview.htm*.

Andrews, T., Curbera, F., Dholakia, H., Goland, Y. et.al. (2003). Business process execution language for web services (bpel4ws) 1.1. (*ftp://www6.software.ibm.com/software/developer/library/ws-bpel.pdf, retrieved April 10, 2005*)

Ardissono, L., Console, L., Goy, A., Petrone, G., Picardi, C., & Segnan, M. (2005). Cooperative model-based diagnosis of web services. *In Proceedings of the 16th international workshop on principles of diagnosis (DX-2005)* (pp. 125–132).

Baresi, L., Ghezzi, C., & Guinea, S. (2006). Towards self-healing compositions of services. In B. J. Krämer & W. A. Halang (Eds.), *Contributions to ubiquitous computing* (Vol. 42). Springer.

Baroni, P., Lamperti, G., Pogliano, P., & Zanella, M. (1999). Diagnosis of large active systems. *Artificial Intelligence*, *110*(1), pp. 135–183.

Berardi, D., Calvanese, D., De Giacomo, G., Lenzerini, M., & Mecella, M. (2003). Automated composition of e-services that export their behavior. In *Proceedings of the 1st int. conf. on service-oriented computing (icsoc'03)* (pp. 43-58).

Canfora, G., Penta, M. D., Esposito, R., & Willani, M. L. (2005). Qos-aware replanning of composite web services. In *Proceedings of IEEE international conference on web services.*

Cordier, M. & Thiébaux, S. (1994). Event-based diagnosis for evolutive systems. In *Proceedings of the fifth international workshop on principles of diagnosis (DX'94)* (pp. 64–69).

Ferrara, A. (2004). Web services: a process algebra approach. In *Proceedings of the 2nd international conference on service oriented computing (icsoc)* (p. 242-251). New York: ACM Press.

Fisteus, J., Fernández, L., & Kloos, C. (2004). Formal verification of bpel4ws business collaborations. In K. Bauknecht, M. Bichler, & B. Prll (Eds.), *Proc. of 5th international conference e-commerce and web technologies (ec-web).*

Foster, H., Uchitel, S., Magee, J., & Kramer, J. (2003). Model-based verification of web service compositions. In *Proc. of eighteenth IEEE international conference on automated software engineering (ase03).*

Fu, X., Bultan, T., & Su, J. (2004). Analysis of interacting bpel web services. In *Proc. of the 13th international world wide web conference (www'04).* ACM Press.

Grastien, A., Cordier, M., & Largouët, C. (2004). Extending decentralized discrete-event modelling to diagnose reconfigurable systems. In *Proceedings of the fifteenth international workshop on principles of diagnosis (DX-04)* (pp. 75–80). Carcassonne, France.

Grastien, A., Cordier, M., & Largouët, C. (2005). Incremental diagnosis of discrete-event systems. In *Proceedings of the sixteenth international workshop on principles of diagnosis (DX-05)* (pp. 119–124). Pacific Grove, California, USA.

Grosclaude, I. (2004). Model-based monitoring of software components. In *Proceedings of the 16th European conf. on artificial intelligence (ECAI'04)* (pp. 1025– 1026).

Hamscher, W., Console, L., & de Kleer, J. (Eds.). (1992). *Readings in model-based diagnosis.* Morgan Kaufmann.

Korel, B. & Laski, J. (1988). Dynamic program slicing. *Information Processing Letters*, 29(3), 155-163.

Lazovik, A., Aiello, M., & Papazoglou, M. (2003). Planning and monitoring the execution of web service requests. *In Proceedings of the 1st int. conf. on service-oriented computing (ICSOC'03)* (pp. 335–350).

Mahbub, K. & Spanoudakis, G. (2004). A framework for requirements monitoring of service based systems. In *Proceedings of the 2nd international conference on service oriented computing* (pp. 84-93).

Narayanan, S. & McIlraith, S. (2002). Simulation, verification and automated composition of web services. In *Proceedings of the eleventh international world wide web conference* (*www-11*) (pp. 77-88).

Ottenstein, K. & Ottenstein, L. (1984). The program dependence graph in software development environment. In *Acm Sigsoft/Sigplan software engineering symposium on practical software development environments* (pp. 177-184).

Ouyang, C., Aalst, W. Breutel, S., Dumas, M., Hofstede, A., & Verbeek, H. (2005). *Formal semantics and analysis of control flow in ws-bpel* (Tech. Rep.). BPM Center Report BPM-05-13. (BPMcenter.org)

Pencolé, Y. & Cordier, M. (2005). A formal framework for the decentralised diagnosis of large scale discrete event systems and its application to telecommunication networks. *Artificial Intelligence Journal*, *164*(1-2), pp. 121–170.

Pencolé, Y., Cordier, M., & Rozé, L. (2002). Incremental decentralized diagnosis approach for the supervision of a telecommunication network. In *Proceedings of 41th IEEE conf. on decision and control (CDC'2002)* (pp. 435–440). Las Vegas, USA.

Pistore, M., Traverso, P., Bertoli, P., & Marconi, A. (2005). Automated composition of web services by planning at the knowledge level. In *Proceedings of the 19th international joint conference on artificial intelligence (ijcai-05)* (p. 1252-1260).

Rao, J. & Su, X. (2004). A survey of automated web service composition methods. In *Proceedings of the first international workshop on semantic web services and web process composition* (*swswpc*).

Reiter, R. (1987). A theory of diagnosis from first principle. *Artificial Intelligence*, *32*(1), pp. 57-95.

Salaün, G., Bordeaux, L., & Schaerf, M. (2004). Describing and reasoning on web services using process algebra. In *Proceedings of the second IEEE int. conf. on web services (ICWS'04)* (pp. 43–51).

Sampath, M., Sengupta, R., Lafortune, S., Sinnamohideen, K., & Teneketzis, D. (1995). Diagnosability of discrete-event systems. *IEEE Transactions on Automatic Control*, *40*(9), pp. 1555-1575.

Schmidt, K. & Stahl, C. (2004). A Petri net semantic for bpel: validation and application. In *Proc. of 11th workshop on algorithms and tools for Petri nets (awpn 04)* (pp. 1-6).

Viroli, M. (2004). Towards a formal foundation to orchestration languages. In *Electronic notes in theoretical computer science* 105 (pp. 51-71). Elsevier.

Weise, M. (1984). Program slicing. *IEEE Transactions on Software Engineering*, *10*(4), pp. 352-357.

Wotawa, F. (2002). On the relationship between model-based debugging and program slicing. Artificial Intelligence, 135, pp. 125-143.

Yan, Y., Pencolé, Y., Cordier, M., & Grastien, A. (2005). Monitoring web service networks in a model-based approach. In *3rd IEEE European conference on web services (ecows05).* Växjö, Sweden: IEEE Computer Society.

ENDNOTES

1 In diagnosis concept, symptom is an observed abnormal behavior, while fault is the original cause of a symptom. For example, an alarm from a smoke detector is a symptom. The two possible faults, a fire or a faulty smoke detector, are the causes of the symptom.

2 $F(c_i)$ is a specific fault mode. When we do not know a specific fault mode, we use $ab(c_i)$ to represent c_i is faulty.

3 When a Web service engine supports multiple instances of a process, different instances are identified with a process ID. Therefore, diagnosis is based on the events for one instance of the process.

4 Sometimes, the exception returns the information about the part *A.InVar.part* is faulty. Then this rule is simplified.

This work was previously published in the International Journal of Web Services Research, Vol. 6, Issue 1, edited by L.J. Zhang, pp. 87-110, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 7.9
Management of Medical Website Quality Labels via Web Mining

Vangelis Karkaletsis
National Center of Scientific Research "Demokritos", Greece

Konstantinos Stamatakis
National Center of Scientific Research "Demokritos", Greece

Pythagoras Karampiperis
National Center of Scientific Research "Demokritos", Greece

Martin Labský
University of Economics, Prague, Czech Republic

Marek Růžička
University of Economics, Prague, Czech Republic

Vojtěch Svátek
University of Economics, Prague, Czech Republic

Enrique Amigó Cabrera
ETSI Informática, UNED, Spain

Matti Pöllä
Helsinki University of Technology, Finland

Miquel Angel Mayer
Medical Association of Barcelona (COMB), Spain

Dagmar Villarroel Gonzales
Agency for Quality in Medicine (AquMed), Germany

ABSTRACT

The World Wide Web is an important channel of information exchange in many domains, including the medical one. The ever increasing amount of freely available healthcare-related information generates, on the one hand, excellent conditions for self-education of patients as well as physicians, but on the other hand, entails substantial risks if such information is trusted irrespective of low

DOI: 10.4018/978-1-60566-218-3.ch010

competence or even bad intentions of its authors. This is why medical Web site certification, also called quality labeling, by renowned authorities is of high importance. In this respect, it recently became obvious that the labelling process could benefit from employment of Web mining and information extraction techniques, in combination with flexible methods of Web-based information management developed within the Semantic Web initiative. Achieving such synergy is the central issue in the MedIEQ project. The AQUA (Assisting Quality Assessment) system, developed within the MedIEQ project, aims to provide the infrastructure and the means to organize and support various aspects of the daily work of labelling experts.

INTRODUCTION

The number of health information websites and online services is increasing day by day. It is known that the quality of these websites is very variable and difficult to assess; we can find websites published by government institutions, consumer and scientific organizations, patients associations, personal sites, health provider institutions, commercial sites, etc. (Mayer et.al., 2005). On the other hand, patients continue to find new ways of reaching health information and more than four out of ten health information seekers say the material they find affects their decisions about their health (Eysenbach, 2000; Diaz et.al., 2002). However, it is difficult for health information consumers, such as the patients and the general public, to assess by themselves the quality of the information because they are not always familiar with the medical domains and vocabularies (Soualmia et.al., 2003).

Although there are divergent opinions about the need for certification of health websites and adoption by Internet users (HON, 2005), different organizations around the world are working on establishing standards of quality in the certification of health-related web content (Winker et.al., 2000; Kohler et.al., 2002; Curro et.al., 2004; Mayer et.al., 2005). The European Council supported an initiative within eEurope 2002 to develop a core set of "Quality Criteria for Health Related Websites" (EC, 2002). The specific aim was to specify a commonly agreed set of simple quality criteria on which Member States, as well as public and private bodies, may build upon for developing mechanisms to help improving the quality of the content provided by health-related websites. These criteria should be applied in addition to relevant Community law. As a result, a core set of quality criteria was established. These criteria may be used as a basis in the development of user guides, voluntary codes of conduct, trust marks, certification systems, or any other initiative adopted by relevant parties, at European, national, regional or organizational level.

This stress on content quality evaluation contrasts with the fact that most of the current Web is still based on HTML, which only specifies how to layout the content of a web page addressing human readers. HTML as such cannot be exploited efficiently by information retrieval techniques in order to provide visitors with additional information on the websites' content. This "current web" must evolve in the next years, from a repository of human-understandable information, to a global knowledge repository, where information should be machine-readable and processable, enabling the use of advanced knowledge management technologies (Eysenbach, 2003). This change is based on the exploitation of *semantic web* technologies. The Semantic Web is "an extension of the current web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation" based on metadata (i.e. semantic annotations of the web content) (Berners-Lee et.al., 2001). These metadata can be expressed in different ways using the Resource Description Framework (RDF) language. RDF is the key technology behind the Semantic Web, providing a means of expressing data on the web in a structured way that can be processed by machines.

In order for the medical quality labelling mechanisms to be successful, they must be equipped with semantic web technologies that enable the creation of machine-processable labels as well as the automation of the labelling process. Among the key ingredients for the latter are *web crawling* techniques that allow for retrieval of new unlabelled web resources, or *web spidering and extraction* techniques that facilitate the characterization of retrieved resources and the continuous monitoring of labeled resources alerting the labelling agency in case some changes occur against the labelling criteria.

The *AQUA* (Assisting QUality Assessment) system[1], developed within the MedIEQ project[2], aims to provide the infrastructure and the means to organize and support various aspects of the daily work of labelling experts by making them computer-assisted. AQUA consists of five major components (each, in turn, incorporating several specialized tools): Web Content Collection (WCC), Information Extraction (IET), Multilingual Resources Management (MRM), Label Management environment (LAM), and Monitor Update Alert (MUA). While WCC, IET and MUA together constitute the web data analysis engine of AQUA, MRM provides them access to language-dependent medical knowledge contained in terminological resources, and LAM handles the generation, storage and retrieval of resulting labels. The user interface of AQUA allows for both entirely manual labelling and labelling based on the results of automatic analysis. In this chapter we will describe the challenges addressed and results achieved by applying the WCC and IET tools to raw web data, as well as the subsequent processes of quality label handling by LAM.

Categories and Quality of Medical Web Content: a Survey

In order to investigate what types of Medical Web Content exist, at the beginning of the project we conducted a survey on a set of Greek health-related websites, classifying them into the following categories: “government organization”, “healthcare service provider”, “media and publishers”, “patient organization / self support group”, “pharmaceutical company / retailer”, “private individual” and “scientific or professional organization”. Apart from the categorization of these websites, we also collected additional information for them in order to construct a *medical web map*. The extra fields of information were the following: “last update”, “language(s)”, “title”, “location”, “description” and “keywords” of the website but also “trust marks: are they present or not”, “trustworthiness (a first estimation on the quality of the medical content: is it reliable?)”, “advertisements: are they present or not?”.

We first collected a few thousands of URLs with the assistance of a *search engine wrapper*. The wrapper queried the Google search engine with several sets of health related keywords, in both Greek and English languages, and collected the resulting websites. From the English keywords’ results we only kept those corresponding to websites originated from Greece. On the resulting Greek URLs’ list, an automated filtering procedure was applied, where duplicates, overlapping and other irrelevant URLs were removed. 1603 URLs remained. Checking manually the remaining URLs, 723 websites were selected for having health-related content. These were then categorized according to the categories mentioned above. The *crawling software*, developed for the purposes of the project, based on machine learning and heuristic methods, extracted the machine detectable information, which is “last update”, “language(s)”, “title”, “location”, “description” and “keywords”.

Table 1. Categorization of Medical Web Content under review

Categories	*URLs*	*Percentage (%)*
Government organizations	15	2%
Healthcare service providers	211	28%
Media and publishers	64	9%
Patient organizations/ self support groups	33	5%
Pharmaceutical companies/ retailers	51	7%
Private individuals	199	28%
Scientific or professional organizations	110	15%
Universities/ research institutions	40	6%
Total	**723**	**100%**

Apparently, the 723 sites examined do not cover the totality of the Greek medical web content. However, they comprise a fair sample of that, which allowed us to make some useful observations with regard to this content.

The majority of websites belong to the *healthcare service provider* category (211 URLs) and to the *private individual* category (199 URLs). This fact reveals that in Greek medical web, the private sector is dominant (which seems reasonable), while the websites coming from the public sector like government organizations and universities/ research institutions are a minority (54 URLs). Furthermore, it is remarkable that a great portion (110 URLs) of the Greek medical web belongs to scientific/professional organizations.

We also noticed that, at the time of the survey, only three websites had a *quality seal*, namely, HON Code (HON, 2001) and all of them belong to the *scientific or professional organization* category. We could argue that the non-conformance to trust mark quality criteria characterizes the Greek medical web as a whole, which demonstrates that Greek online medical content providers are not familiar with the quality labelling aspect. Thus, the quality of the content of Greek medical websites appears to be doubtful. To support this, note that the HTML tags for "description" and "keywords" (which the crawler reads automatically), were found as either empty or containing misleading information in most Greek medical pages, while, for example, a quick look into a portion of the German medical web showed the opposite. Concluding, only few Greek medical websites conform to the biggest part of the selected criteria as to be considered of good quality.

We also conducted analogous but less elaborate studies for other 'less-spoken' languages that are involved in the MedIEQ project but not covered by the partner labelling agencies, namely Czech and Finnish. The first observations of the Czech and Finnish medical web maps seem to confirm the hypotheses formed based on the analysis of Greek websites detailed above.

Thus, the establishment of mechanisms/infrastructures for the quality certification of health related websites is quite critical. Its positive role would amount to forcing health content providers to the following directions:

- For *existing* online medical content: conform to generally accepted quality criteria defined by experts. For online medical content *planned* to be published: adapt to specific standards (presence of detailed information on the content provider, authorship information, last update, contact data, etc.).
- *High-quality* websites, already trusted by health information consumers, would

clearly boost the opinion that the web is not an advertising-oriented or dangerous space, but a powerful source of information and must be considered as such. In the same direction, the national medical sector could be motivated to develop web resources of quality, extending the usefulness of the medium and eventually attracting a larger amount of users.

The MedIEQ project aims to directly contribute to this direction.

STATE OF THE ART IN HEALTH WEB QUALITY LABELLING

Two major approaches currently exist concerning the labelling of health information in the internet: a) *filtering portals* (organizing resources in health topics and providing opinions from specialists on their content) and b) *third-party certification* (issuing certification trustmarks or seals once the content conforms to certain principles). In general, and in both approaches, the labelling process comprises three tasks that are carried out entirely or partially by most labelling agencies:

- *Identification* of new web resources: this could happen either by active web searching or on the request of the information provider, i.e. the website responsible actively asks for the review in order to get a certification seal.
- *Labelling* of the web resources: this could be done with the purpose of awarding a certification seal or in order to classify and index the web resources in a filtering portal.
- *Re-reviewing* or *monitoring* the labeled web resources: this step is necessary to identify changes or updates in the resources as well as broken links, and to verify if a resource still deserves to be awarded the certification seal.

This is the general case; eventually, any particular agency can integrate additional steps which may be necessary in its work. The two labelling agencies participating in MedIEQ, Agency for Quality in Medicine – AQuMed (http://www.aezq.de) and Web Mèdica Acreditada - WMA (http://wma.comb.es), represent the two approaches mentioned above: AQuMed maintains a filtering portal while WMA acts as a third-party certification agency.The indexing and labelling process in AQuMed consists of five steps:

1. *Inclusion of a new resource.* There are two ways through which a new resource can be identified for indexing in AQuMed database. The first one is through internet search and the second one is through a direct request from the information provider. The websites are selected according to general criteria: content, form and presentation should be serious, authorship, sponsorship and creation/update date should be clear, and only websites without commercial interest should be indexed.
2. *Website classification.* Previously unlabelled websites are classified into four groups: treatment information, background information, medical associations/scientific organizations and self-help/counseling organizations. Only the sites with treatment information proceed to the next step.
3. *Evaluation.* Sites with treatment information are evaluated using the DISCERN (DISCERN, 2008) and Check-In (Sanger, 2004) instruments. DISCERN is a well-known user guidance instrument, and Check-In was developed by AQuMed in collaboration with the "Patient Forum" of the German Medical Association. Check-In is based on DISCERN and the AGREE (AGREE, 2004)

instrument for critical evaluation of medical guidelines.

4. *Confirmation.* The database administrator has to confirm the result of the evaluation. It can be modified, erased, or simply confirmed.
5. *Feedback to the information provider.* AQuMed sends an e-mail with the result of the evaluation in the case of sites with treatment information and with the information about the admission into the AQuMed database in the case of other categories.

AQuMed's database is periodically populated through new internet searches and is regularly examined for broken links. The evaluated web resources are also periodically re-reviewed in order to identify changes against the criteria or other updates.Similarly, the complete certification process in WMA consists of the following four steps:

1. The person in charge of a website sends a (voluntary) request to WMA in order to initiate the process. Using the online application form, the interested party provides certain information to WMA and has the chance to auto-check the WMA criteria based on the Code of Conduct and the Ethical Code;
2. The WMA Standing Committee assesses the website based on the WMA criteria (medical authorship, updating, web accessibility, rules in virtual consultation, etc.), and issues recommendations;
3. WMA sends a report to the person in charge who implements the recommendations;
4. When the recommendations have been implemented, it is possible to obtain the seal of approval. In such a case, WMA sends an HTML seal code to be posted on the accredited website. In addition, WMA includes the site's name and URL to the index of accredited websites and an RDF file is generated.

EXPERIMENTAL COLLECTION OF LABELLING CRITERIA

In the MedIEQ project we decided to develop a representative collection of *labelling criteria*, which would reflect the needs of the *labelling agencies* involved in the project consortium and at the same time provide an adequate proof of concept for our general methodology for computer-assisted labelling. It is important to stress that the methodology and software tools are to a large degree independent of the concrete criteria and thus could be easily adapted to different criteria used by various agencies. Such adaptation is also eased by the fact that the criteria specification was also influenced by the analysis of criteria used by other organizations such as HON, and thus has significant overlap with them.

The set of labelling criteria used in MedIEQ (36 in total, organized in 10 different categories) is shown in Table 2. For each of these criteria, the AQUA system aims to identify and extract relevant information to be proposed to the expert (i.e. automatically provide information otherwise searched for manually within the site). The expert can accept or modify AQUA's suggestions and generate a quality label on the fly.

THE AQUA SYSTEM OVERVIEW

Development Objectives

Taking into account WMA and AQuMed approaches, the AQUA tool (Stamatakis et. al., 2007) was designed to support the main tasks in their labelling processes, more specifically:

1. Identification of unlabelled resources having health-related content
2. Visit and review of the identified resources
3. Generation of content labels for the reviewed resources

Table 2. The set of criteria examined in MedIEQ

ID	Criterion Name	Description
1. Resource Defining Information		
1.1	Resource URI	Includes information identifying/describing the resource. Concerning the resource URI: a) whether the resource's URI is valid or not and b) in case it redirects to external domains, are these domains between those specified when the resource was added? The rest is information like the resource's last update, its title and the language(s) in which content is provided.
1.2	Resource title	
1.3	Resource last update	
1.4	Resource language(s)	
2. Ownership / Creatorship		
2.1	Organization name(s) (owner)	The user should know who is behind the resource in order to judge by himself the credibility of the provided information. Therefore, information like the name(s) of the organization(s) providing the information and the type of this(these) organization(s) should be available. At the same time, the name(s), title(s) (e.g. MD, PhD, Dr, etc.) and contact details of website responsible(s), to contact in case of questions on health related issues, as well as the name(s) and contact details of the webmaster(s) should be available.
2.2	Organization type(s) (owner)	
2.3	Responsible name(s)	
2.4	Responsible title(s)	
2.5	Responsible(s) contact details	
2.6	Webmaster name(s)	
2.7	Webmaster(s) contact details	
3. Purpose / mission		
3.1	Purpose / mission of the resource provided	It has to be clear for the user which is the goal and motivation of the provided information and for what kind of users it was created e.g. adults, children, people with diabetes, etc.
3.2	Purpose / mission of the owner(s) provided	
3.3	Target / intended audience(s)	
3.4	Statement declaring limitation of the provided information	
4. Topics / Keywords		
4.1	Topics / Keywords (UMLS)	Mapping of the resource's content to concepts from the UMLS Metathesaurus.
5. Virtual consultation		
5.1	VC service available	A virtual consultation (VC) service is an online service allowing the user to ask questions and/or send/upload information on health related issues asking for advice. The name(s) and details of the person(s) responsible(s) for this service should also be clearly mentioned. Moreover, a declaration that VC is only a supporting means that cannot replace a personal consultation with a physician should be provided.
5.2	VC responsible name(s)	
5.3	VC responsible(s) contact details	
5.4	Statement declaring limitation of the VC service	
6. Funding / Advertising		

Table 2. continued

ID	*Criterion Name*	*Description*
6.1	Statement declaring sources of funding (sponsors, advertisers, etc.)	Health web resources should disclose possible conflicts of interest. For this reason it is important to know how and by whom a web resource is funded. If there are any sponsors, it has to be clear who they are. Furthermore, it should be stated that sponsors do not have any influence on the content. Additionally, it has to be known whether the web resource hosts or not advertising material in whatever format. In case that happens, such material should be clearly distinguished from informative material. Furthermore, information on resource's policy with regard to advertising must be easily accessible and clear.
6.2	Name(s) of funding (sponsoring) organization(s)	
6.3	Statement declaring limitation of influence of sponsors on content	
6.4	Advertising present	
6.5	Are advertisements clearly separated from editorial content?	
6.6	Policy with regard to advertisement	
7. Other Seal or Recommendation		
7.1	Other seal(s) present	Are there other seals identified in the resource? Indicates that the resource already conforms to other, known quality criteria. Identifiers for other seals: a) Real seals: WMA, HONcode, pWMC, URAC, eHealth TRUST-E, AFGIS, b) Filtering health portals (a resource is recommended by): AQUMED, Intute, WHO ("Vaccine Safety Net")
7.2	Which other seal(s)?	
8. Information Supporting Scientific Content		
8.1	References, bibliography (with links to literature)	Regarding the provided specialized health information (scientific parts of the resource) it is relevant to know if it is based on scientific books, medical journal articles, etc. For this, scientific articles or documents should include a references or bibliography section. Additionally, it is important to know if such information is up-to-date (publication and last modification dates are required) and who is the author of such content (author(s) name(s) and contact details are required for pages/documents providing scientific information).
8.2	Publication / creation date	
8.3	Last revision / modification date	
8.4	Author name(s)	
8.5	Author(s) contact details	
8.6	Editorial policy	
9. Confidentiality / privacy policy		
9.1	Explanation on how personal data (visitor coordinates, e-mail messages, etc.) is handled	Internet users are much concerned about protection of their privacy and personal data. For this reason the resource should provide a confidentiality/privacy policy ensuring that personal data (visitor coordinates, e-mail messages, etc.) is safely handled, describing how these data are handled.
10. Accessibility		
10.1	Accessibility level	The resource is examined upon various accessibility criteria and information on its accessibility level (whether the resource is of level A, AA or AAA) is deduced.

4. Monitoring the labeled resources

Compared to other approaches that partially address the assessment process (Griffiths et. al., 2005; Wang & Liu, 2006), the AQUA system is an integrated solution. AQUA aims to provide the infrastructure and the means to organize and support various aspects of the daily work of labelling experts by making them computer-assisted. The steps towards this objective are the following:

- **Step 1:** Creating machine readable labels by:
  - Adopting the use of the RDF model (W3C, 2004) for producing machine-readable content labels; at the current stage, the RDF-CL model (W3C, 2005) is used. In the final version of AQUA, another model called POWDER, introduced by the recently initiated W3C Protocol for Web Description Resources (POWDER) working group (W3C, 2007), will be supported.
  - Creating a vocabulary of criteria, consolidating on existing ones from various Labelling Agencies; this vocabulary is used in the machine readable RDF labels.
  - Developing a label management environment allowing experts to generate, update and compare content labels.
- **Step 2:** Automating parts of the labelling process by:
  - Helping in the identification of unlabelled resources.
  - Extracting from these resources information relative to specific criteria.
  - Generating content labels from the extracted information.
  - Facilitating the monitoring of already labeled resources.
- **Step 3:** Putting everything together; AQUA is implemented as a large-scale, enterprise-level, web application having the following three tiers:
  - The user tier, including the user interfaces for the labelling expert and the system administrator.
  - The application tier where all applications run.
  - The storage tier consisting of the MedIEQ file repository and the MedIEQ database.

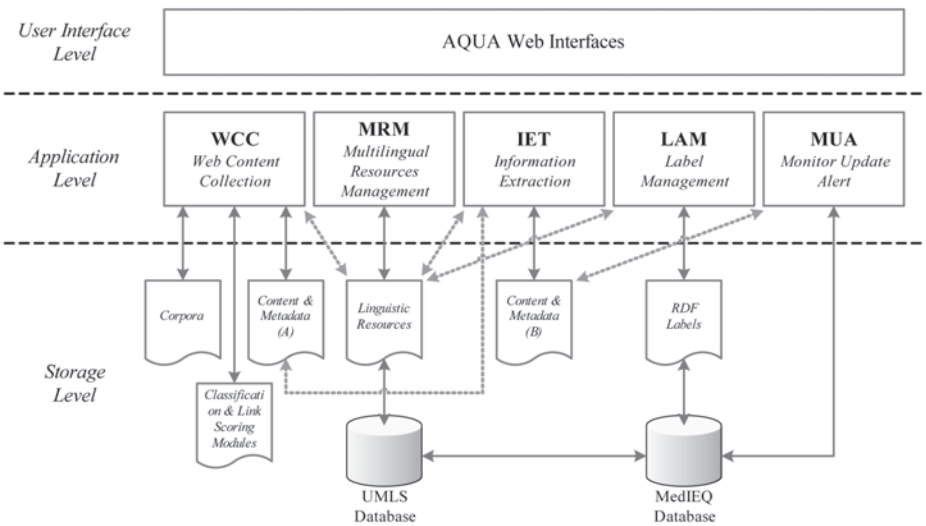
System Architecture

AQUA addresses a complex task. However, various design and implementation decisions helped MedIEQ partners keep AQUA extensible and easy to maintain. The main characteristics of its implementation include:

a. Open architecture
b. Accepted standards adopted in its design and deployment
c. Character of large-scale, enterprise-level web application
d. Internationalization support

AQUA incorporates several subsystems (see the application level in Figure 1) and functionalities for the labelling expert. The *Web Content Collection* (WCC) component identifies, classifies and collects online content relative to the criteria proposed by the labelling agencies participating in the project. The *Information Extraction Toolkit* (IET) analyses the web content collected by WCC and extracts attributes for MedIEQ-compatible content labels. The *Label Management* (LAM) component generates, validates, modifies and compares the content labels based on the schema proposed by MedIEQ. The *Multilingual Resources Management* (MRM) gives access to health-related multilingual resources; input from such resources is needed in specific parts of the WCC, IET and LAM toolkits. Finally, *Monitor-Update-Alert* (MUA) handles auxiliary but important jobs

Figure 1. Architecture of the AQUA system



like the configuration of monitoring tasks, the MedIEQ database updates, or the alerts to labelling experts when important differences occur during the monitoring of existing content labels.

While the first prototype, made operational in autumn 2007, only addresses the certification of new resources and covers two languages (English and Spanish), the full version of the system will also enable monitoring of already labeled resources and will cover 7 languages in total.

Figure 1 shows all the possible data flows in AQUA (dashed arrows): a) From WCC to IET: pages collected by WCC, once undergone a first-level extraction by WCC (extraction of metadata 1), are then forwarded to IET for further processing (extraction of metadata 2); b) From IET to MUA: MUA takes all metadata collected by both WCC and IET and updates the MedIEQ database; c) From MRM to WCC, IET, LAM: custom vocabularies, generated by the MedIEQ users through MRM interface, can be accessed from other toolkits (WCC, IET, LAM), where the user may need them.

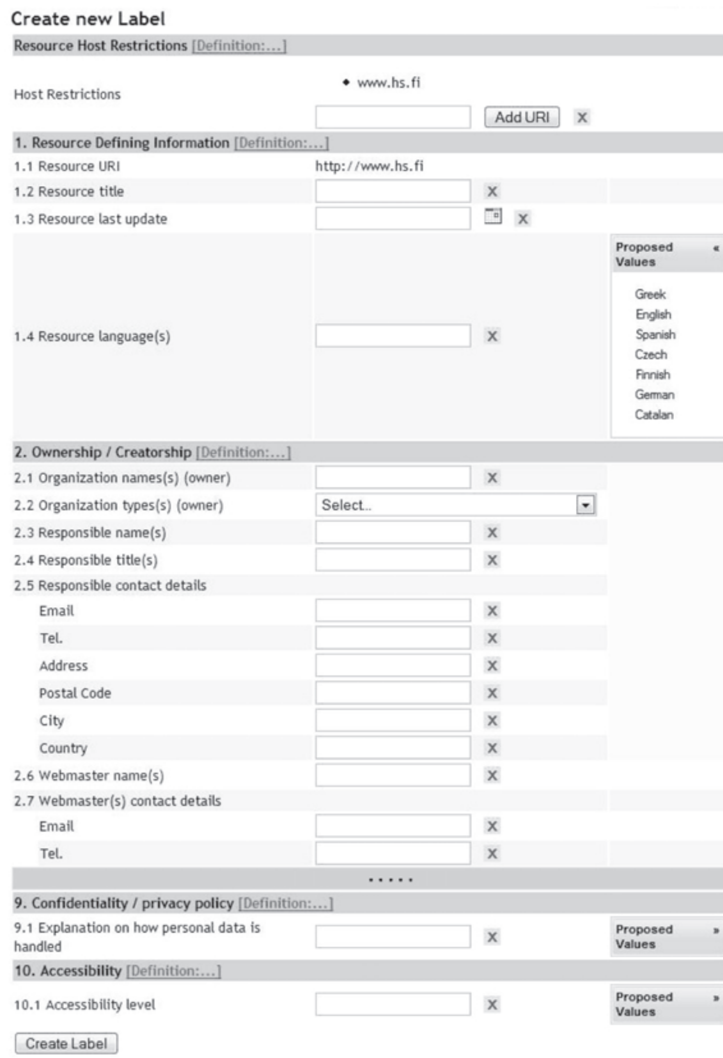
The following two sections are devoted to a more detailed description of AQUA, namely of its (manual) label management components and of its automated labelling support components.

AQUA LAM COMPONENT: CREATING MACHINE-READABLE LABELS

Representation Formalism for Machine-Readable Labels

To make content labels machine-readable the use of the RDF model is adopted. At the current stage, the RDF-CL model is used. The RDF-CL model was issued by the EC-funded project Quality Assistance and Content Description (QUATRO) (www.quatro-project.org); it is currently being refined by the W3C Protocol for Web Description Resources (POWDER) working group (W3C, 2007). POWDER is expected to be completed before the end of the

Figure 2. The AQUA label management environment (LAM) interface



MedIEQ project and the plan is to use it in the final version of AQUA.

User Interaction with the Label Management Environment

The *label management* interface and associated tools, together called LAM, allows experts to generate, update and compare content labels. From within the LAM user interface the user is able to a) generate new RDF labels from information automatically extracted by other AQUA tools, b) manually fill the relevant fields and generate new RDF labels, c) edit and update existing RDF labels, and d) compare RDF labels among themselves.

The user interface to generate/edit a label is a *web form* (see Figure 2) with input boxes, single and multiple select boxes, links and buttons. It is split into two distinct areas. The first part lets the user *constrain* the application of the label to certain hosts by explicitly declaring the host URIs or by adding regular expressions that properly identify them. Multiple hosts can be defined. Regular expressions for more fine-grained addressing can be defined as well. These definitions can be combined via the union and intersection operators and thus create rules that link different parts of a web resource with different labels.

The second part is where the *label properties* are assigned *values*. The label properties are the actual descriptors of a web resource, mapping the labelling criteria. A set of label descriptors can be linked with a set of host restrictions defined in the first part. Related properties are grouped to make the user filling them easier.

Once the user has filled the label metadata, restrictions and properties, s/he can *save* the label. There is a notification field that alerts the user if the label already exists in the system, and its changes are tracked by the AQUA version control system. In this case the user can save the label as a *revision* of an existing label. If the label is new, the user just selects to save the label. In both cases the user has the option to download an RDF/XML serialized form of the label. This serialized label can be assigned to the web resource by the site webmaster.

AQUA WCC+IET: AUTOMATING PARTS OF THE LABELLING PROCESS

Locating Unlabeled Web Resources

The AQUA *crawling* mechanism is part of the *web content collection* environment (WCC) (Stamatakis et. al., 2007). Its AQUA interface is shown in Figure 3. The Crawler searches the Web for health-related content that does not have a content label yet (at least not a label found in MedIEQ

Figure 3. Configuring the MedIEQ Crawler from the AQUA interface

records). It is a meta-search engine that exploits results returned from known search engines and directory listings from known Web directories. All collected URLs from all sources are merged and filtered, and a pre-final URLs list is returned. The merging / filtering process: a) removes possible duplicates, b) ignores sub-paths of URLs already in list, and c) removes URLs already having a content label (the Crawler consults the MedIEQ database for this).

The crawling process becomes even more focused with the aid of a *content classifier*, inductively trained to distinguish health content from non-health content. This classification component visits every URL from the merged / filtered pre-final URL list and checks its contents, thus filtering out some more entries.

The current version of the AQUA Crawler queries Google and Yahoo! *search engines* (with terms proposed by the user) and explores Web directories (again proposed by the user). By merely using general-purpose search engines, the Crawler inevitably inherits their shortcomings. Therefore, aiming to further enhance our Crawler, we also include two search mechanisms specialized to the *health domain*: one provided by HON (www.hon.ch) and another by Intute's Health and Life Sciences branch (www.intute.ac.uk). The Crawler interface is shown in Figure 3.

Browsing Medical Knowledge Sources

One of the main requirements when working with medical web resources, is to identify and classify them based on standardized medical terms. Such terms (knowledge sources) have been globally defined by the Unified Medical Language System (UMLS) (www.nlm.nih.gov/ research/ umls/). UMLS provides a wide set of linguistic health resources, well maintained and up-to-date, containing health concepts and relations between concepts and between resources.AQUA incorporates a module called Multilingual Resources Management Toolkit (MRM) that aims to support labelling experts in:

- Easily accessing and browsing selected "knowledge sources" form the variety that UMLS provides
- Creating new, custom resources, to better support the labelling process

MRM is an environment from which linguistic resources, either UMLS-supported or not (custom or user generated) in different languages can be managed. MRM provides a user-friendly environment for accessing and managing both UMLS "knowledge sources" and custom resources (see Figure 4).

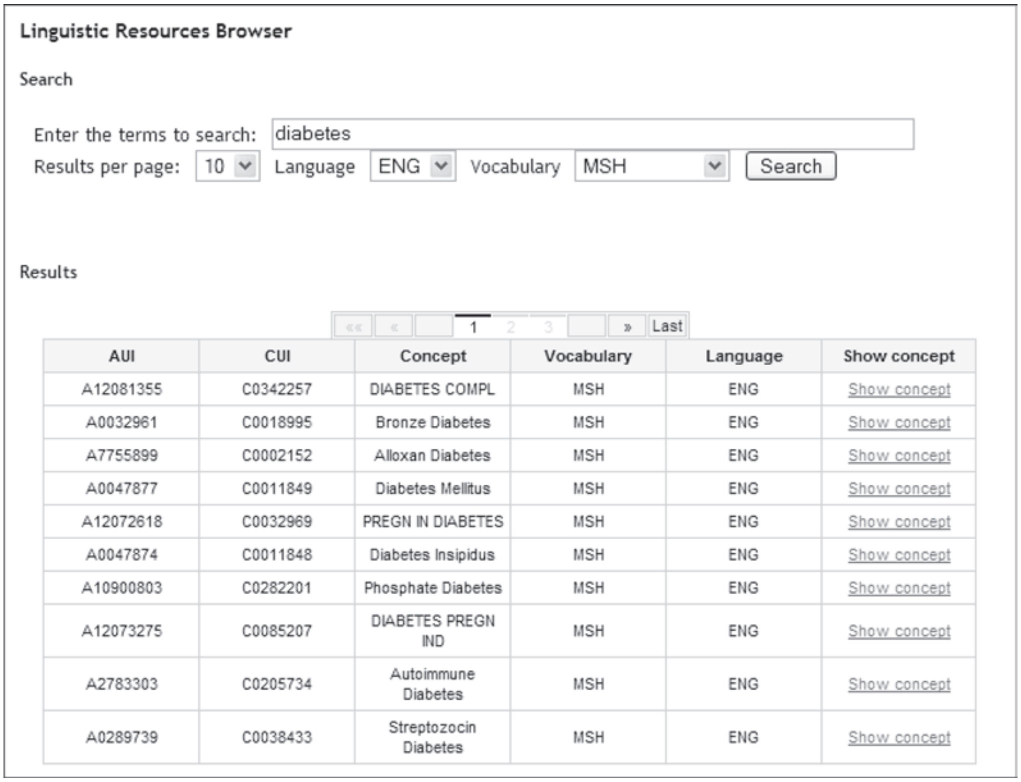
Spidering the Website

While the Crawler proceeds from the initial user's content collection requirement to the identification of a relevant website as a whole, the *Spider*, in turn, examines individual pages of the site. The sites whose URLs are obtained from the Crawler are processed by the Spider one-by-one in several independent threads. Unreachable sites/pages are revisited in next run.

Since not all the pages of a web site are interesting for the labelling process, the Spider utilizes a content classification component that consists of a number of *classification modules* (statistical and heuristic ones). These modules decide which pages contain interesting information. Each of them relies on a different classification method according to the classification problem on which it is applied. Pages identified as belonging to classes relevant to the labelling criteria are stored locally in order to be exploited by the Information Extraction subsystem.

One of the main classification modules of the Spider is the "UMLS/MeSH categoriser", called POKA. POKA (http://www.seco.tkk.fi/tools/poka/) is a tool for automatic extraction of ontological resources (RDF, OWL, SKOS) from text

Figure 4. Browsing medical knowledge sources with AQUA



Linguistic Resources Browser

Search

Enter the terms to search: diabetes

Results per page: 10 Language ENG Vocabulary MSH Search

Results

«« « 1 2 3 » Last

AUI	CUI	Concept	Vocabulary	Language	Show concept
A12081355	C0342257	DIABETES COMPL	MSH	ENG	Show concept
A0032961	C0018995	Bronze Diabetes	MSH	ENG	Show concept
A7755899	C0002152	Alloxan Diabetes	MSH	ENG	Show concept
A0047877	C0011849	Diabetes Mellitus	MSH	ENG	Show concept
A12072618	C0032969	PREGN IN DIABETES	MSH	ENG	Show concept
A0047874	C0011848	Diabetes Insipidus	MSH	ENG	Show concept
A10900803	C0282201	Phosphate Diabetes	MSH	ENG	Show concept
A12073275	C0085207	DIABETES PREGN IND	MSH	ENG	Show concept
A2783303	C0205734	Autoimmune Diabetes	MSH	ENG	Show concept
A0289739	C0038433	Streptozocin Diabetes	MSH	ENG	Show concept

documents. In the MedIEQ framework, POKA is used to find relations between medical web content and medical vocabularies such as MeSH to facilitate categorization of web resources. The POKA system is used as a component of the web spidering tool where the spider traverses health web sites by gathering internal links and visiting the corresponding web pages one by one. POKA is then harnessed to find medical terminology inside these pages by matching content with the MeSH vocabulary.

Extracting Information Relative to Criteria

MedIEQ continues and builds upon the work of previous projects in the area of *information extraction* (IE) (Karkaletsis et.al. 2004; Rainbow, 2005; Labsky & Svatek, 2006). The AQUA IE toolkit (IET) employs a set of components responsible for the extraction of elementary information items found in each document and for the integration of these items into a set of semantically meaningful objects called *instances*. An instance (of certain general class) can be for example the set of contact information about a health provider or the set of bibliographic information about a scholarly resource referred to on the website.

The core IE engine currently used within IET is the *Ex* system (Labsky et al., 2007), which relies on combination of so-called *extraction ontologies* with exploiting local *HTML formatting regularities* and the option of embedding *trainable classifiers* to perform selected extraction subtasks. IET is built as a generic information extraction toolkit that supports changes and additions to the utilized labelling schemes. In this way, IET can also be used for IE using third-party labelling schemes and within other domains.

Monitoring of Already Described Resources

Another part of AQUA, called MUA (from Monitor-Update-Alert), handles problems such as the *configuration* of *monitoring tasks*, the necessary MedIEQ *repository updates* and the *alerts* to labelling experts when important differences (relative to the quality criteria) occur during the monitoring of previously labeled sites. MUA thus extends the functionality of the content collection and extraction toolkits by shifting from a one-shot scenario to that of continuous monitoring.

MUA is currently in its design phase. Fully functional implementation is envisaged in the late phase of the MedIEQ project (mid-2008).

PRELIMINARY EVALUATION OF AQUA

Locating Unlabeled Web Resources

In this section, we summarize evaluation results on Crawler's content classification component. For this evaluation, we used an English corpus, consisting of 1976 pages (944 positive & 1032 negative samples), all manually annotated. Three different classifiers have been tested (SMO, Naïve Bayes and Flexible Bayes). All 1-grams, 2-grams and 3-grams were produced and the best of them according to information gain were selected (see Table 3). Best performance was achieved with 1-grams and HTML tags removed.

The relatively low performance of the content classifiers is justified by the fact that it is difficult, even for humans, in various cases to assess whether a website has health-related content or not.

Spidering the Website

The classification mechanism our Spider exploits has been examined using statistical classification techniques, for the criteria listed in Table 4. In addition, for the last criterion, a method based on heuristic detection was examined.

Several learning schemes, decision trees, naive Bayes and supported vector machines (SMO) were tested. The performance of the SMO classifier, which provides the best results, is presented in Table 5. As expected, the most difficult criterion for classification purposes is the target audience, being a highly subjective one.

Extracting Information Relative to Criteria

Table 6 shows preliminary results for extraction of *contact information*. Data sets were collected through website crawling and spidering, contact pages were identified and manually annotated for English (109 HTML pages), Spanish (200) and Czech (108). The collections contained roughly 7000, 5000 and 11000 named entities, respectively. The *contact extraction* ontologies (one per language with shared common parts) were developed based on seeing 30 randomly chosen documents from each dataset and evaluated using

Table 3. Classification performance results for content classification

	1-grams (Tags removed)		
	Prec.	**Rec.**	**Fm.**
NB	0.75	0.63	**0.68**
FB	0.73	0.55	**0.62**
SMO	0.75	0.61	**0.67**

Table 4. The MedIEQ criteria upon which our classification components were evaluated

Criterion	*MedIEQ approach*
The target audience of a website	Classification among three possible target groups: adults, children and professionals
Contact information of the responsible of a website must be present and clearly stated	Detection of candidate pages during the spidering process and forwarding for information extraction
Presence of virtual consultation services	Detection of parts of a website that offer such services during the spidering process
Presence of advertisements in a website	Detection of parts of a website that contain advertisements during the spidering process

Table 5. SMO performance

	English			**Spanish**		
Category	**Precision**	**Recall**	**Fm**	**Precision**	**Recall**	**Fm**
Contact Info	0.84	0.96	0.90	0.80	0.65	0.72
Advertisements	0.87	0.80	0.83	0.77	0.72	0.75
Virtual Consultation	0.87	0.87	0.87	0.75	0.58	0.65
Adults	0.78	0.75	0.77	0.65	0.64	0.65
Children	0.80	0.78	0.79	-	-	-
Professional	0.77	0.81	0.79	0.62	0.63	0.62

the remaining documents. Extraction ontologies utilize nested regular patterns at word, character and HTML tag level. They also refer to gazetteers such as lists of city names, common first names and surnames. Each ontology contained about 100 textual patterns for the context and content of attributes and also for the single extracted ’contact’ class, attribute length and data type constraints and several axioms. For the results below we did not exploit trainable classifiers; their meaningful combination with the manually authored extraction knowledge is still work-in-progress, and when applied standalone, their results were so far slightly inferior to those achieved via extraction ontologies. We attribute this observation to small amount and large heterogeneity of training data.

The effort spent on developing and tuning the ontologies was about 2-3 person-weeks for the initial, English ontology, and 1-2 person weeks for its customization to Spanish and Czech. In the strict mode of evaluation, only exact matches are considered to be successfully extracted. In the loose mode, partial credit is also given to incomplete or overflown matches; e.g. extracting ’John Newman’ where ’John Newman Jr.’ was supposed to be extracted will count as a 66% match (based on overlapping word counts). Table 6 shows results in ‘strict/loose’ order. Some of the performance numbers below may be impacted by a relatively low inter-annotator agreement (English and Spanish datasets are still being cleaned to remove inconsistencies).

AQUA Usability Evaluation

The 1st AQUA prototype was also evaluated by the labelling organizations participating in the MEDIEQ project (namely, WMA and AQUMED). The primary goal of this evaluation was to conclude with a functional prototype that has the potential to be fully integrated within the day-to-day activities of a labelling organization. To this end, a

Table 6. Results of IET for contact information[3]

	English			Spanish			Czech		
Attribute	**Precision**	**Recall**	**Fm**	**Precision**	**Recall**	**Fm**	**Precision**	**Recall**	**Fm**
Degree/Title	71/78	82/86	76/82	-	-	-	87/89	88/91	88/90
Name	66/74	51/56	58/64	71/77	81/86	76/81	74/76	82/83	78/80
Street	62/85	52/67	56/75	71/93	46/58	56/71	78/83	66/69	71/75
City	47/48	73/76	57/59	48/50	77/80	59/61	67/75	69/79	68/77
Zip	59/67	78/85	67/75	88/91	91/94	89/93	91/91	97/97	94/94
Country	58/59	89/89	70/71	67/67	78/78	72/72	64/66	87/96	74/78
Phone	97/99	84/87	90/93	84/89	91/96	87/92	92/93	85/85	88/89
Email	100/100	99/99	100/100	94/95	99/99	96/97	99/99	98/98	98/98
Company	57/81	37/51	44/63	-	-	-	-	-	-
Department	51/85	31/45	38/59	-	-	-	-	-	-
Overall	70/78	62/68	66/72	71/76	81/86	76/80	81/84	84/87	82/84

parallel technical improvement action took place, refining given functionalities. The main objective of the extra technical improvement action was to enhance the overall system workflow, so as to better match the day-to-day practice. The specifications for these technical refinements were given by an iterative feedback process with the MedIEQ labelling organizations, during the evaluation. It must be noted that the current interim version of AQUA was well received by both labelling organizations participating in the Usability Evaluation testing, and that they expressed their confidence that AQUA will be fully integrated within their day-to-day labelling activities.

CONCLUDING REMARKS

Other attempts to automatically assess health information in the internet exist but address the assessment process only partially. The Automated Quality Assessment procedure (AQA) (Griffiths et. al., 2005) ranks depression websites merely according to their evidence-based quality. The Automatic Indicator Detection Tool (AIDT), presented in a recent study (Wang & Liu, 2006), is suggested as a complementary instrument for the assessment of health information quality. AIDT is evaluated upon the automatic detection of pre-defined indicators that correspond to a number of technical quality criteria. However, AIDT focuses on a narrow scope of extraction techniques only, and does not address the assessment process as a whole. In contrast, the AQUA approach seems to be unique in covering the whole workflow of labelling agencies and employing a comprehensive and flexible collection of automated tools.

Assessing the quality of health-related information published on the internet is a task with great importance for the quality of the healthcare itself, due to a large proportion of patients as well as medical practitioners nowadays using the internet as a high-coverage information resource. It is at the same time a complex task as it has to examine the conjunction of a number of different aspects. Various initiatives around the world have attempted to codify these aspects into criteria, principles, codes of conduct, etc. Health specialists review online health resources and label them, either by issuing certification trustmarks or by including them in a thematic health portal. However this work can be proven quite tedious even for experienced users. Additionally, as it currently relies on manual effort, the labelling process is very time-consuming. Instruments to assist certain parts of the work exist; they however focus on specific problems only and none of them addresses the assessment process as a whole. In this context, efforts such as the MedIEQ project will bring wide reusability to content labels in the health domain by giving them machine-readable semantics and by providing services, such as those of the AQUA system, for creating and exploiting these machine-readable labels.

From the knowledge technology research viewpoint, the added value of MedIEQ is in employing existing technologies in a novel application: the automation of the labelling process in health-related web content. These technologies are *semantic web* technologies for describing web resources and *web search* (crawling and spidering) *and mining* (classification and information extraction) technologies for collecting domain-specific web content and extracting information from it. Experimental results for the mining components, investigating the performance of different inductive-learning-based as well as knowledge-engineering-based methods, are promising.

REFERENCES

W3C. (2004). *Resource Description Framework (RDF)*. Retrieved from http://www.w3.org/TR/rdf-schema/

W3C. (2005). *RDF-Content Labels (RDF-CL)*. Retrieved from http://www.w3.org/ 2004/12/q/doc/content-labels-schema.htm

W3C. (2007). *Protocol for Web Description Resources (POWDER)*. Retrieved from http://www.w3.org/2007/powder/

AGREE. Appraisal of Guidelines Research and Evaluation (AGREE), 2004. Available Online at: http://www.agreecollaboration.org/instrument/

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, (May): 2001.

Curro, V., Buonuomo, P. S., Onesimo, R., Vituzzi, A., di Tanna, G. L., & D'Atri, A. (2004). A quality evaluation methodology of health web-pages for non-professionals. *Medical Informatics and the Internet in Medicine*, *29*(2), 95–107. doi:10.1080/14639230410001684396

Diaz, J. A., Griffith, R. A., Ng, J. J., Reinert, S. E., Friedmann, P. D., & Moulton, A. W. (2002). Patients'use of the Internet for medical Information. *Journal of General Internal Medicine*, *17*(3), 180–185. doi:10.1046/j.1525-1497.2002.10603.x

DISCERN. (2008). *DISCERN: Quality criteria for consumer health information*. Retrieved from http://www.discern.org.uk/.

European Commission. (2002). *eEurope 2002: Quality Criteria for Health related Websites*. Retrieved from http://europa.eu.int/information_society/eeurope/ehealth/ doc/communication_acte_en_fin.pdf.

Eysenbach, G. (2000). Consumer health informatics. *BMJ (Clinical Research Ed.)*, *320*(4), 1713–1716. doi:10.1136/bmj.320.7251.1713

Eysenbach, G. (2003). The Semantic Web and healthcare consumers: A new challenge and opportunity on the horizon? *J Health Techn Manag*, *5*, 194–212. doi:10.1504/IJHTM.2003.004165

Griffiths, K. M., Tang, T. T., Hawking, D., & Christensen, H. (2005). Automated assessment of the quality of depression Web sites. *Journal of Medical Internet Research*, *7*(5), e59. doi:10.2196/jmir.7.5.e59

HON, Health on the Net Foundation. (2001). *HONCode*. Retrieved from http://www.hon.ch

HON, Health on the Net Foundation. (2005). *Analysis of 9th HON Survey of Health and Medical Internet Users Winter 2004-2005*. Retrieved from http://www.hon.ch/Survey/Survey2005/res.html

Karkaletsis, V., Spyropoulos, C. D., Grover, C., Pazienza, M. T., Coch, J., & Souflis, D. (2004) A Platform for Crosslingual, Domain and User Adaptive Web Information Extraction. In *Proceedings of the European Conference in Artificial Intelligence* (ECAI), Valencia, Spain; p. 725-9.

Kohler, C., Darmoni, S. D., Mayer, M. A., Roth-Berghofer, T., Fiene, M., & Eysenbach, G. (2002). MedCIRCLE – The Collaboration for Internet Rating, Certification, Labelling, and Evaluation of Health Information. Technology and Health Care, Special Issue: Quality e-Health. *Technology and Health Care*, *10*(6), 515.

Labsky, M., & Svatek, V. (2006, June) Information Extraction with Presentation Ontologies. In: ESWC'06 *Workhshop on Mastering the Gap: From Information Extraction to Semantic Representation*, Budva, Montenegro.

Labsky, M., Svatek, V., Nekvasil, M., & Rak, D. (2007). The Ex Project: Web Information Extraction using Extraction Ontologies. In: *Proc. PriCKL'07, ECML/PKDD Workshop on Prior Conceptual Knowledge in Machine Learning and Knowledge Discovery*. Warsaw, Poland, October 2007.

Mayer, M. A., Leis, A., Sarrias, R., & Ruíz, P. (2005). Web Mèdica Acreditada Guidelines: realiability and quality of health information on Spanish-Language websites. In Engelbrecht R et al. (ed.). *Connecting Medical Informatics and Bioinformatics*. Proc of MIE2005, 1287-92.

Rainbow, University of Economics Prague, Knowledge Engineering Group. (2005). *Reusable Architecture for Intelligent Brokering Of Web information access (Rainbow)*. Retrieved from: http://rainbow.vse.cz/descr.html

Sanger, S. (Ed.). (2004). *Check-In*. Retrieved from http://www.patienten-information.de/content/informationsqualitaet/informationsqualitaet/images/check_in.pdf

Soualmia, L. F., Darmoni, S. J., Douyère, M., & Thirion, B. (2003). Modelisation of Consumer Health Information in a Quality-Controled gateway. In Baud R et al. (ed.). *The New Navigators: from Professionals to Patients*. Proc of MIE2003, 701-706.

Stamatakis, K., Chandrinos, K., Karkaletsis, V., Mayer, M. A., Gonzales, D. V., Labsky, D. V., et al. (2007) AQUA, a system assisting labelling experts assess health Web resources. In *Proceedings of the 12th International Symposium for Health Information Management Research* (iSHIMR 2007), Sheffield, UK, 18-20 July, 75-84.

Stamatakis, K., Metsis, V., Karkaletsis, V., Ruzicka, M., Svátek, V., Amigó, E., & Pöllä, M. (2007). Content collection for the labelling of health-related web content. In *Proceedings of the 11th Conference on Artificial Intelligence in Medicine* (AIME 07), LNAI 4594, Amsterdam, 7-11 July, 341-345.

Wang, Y., & Liu, Z. (2006, May 31). Automatic detecting indicators for quality of health information on the Web. *Int J. Med Inform*.

Winker, M. A., Flanagan, A., & Chi-Lum, B. (2000). Guidelines for Medical and Health Information Sites on the Internet: principles governing AMA Web sites. American Medical Association. *Journal of the American Medical Association*, *283*(12), 1600–1606. doi:10.1001/jama.283.12.1600

KEY TERMS AND DEFINITIONS

Crawling: A web crawler is a program or automated script which browses the World Wide Web in a methodical, automated manner. This process is called web crawling. Web crawlers are mainly used to create a copy of all the visited pages for later processing.

Information Extraction: Automatic assignment of meaning to elementary textual entities and possibly more complex structured objects.

Metadata: Data that describes information about either online or offline data. Information that characterizes the who, what, where, and how related to data collection. Often, the information refers to special tagged fields in a document that provide information about the document to search engines and other computer applications. Web pages often include metadata in the form of meta tags. Description and keywords meta tags are commonly used to describe the Web page's content. Most search engines use this data when adding pages to their search index.

Resource Description Framework (RDF): Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model, but which has come to be used as a general method of modeling information through a variety of syntax formats. The RDF metadata model is based upon the idea of making statements about Web resources in the form of subject-predicate-object expressions, called triples in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object.

Semantic Web: The Semantic Web is an evolving extension of the World Wide Web in which the semantics of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the web content. It derives from W3C director Tim Berners-Lee's vision of the Web as a universal medium for data, information, and knowledge exchange. At its core, the semantic web comprises a set of design principles, collaborative working groups, and a variety of enabling technologies. Some elements of the semantic web are expressed as prospective future possibilities that have yet to be implemented or realized. Other elements of the semantic web are expressed in formal specifications.

Spidering: A web spider is a complementary mechanism/tool to a web crawler. Web crawlers are mainly used to create a copy of all the visited pages for later processing, whereas, web spiders are used to gather specific types of information from Web pages. Many sites, in particular search engines, use spidering as a means of providing up-to-date data.

Web Mining: Web mining is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. Web usage mining is the application that uses data mining to analyse and discover interesting patterns of user's usage data on the web. Web content mining is the process to discover useful information from the content of a web page. The type of the web content may consist of text, image, audio or video data in the web. Web structure mining is the process of using graph theory to analyse the node and connection structure of a web site.

ENDNOTES

1. http://www.medieq.org/aqua/welcome.seam
2. http://www.medieq.org
3. At the time of writing, degrees were not annotated as part of the Spanish collection and results for company and department names for Spanish and Czech were still work in progress.

Chapter 7.10
User Facing Web Services in Portals

Jana Polgar
NextDigital, Australia

ABSTRACT

In SOA framework, Portal applications aggregate and render information from multiple sources in easily consumable format to the end users. Web services seem to dominate the integration efforts in SOA. Traditional data-oriented web services require portlet applications to provide specific presentation logic and the communication interface for each web service. This approach is not well suited to dynamic SOA based integration of business processes and content. WSRP 2.0 aim at solving the problem and providing the framework for easy aggregation of presentation services. Is not practical to publish portlets locally if the organisation wishes to publish their portlets as web services to allow their business partners using these services in their portals. UDDI extension for WSRP enables the discovery and access to user facing web services while eliminating the need to design local user facing portlets. Most importantly, the remote portlets can be updated by the web service providers from their own servers.

VISION FOR USER-FACING PORTLETS

Web services introduced the means for integrating and sharing business processes via the Internet. WSRP (WSRP specification version 1 (2003)) goal is to extend the integration further by providing framework for sharing web service presentation components. WSRP specification formulated a standard protocol which enables all content and application providers to create web services, generate their presentation faces as HTML fragments and offer them to the consumers to be plugged into their local portals.

Portals and portlets (JSR 168 (2005)) provide specific presentation logic to aggregate data from multiple sources which could be legacy systems, Enterprise Information Systems (EIS), local or remote web services, or EIS with exposed web service interfaces. The first draft of JSR 286 (JSR 286 (2008) brings new features to the Java portlets capabilities introduced by WSRP 2.0 (WSRP Specification version 2.0 (2008)). JSR 286 new features include:

- Interportlet communication: coordination between portlets and allow building composite applications based on portlet components;
- Shared render parameters enable to specify which render parameters they can share with other portlets;
- Resource serving feature enables portlets to serve resources within the portlet context;
- Frameworks for better support for JSF and Struts
- Alignment with WSRP 2.0
- Better user experience using AJAX patterns
- Portlet filters to selectively define the portlets which can transform the content of portlet requests and responses on the fly.

The WSRP specification is intended for presentation-oriented web services, user-facing web services that can be easily integrated with portals. They let businesses provide content or applications without requiring any manual content or application-specific adaptation by portal presentation logic. It is envisaged that in the near future portals will easily aggregate WSRP services without any programming effort. The only effort required is the actual deployment of remote portlets in the local portal server (Hepper, S and Hesmer, S. (2003)). We are not taking into account the effort needed for the "implementation", that is the design of the portal page which is needed in any case.

The WSRP specification (WSRP specification version 1 (2003) and WSRP 2.0 are the effort of the working group at OASIS (http://www.oasis-open.org/committees/wsrp). It aims to provide a set of options for aggregating user-facing web services (remote portlets) from multiple remote web services within one portal application. WSRP standard has been conceived for implementing simple services. The developer of the portlet provides the markup fragments to display web service data. The current version allows for more complex services that require consumer registration, support complex user interaction, and operate on transient and persistent state maintained by the service provider. Before looking at the functionality of WSRP, note that what WSRP refers to as a portlet is the combination of a portlet implementation and any configuration data that supports the implementation. WSRP 2.0 (WSRP Specification version 2.0 (2008) is closely aligned with the JSR286 thus providing the framework for publishing JSR286 portlets as web services.

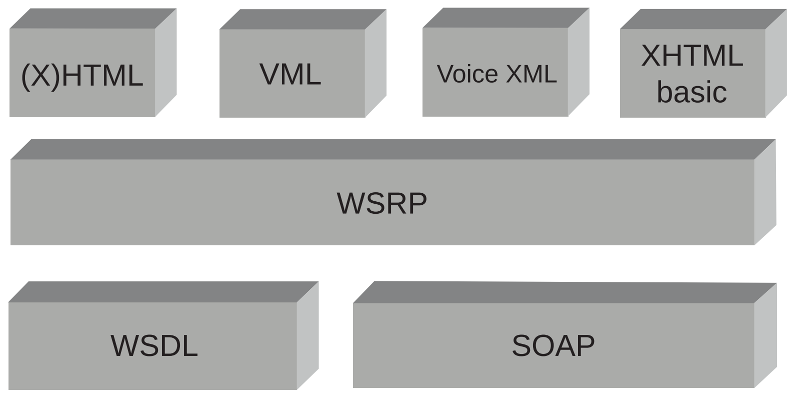
WSRP AND WSRP RELATED STANDARDS

WSRP defines the notion of valid fragments of markup based on the existing markup languages such as HTML, (X)HTML, VoiceXML, cHTML, etc (Figure 1). For markup languages that support CSS (Cascading Style Sheet) style definitions, WSRP also defines a set of standard CSS class names to allow portlets to generate markup using styles that are provided by WSRP compliant portals such that the markup assumes the look and feel of the consuming portal.

WSRP is fully integrated with the context of the web services standards stack. It uses WSDL additional elements to formally describe the WSRP service interfaces and requires that at least SOAP binding be available for invocations of WSRP services. WSRP also defines the roles of web service *producers* and *consumers*. Both *producers* and *consumers* use a standard protocol to provide and consume web services for user facing portlets. The WSRP specification requires that every *producer* implement two required interfaces, and allows optional implementation of two others:

1. **Service Description Interface (required):** This interface allows a WSRP *producer* to advertise services and its capabilities to

Figure 1. WSRP related standards



consumers. A WSRP *consumer* can use this interface to query a *producer* to discover what user-facing services the *producer* offers.

2. **Markup Interface (required):** This interface allows a *consumer* to interact with a remotely running portlet supplied by the *producer*.
3. **Registration Interface (optional):** This interface serves as a mechanism for opening a dialogue between the *producer* and *consumer* so that they can exchange information about each others' technical capabilities.
4. **Portlet Management Interface (optional):** This interface gives the *consumer* control over the life cycle methods of the remote portlet.

URL generation concept: To support user interaction, all the URLs embedded in the markup fragment returned by the remote *producer* service must point back to the *consumer* application. Therefore, the *consumer* needs to send a URL template as part of the invocation of the `getMarkup()` method. For example, the consumer may send the URL template with two variables: `navigationState` and `sessionID`:.

```
http://neptune.monash.edu.au/myApp?ns={navigationState}&si={sessionID}
```

The *producer* responsibility is to generate a markup fragment in which all the interaction URLs must point back to the *consumer*. The *producer* generates a link pointing to the URL replacing the template variables `navigationState` and `sessionID` with concrete values:

```
http://neptune.monash.edu.au/myApp?ns=page2&si=4AHH55A
```

Alternatively, the predetermined pattern allows the *producer* to create URLs that is compliant with this pattern. The *consumer* then parses the markup and rewrites variable parts of URL to point back to the application.

ROLE OF PRODUCERS AND CONSUMERS

WSRP is a protocol in which the interaction always occurs between two web applications or web services. The *consumer* application acts as a client to another application called *producer*. The *producer* provides end-user-facing (also called presentation services) web services in the form of remote portlets. These remote portlets are aggregated into the *consumer's* portal page in the same way as local portlets.

Let's start with comparing WSRP with a web services application. The web based application *consumer* uses HTTP, SOAP and browsers to interact with remote servers hosting web services. In response they receive web service raw **data** needed to create the markup (typically HTML or HTML form). The input data are posted by submitting the form via a browser.

HTTP protocol is also utilized with WSRP. *Consumers* can be seen as intermediaries that communicate with the WSRP *producers. Consumers* gather and aggregate the **markup** delivered by local as well as remote portlets created by the *producers* into a portal page. This portal page is then delivered over SOAP and HTTP to the client machine (PC or a workstation). The *consumer* is responsible for most of the interactions with the remote systems, ensuring user privacy and meeting the security concerns with regard to the processing information flow.

In a sense of additional capabilities, today's *consumers* of WSRP are more sophisticated than simple web service clients:

1. *Consumer* aggregates multiple interface components (local and remote portlets) into a single page. In addition, features like personalization, customization and security are also available for remote portlets;
2. The aggregation into a single page is not straightforward since it involves applying *consumer*-specific page layouts, style and skins to meet the end-user requirements. Therefore, the *consumer* must have knowledge of "presenting" related features in remote portlets to apply customization and rendering.
3. The *consumer* can aggregate content produced by portlets running on remote machines that use different programming environments, like J2EE and .NET.
4. *Consumers* are able to deal with remotely managed sessions and persistent states of WSRP web services.

The *producer* is responsible for publishing the service and portlet capabilities descriptions in some directory e.g. UDDI. It allows the *consumer* to find the service and integrate it into portal. The purpose of the portlet capabilities description is to inform the *consumer* about features each portlet offers. *Producer's* major responsibilities are listed below:

5. *Producers* are capable of hosting portlets (they can be thought of as portlet containers). Portlets generate markup and process interactions with that markup;
6. *Producers* render markup fragments which contains web service data.;
7. *Producers* process user interaction requests; and
8. *Producers* provide interfaces for self description, and portlet management.

The *consumer* can optionally *register* with the *producer.* The *producer* is responsible for specifying whether the registration is required. Typical registration contains two types of data: *capabilities* (for example, window states and modes the *producer's* remote portlets support), and *registration properties* (required data prescribed in the service description). Upon successful registration, the *consumer* receives a unique registration handle. This handle allows all portlets to be scoped to fit to the local portal. Optionally, the *consumer* may provide the credentials to the *producer.*

Portlet management is an optional interface implemented by the *producer.* It allows the *consumer* to manage the lifecycle of portlets exposed in the service description. These exposed portlets can be cloned and customized at the *consumer* portal. Note that the original portlets exposed in the service description cannot be modified.

Important points to note is that WSRP based web services are synchronous and UI-oriented. *Consumers* can invoke the web service in the usual way and interact with the service UI. The typical browser-server interaction protocol is then translated into protocol suitable for *consumers* of user facing web services. A typical processing would consist of the following steps:

- The web service interfaces exposed by the *producer* to the *consumer* are described using Web Services Description Language (WSDL). WSDL is the mandatory interface between the client and service that enables the client to bind to the service and use it;
- Optionally, *consumers* can be registered in a *producer's* portal;
- Portal detects the remote portlet on its page and sends getMarkup() message to the *producer.* The markup interface supports end user interaction and it is another mandatory interface in WSRP;
- In response it receives a HTML fragment from the *producer;*
- Portal (*consumer*) aggregates the fragment into the portal page; and
- Optional functionality is the use of the portlet management. The portlet management defines operations (API) for cloning, customizing and deleting portlets.

The actual interaction between WSRP *consumers* and *producers* is more complex. We assume that the user can dynamically add a portlet to the portal page. In response, the portal invokes the WSRP remote service. This action specifies a new portlet instance that allocates a corresponding portlet instance on the portal side. When a user wants to view this portlet, the portal obtains the WSRP markup that defines the fragment to be displayed. The returned markup contains portlet action links and/or a portlet session identifier. When the user clicks on the link (*Click-on-Action*), a request goes from the browser to the portal. The portal maps the request into the invocation of the WSRP service. The capability to maintain the session identity is provided through the parameters that are passed, such as the session ID. This allows the WSRP service to look up the previous session details. When the user does not want to access the WSRP service any more, the session is closed, the portlet is removed, and its instance is destroyed.

WSRP PROCESSING SCENARIOS

The goal of WSRP is to make implementation of remote web services and access to the remote content easy. WSRP service scenarios come in several flavours ranging from simple view to complex interactions and configurations. Please note that our examples are based on IBM's WebSphere 5.1 Portal server. Some of the operations could be implemented differently on IBM Websphere 6.1 Portal or on other vendors' platforms. There are typically three different situations to deal with remote portlets: simple case of just processing view portlet, user interaction and dealing with the state information, and handling of configuration and customization.

REGISTRATION PROCESS

We have to start with two steps which have to be performed in all scenarios at the *consumer* portal:

Registering with the producer portal allows the *producer* to be known to the consumer and make available the list of WSRP services that could be consumed by the consumer portal. There are possible situations:

- Consumer has *online* access to the *producer.* In this scenario it is possible to use the XML configuration interface to configure new *producer* and remote web services. If in-band registration is supported in the producer, the consumer can register through the WSRP registration port type (register() call).

   a. If in-band registration is not supported by the producer, the consumer administrator must manually obtain the registration handle from the *producer*'s administrator.
   b. If the registration is required by the *producer*, it is necessary to implement a registration validation process for informing the producer whether a registration data from the consumer are valid.
- If the *consumer* works *offline* with regard to the *producer,* only the XML configuration interface can be used to create a *producer.*

Consuming the WSRP service allows you to integrate WSRP services from registered *producers* into the *consumer* portal and interact with them as they were local portlets.

WSRP 2.0 provides additional APIs relevant to the portlet lifetime: `set|RegistrationLifetime` and `getRegistrationLifetime` which allow the management of the registration.

SIMPLE VIEW PORTLET

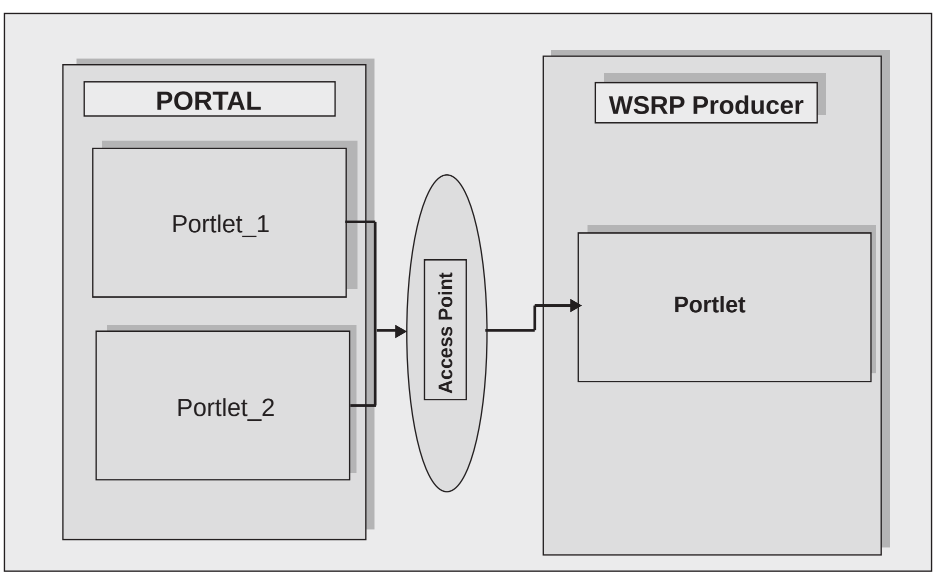
In our simple View portlet example, we assume that the web service requires only to be viewed by the end-user. Portlet has to be rendered and no interaction or forms are implemented.

Based on our description of available APIs, we need only `getMarkup()`operation to be implemented (Figure 2). This operation returns WSRP markup fragment which is then aggregated in the portal page.

INTERACTIVE SERVICE WITH TRANSIENT CONVERSATIONAL STATE

In this scenario, we need the WSRP implementation to support user interaction and maintain the conversational state of the application. Similarly to servlets (Servlets Specification 2.4 (2004)), the WSRP protocol operates over stateless HTTP. In order to generate correct responses, the application must be stateful and maintain its state. The state

Figure 2. Simple view portlet



may span across several request/response cycles. The WSRP protocol distinguishes between two states: transient and persistent (Figure 3). `Navigational state` is used when *producer* requires generation of markup for the portlet, several times during its conversation with the *consumer*. This state locally encapsulates required data needed to keep track of the conversation about the current state of the portlet. It means that the *producer* does not hold the transient state locally and the user can store or bookmark the URL using the navigational state. The state is stored with the URL only and both *page refresh* and *bookmarked pages* generate the output the end user expects. The session state is maintained using `sessionID` which is generated when the portlet initializes the session for a particular end-user. During the interaction the `sessionID` is moved between the *producer* and *consumer*.

The persistent state survives the conversation and will cease to exist only when either *consumer* or *producer* are discarded. The persistent state is the property exposed by the *producer* via the portlet management interface. In the case of registration (`Consumer Registration`), the registration state is maintained with the help of the `registrationHandle` generated during the consumer registration. WSRP protocol allows the consumer to customize the portlet and keep its state using `portletHandle`.

As an example we use again the University course offerings service that provides an overview of subjects offered in different semesters and allows users to click on the course offerings to navigate to the individual subjects and then on a "back-link" navigate back to the course offerings. Such a service should maintain conversational state within a *WSRP Session* to always display the correct view for a particular user and return a session ID for an internally managed session in each response of the `getMarkup()` operation (Figure 4). The markup returned may also contain links that will trigger invocations of the `performBlockingInteraction()` operation. This operation allows the portlet to perform logical operations updating state that could be shared with other portlets at the *producer*.

INTERACTIVE SERVICE CONTAINING PERSISTENT DATA

Let us consider a remote service that maintains configuration data that can be associated with

Figure 3. WSRP states

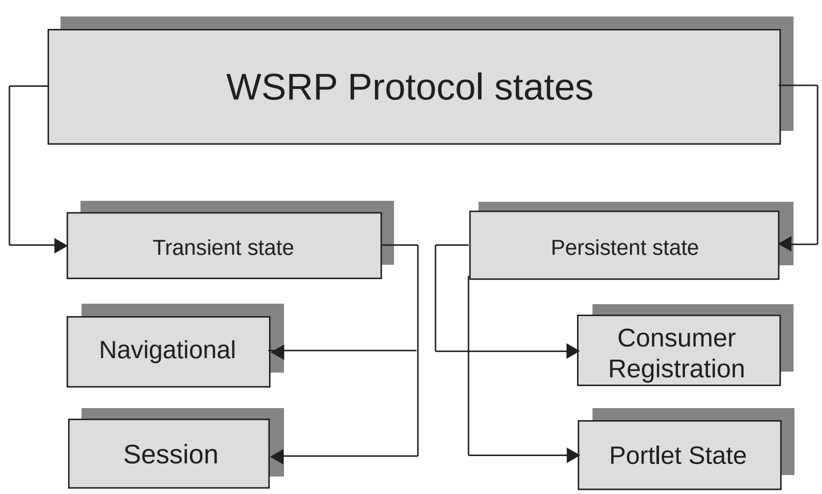
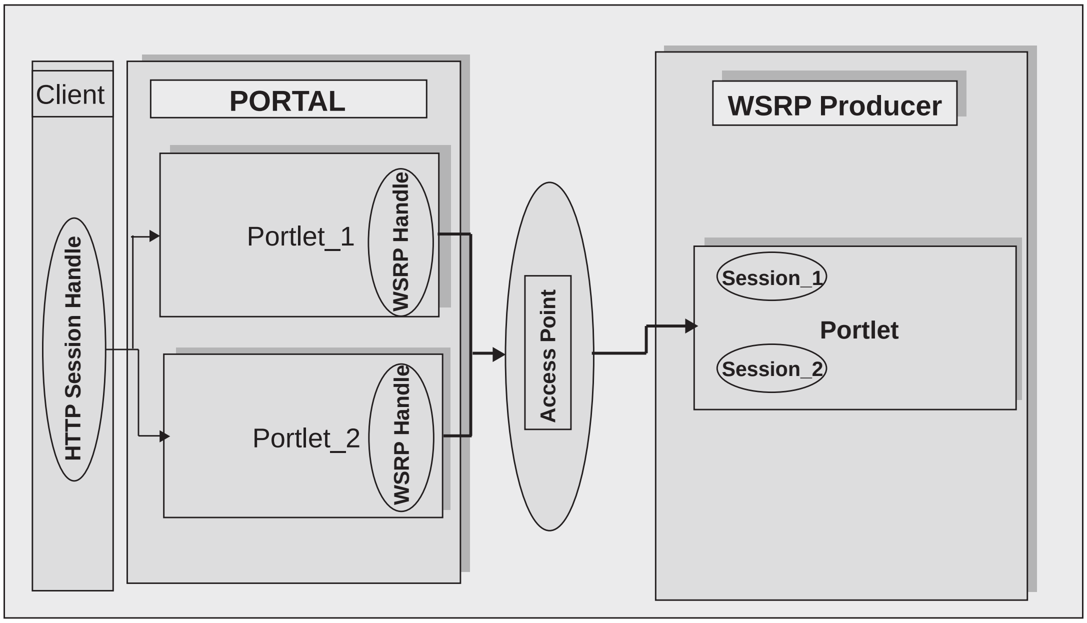


Figure 4. Conversational interactive services



individual portlets available from the *producer*. An example for such a service is a tutorial allocation service that allows individual users to define their own personal schedules for tutorials. This situation requires the implementation of configuration data and ability to retain application persistent state for the end user.

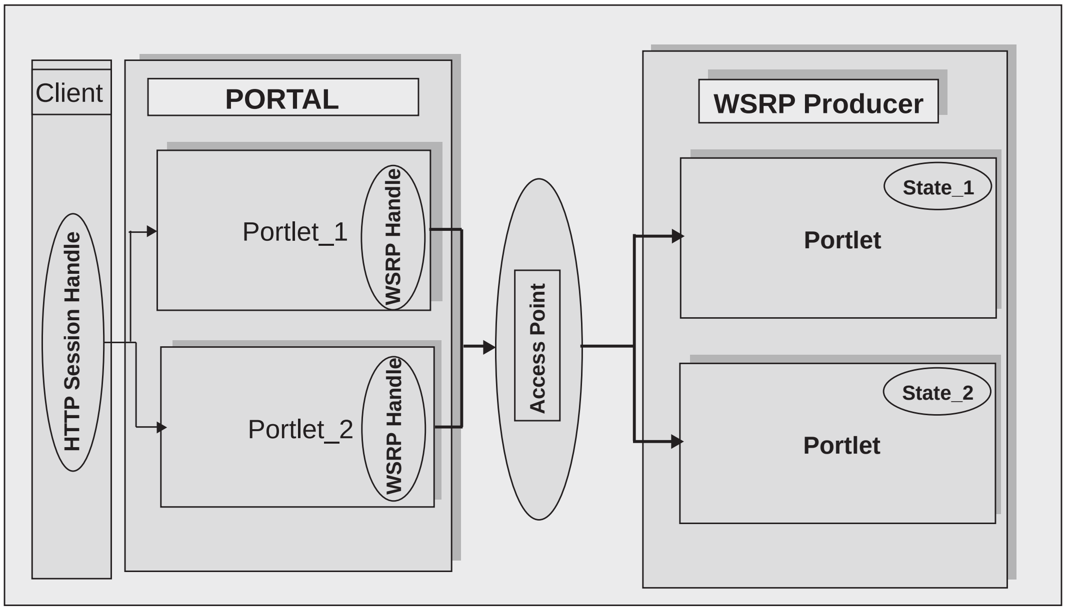
Since customization of portlets is not available in WSRP protocol, the *consumers* create new portlets using `clonePortlet` (Figure 5), specifying an existing portlet – either a producer offered portlet or one previously cloned by the consumer. The new portlet will be initialized with the same configuration data as the existing portlet. New portlets can also be cloned during the processing of a `performBlockingInteraction()` method. This is enabled when the *consumer* sets a flag preventing the user to customize the configuration data of the supplied portlet. The clone operation returns a portlet with updated configuration data and the customization is allowed. The portlet implementation can also make an attempt to update its configuration. This attempt typically results in the *producer* cloning the configuration data and applying the update to the cloned configuration. In either of these cases, the consumer obtains a handle (`portletHandle`) for referring to the new portlet when calling the *producer*.

When a portlet is no longer needed, it can be discarded by calling `destroyPortlets()`, passing the portlet handle. At this point, all persistent data can be discarded as well.

INTERACTIVE SERVICE CONTAINING CONFIGURATION DATA AND MAINTAINING SESSION

The *producer* may need to use both configuration data and transient session state to satisfy the application requirements. Several remote sessions may be associated with a portlet at any given time. For example, many remote sessions to the same portlet may exist for a *consumer* that is a portal with shared pages referencing the portlet and being used concurrently by multiple end users (Figure 6).

Figure 5. Interactive service with configuration data



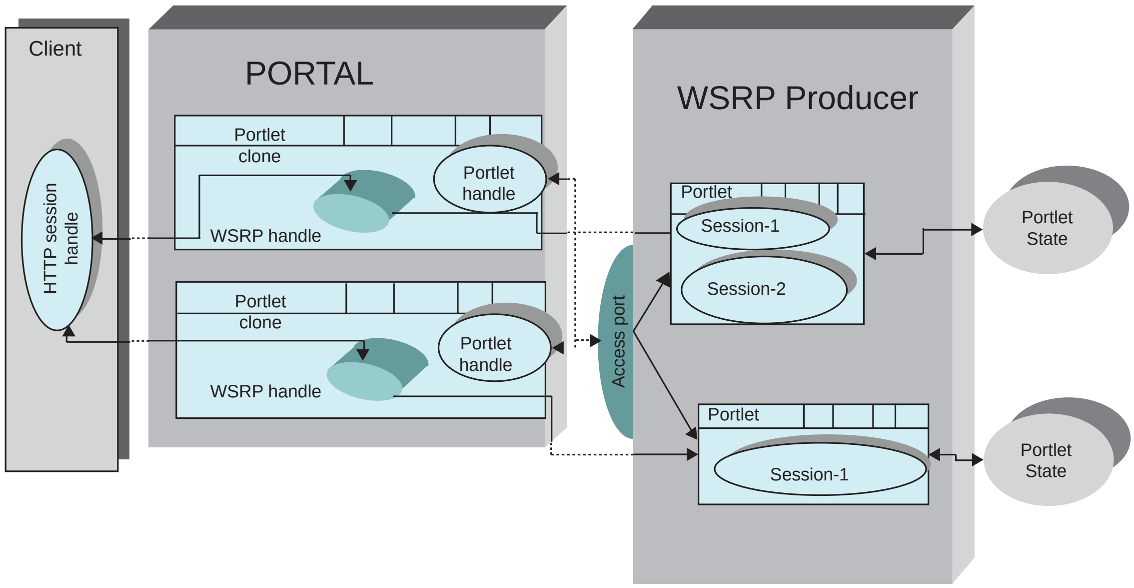
A typical information flow pattern starts with the end-user adding the remote portlet to a page. This is done for example by portal administrators via administration interface or XML configuration interface. The portlet invokes `clonePortlet()` operation on the remote service specifying an existing portlet and optionally including pre-configuration data. In return it obtains a new portlet handle (`portletHandle`) that it stores together with a newly created portlet instance on the portal database. The reason for cloning is that the original portlets exposed in the service description cannot be customized.

In the view mode, the portal determines the portlet handle (`portletHandle`) and uses it to make a call to the `getMarkup()` operation of the remote service. The operation returns the HTML fragment to be aggregated and displayed in the page within a `doView()` operation.. The response may contain action links, and could include a session handle (`sessionID`) if the portlet wants to maintain the conversation state. The portal typically needs to rewrite any action links to point to the *consumer* site and must store any returned session handle in a manner that allows it to be used on subsequent requests.

When the user clicks on an action link in the markup, a HTTP request is sent from the browser to the portal. The portal processes the request and maps it to an invocation of the `performBlockingInteraction()` operation of the remote service and passes the `sessionID` which allows the remote service to look up the associated session state. In the `performBlockingInteraction()` invocation, the remote service typically changes the state. When the `performBlockingInteraction()` operation returns, the portal refreshes the page. This results in an invocation of `getMarkup()` on all the portlets on the page and starts a new user-interaction cycle.

When an end user is finished with a portlet instance and discards it from a portal page, the portal recovers the handle of the portlet which is no longer needed and invokes `destroyPortlets()` on the remote service. The remote service discards the portlet and is free to release any resources associated with this portlet.

Figure 6. Interactive service with configuration data and session maintenance



RESTFUL WEB SERVICES

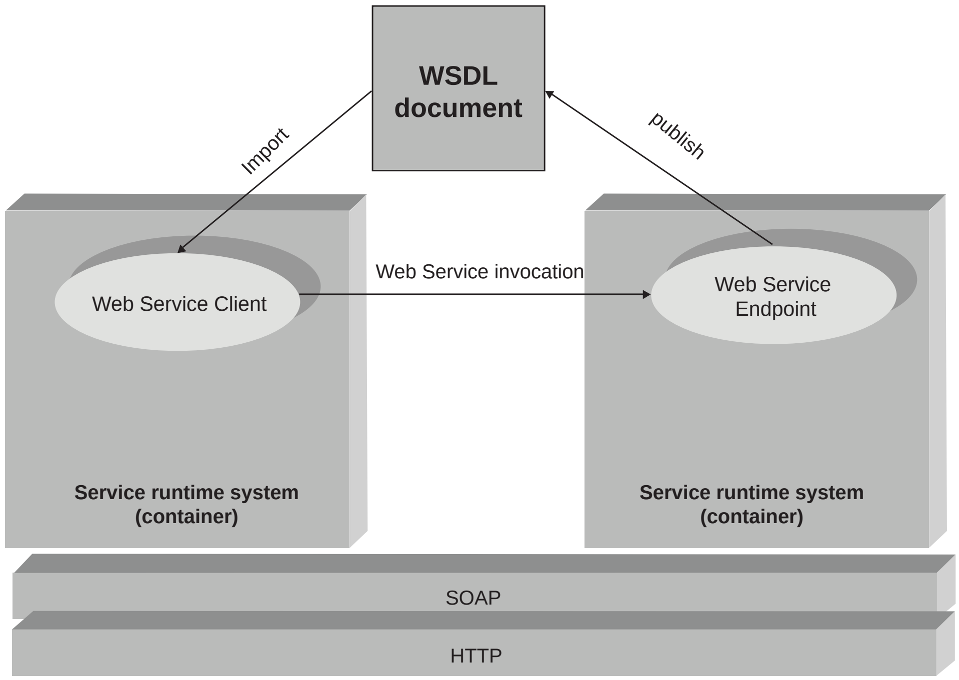
Data oriented web services are characterised by their complexity. Their development involves implementing various infrastructural components (WSDL, SOAP). Web services solution has to invest in creating a robust Web service infrastructure model. From the development point of view, it becomes increasingly complex to design and learn the technology. Presentation oriented services such as WSRP based services provide relief from the complexity of the infrastructure. The newly introduced features from JSR 286 provide sufficient flexibility in terms of inter-portlet communication and event processing, AJAX use, and resource serving capability. The presentation logic embedded in the remote portlet takes care of the easy rendering.

The new wave in web services are the RESTful Web services characterised by a simple XML-over-HTTP transmission. The RESTful services encapsulate data in a simple XML form and transport it over HTTP the same way as a Web page request. It takes full advantage of the REST architecture style which is related to a Web resource. In turn, this Web resource is a representation identified by a Uniform Resource Indicator (URI). The resource can be any persistent entity, and queries or updates the resource are applied through the URI and therefore influence a state change in its representation. In REST, a user to invoke operations on a Web resource using HTTP request methods in a Web service style. REST is closely associated with HTTP and leverages all HTTP features, such as methods, headers, and types.

ROLE OF UDDI IN WEB SERVICES

Portlets (JSR 168 (2005)) provide user interface to data delivered from web services. Before we explain the remote portlet publishing and discovery process in UDDI, we need to refresh the concept of publishing and discovering the web services in UDDI (Hugo Haas, P. L. H., Jean-Jacques Moreau, David Orchard, Jeffrey Schlimmer, Sanjiva Weerawarana (2004)). Web services expose their interfaces by registering in

Figure 7. Publish-Find-Bind Mechanism in UDDI



UDDI (UDDI Specifications (2005)). The web service consumer must find the service, bind to it and invoke the service. The basic mechanism for publishing and discovering data – oriented Web services is in Figure 7.

Regardless of whether the web service will be accessible to a single enterprise or to other companies (public access), the details about the service (its interface, parameters, location, etc.) must be made available to *consumers*. This is accomplished with a WSDL description of the Web service and a Web service directory where the details of the Web service are published (refer to Web Services Description Language (WSDL)). There are three steps which have to be performed in order to discover and use a web service published in the UDDI:

Publishing web service (step 1): In order to be accessible to interested parties, the web service is published in a Registry or web service directory. There are several choices regarding where to publish a web service:

1. If the web service is intended for the general public then a well-known registry is recommended. Consequently the WSDL description together with any XML schemas referenced by this description is made public.
2. The web service intended for enterprise use over an intranet should be published in a corporate registry only. No public access from the outside of the firewall is required.
3. Finally, providing all clients are dedicated partners in business, and there is an existing agreement on usage of this service, the web service can be published on a well-known location on the company server - with proper security access protection. Such a server would be placed on the public side of the company firewall but it would allow limited access, similar to a B2B Web server.
4. Web services directories are made up of a repository and the taxonomies (classification of registered entities for easier search) as-

sociated with them. There are no restrictions on publishing the web service in multiple registries, or in multiple categories.

Discovery of web service (step 2): Registry implementations can differ but there are some common steps, outlined below, that the client must perform before it can discover and bind (step 3) to the service:

1. The client must determine how to access the web service's methods, such as determining the service method parameters, return values, and so forth. This is referred to as *discovering the service definition interface*.
2. The client must locate the actual web service (find its address). This is referred to as *discovering the service implementation*.

Bind to the web service and invoke it (step 3): The client must be able to bind to the service's specific location. The following types of binding may occur:

1. Static binding during client development or at the deployment time.
2. Dynamic binding (at runtime).

From the client point of view, the binding type and time play important roles in possible scenarios relevant to the client's usage of the web service. The following situations are typical:

1. A web service (WSDL and XML schemas) is published in well-known locations. The developers of the application that use the service know the service, its location, and the interface. The client (which is a process running on a host) can bypass the registry and use the service interfaces directly. Alternatively, the client knows the location and can statically bind to the service at the deployment time.
2. The web service expects its clients to be able to easily find the interface at build time. These clients are often generic clients. Such clients can dynamically find the specific implementation at runtime using the registry. Dynamic runtime binding is required.

Development of web service clients requires some rules to be applied and design decisions to be made regarding which binding type is more appropriate for the given situation (static or dynamic binding). Three possible cases are discussed:

1. *Discovering the service interface definition*: If we are dealing with a known service interface, and the service implementation is known (no registry is required), the actual binding should be static.
2. *Discovering the service implementation*: In this case, static binding is also appropriate because we know the interface. We need to discover the service implementation only at build time.
3. The client does not know the service interface and needs to discover the service interface dynamically at build time. The service implementation is *discovered dynamically at runtime*. This type of invocation is called Dynamic Invocation Interface (DII). In this case, the binding must be dynamic.

Each WSDL description of the service published in UDDI must contain the following six elements: definitions, types, message, portType, binding, and service. The main elements of the UDDI data model are listed below (Figure 8):

- `businessEntity` represents the physical company which registered the services with UDDI;
- `businessService` represents a specific service offered by a company;
- `bindingTemplate` contains instructions for service invocation;

- publisherAssertion structure allows businesses to publish relationships between businessEntities within the company; and
- `tModel` is a structure similar to a database table. It contains the following information about an entity: the name, description, URL, and the unique key.

The relationships between the description and actual registered structures are outlined in Figure 9. The `portType` is represented by a UDDI structure called `tModel`. This `tModel` is categorized using unified *Category System* and the WSDL `EntityType` structure. The relevant *Category System* is known as WSDL `portType tModel` category and distinguishes it from other types of `tModels` with which the service might be associated.

A WSDL binding is also represented by a `tModel` structure. This is the binding `tModel` structure. This kind of categorization uses the same *Category System* as the `portType` tModel, but with a different key value to differentiate a binding tModel from a portType tModel.

The WSDL may represent a web service interface for an existing service. However, there may be an existing UDDI `businessService` that is suitable, and WSDL information can be just added to that existing service. If there is no suitable existing service found in the UDDI registry, a new `businessService` must be created. Finally, the WSDL binding port is represented by UDDI `bindingTemplate`. A WSDL service may contain multiple ports. These ports are exactly mirrored by the containment relationship in a UDDI `businessService` and its `bindingTemplates`.

REGISTERING WSRP SERVICES AS REMOTE PORTLETS IN UDDI

WSRP *producer* is considered as a web service on its own, exposing multiple `Bindings` and `Port-`

Figure 8. UDDI model composition

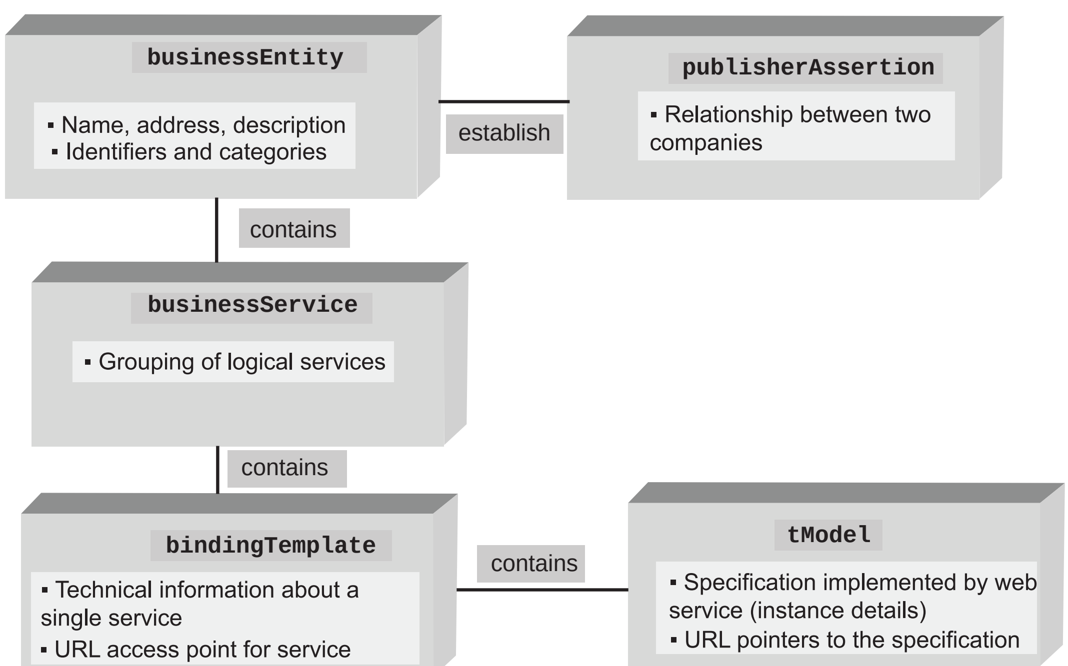
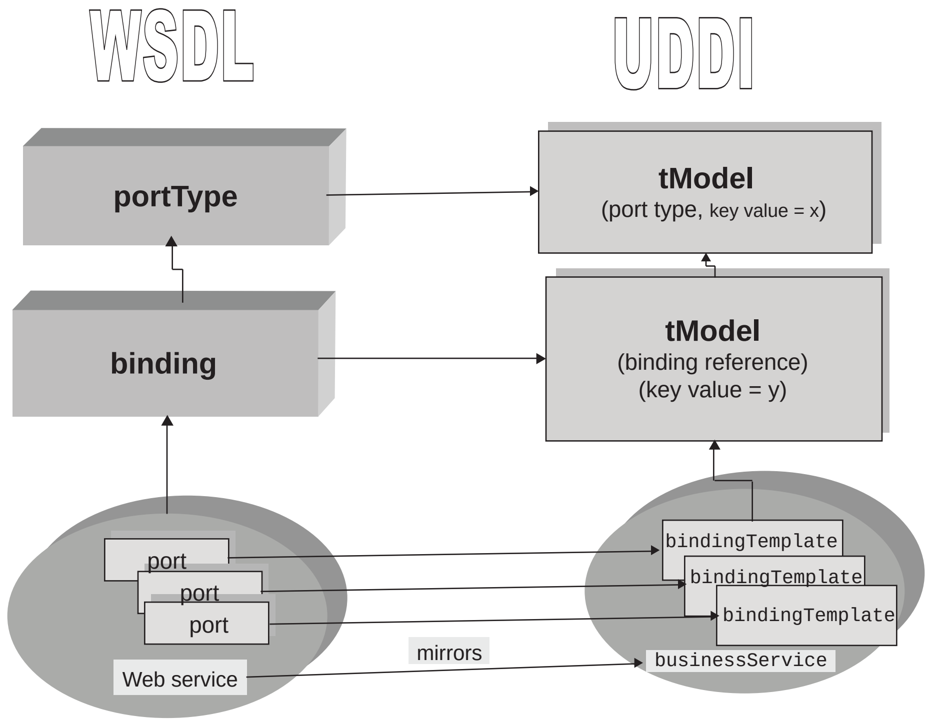


Figure 9. Mapping from WSDL to UDDI



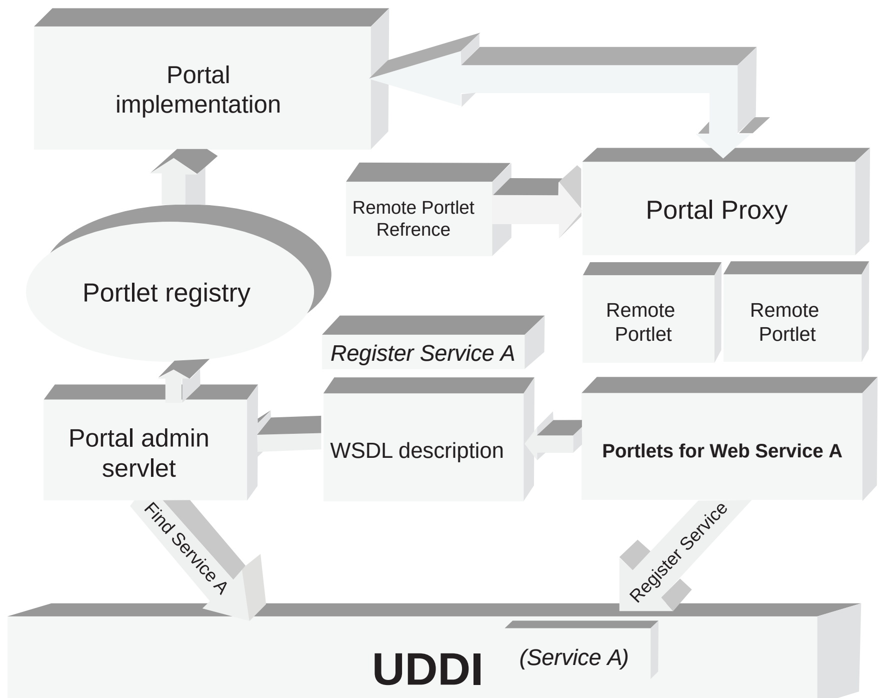
`Types`. It is described through the WSRP WSDL services description and some additional portlet types. Portlets are not fully fledged services, they are only HTML fragments. Therefore, they do not expose `PortType, binding` template and access points. The portlet is exposed by its *producer* and *consumer* interacts indirectly with remote portlets using the *producer's* infrastructure. The remote portlet is addressed by a `portletHandle` defined within the *producer's* scope.

Figure 10 shows an example how a portal finds and integrates a remote portlet published in the UDDI. Content or application providers (known as WSRP *producers*) implement their service as WSRP service and publish it in a globally accessible directory. *Producer's* WSDL description provides the necessary information about remote service actual end-points. The directory lets the *consumers* easily find the required service. Directory entries, published in WSDL format, briefly describe the WSRP components and offer access to details about the services. The portal administrator uses the portal's published functions to create remote portlet web service entries in the portal local registry. Furthermore, the portlet proxy binds to the WSRP component through SOAP, and the remote portlet invocation (RPI) protocol ensures the proper interaction between both parties.

Typical discovery and binding steps are summarized below:

- A provider offers a set of portlets and makes them available by setting up a WSRP *producer* and exposing them as remote portlets. These portlets are then made available to other businesses by publishing them in a UDDI registry. The provider may perform

Figure 10. Publishing and locating remote portlets with the UDDI



the publishing task either through a custom built user interface or through the interface provided by a UDDI Server.

- End-user wants to add a portlet to his own portal. Using the tools provided by his portal (for example portal administrative interface or a custom-written XML interface[1]), he/she searches for remote portlets. After finding the suitable remote portlet, these portlets can be added to the portal pages. Alternatively, a portal administrator could search the UDDI registry for portlets and make them available to end-users by adding them to the portal's internal database.
- The user can now access the page containing newly added and running remote portlets. Behind the scenes, the portal is making a web service call to the remote *producer*, and the *producer* is returning a markup fragment with the required data for the portal to render on the portal page.

In order to provide necessary information about remote portlets, WSRP extended the definition of the bind namespace for `portTypes` and SOAP binding. The following extensions are defined (WSRP specification version 1 (2003). This WSDL defines the following `portTypes` (normative definitions):

- **WSRP_v1_Markup_PortType:** This is the port on which the Markup Interface can be accessed. All *producers* must expose this `portType`.
- **WSRP_v1_ServiceDescription_PortType:** This is the port on which the Service Description Interface can be accessed. All *producers* must expose this `portType`.
- **WSRP_v1_Registration_PortType:** This is the port on which the Registration Interface can be accessed. Only *producers* supporting in-band registration of *consumers* need expose this `portType`.

- **WSRP_v1_PortletManagement_PortType:** This is the port on which the Management Interface can be accessed. *Producers* supporting the portlet management interface expose this `portType`. If this `portType` is not exposed, the portlets of the service cannot be configured by consumers.

SOAP bindings for these `portTypes` are listed below:

1. **WSRP_v1_Markup_Binding_SOAP:** All *producers* must expose a port with this binding for the `WSRP_v1_Markup_PortType` (the `Markup portType`).
2. **WSRP_v1_ServiceDescription_Binding_SOAP:** All *producers* must expose a port with this binding for the `WSRP_v1_ServiceDescription_PortType` (`ServiceDescription portType`).
3. **WSRP_v1_Registration_Binding_SOAP:** *Producers* supporting the `Registration portType` must expose a port with this binding for the `WSRP_v1_Registration_PortType`.
4. **WSRP_v1_PortletManagement_Binding_SOAP:** *Producers* supporting the `PortletManagement portType` must expose a port with this binding for the `WSRP_v1_PortletManagement_PortType`.

Web service is typically represented by several remote portlets and relevant WSDL description (Figure 11) which contains pointers to all required and optional WSRP portlet interfaces (e.g. registration interface, service description, etc.) in the form of a `portType`.

In essence, WSRP *producers* are web services. They expose `PortTypes` and `bindings` which the *consumers* can use to access and interact with. It means that the process of publishing a *producer* corresponds to publishing a web services together with associated portlet metadata. Besides the `portletHandle`, the `Portlet Title` and textual description, all further portlet metadata are missing in the UDDI. These remaining metadata must be retrieved from the respective ports (`ServiceDescription portType` or `PortletManagement portType`).

Presentation oriented service has been developed to ease the burden of complexity of data oriented services. Specifically, to eliminate the need of developing the presentation logic at the consumer site. It is still using SOAP as main transport feature. There is still need to take into account the binding to service markup and service description.

SUMMARY AND CRITICAL LOOK AT WSRP

WSRP can be used to create powerful portal services from originally non-portal-centric applications. WSRP provides easy access to remote web services and their user-facing representations. Web services offer a mechanism to create remotely accessible and platform independent services. Portlet standard - JSR 168 - complements this mechanism by defining a common platform and APIs for developing user interfaces in the form of portlets. WSRP enables reuse of these portlets. Only one generic proxy is required to establish the connection. The WSRP could be used to facilitate the development of an entire network of presentation-oriented web services. It would allow the portal users easily discover and use any number of remote services. There is no need to develop custom adapters, build client interfaces, and spend time locally deploying the customized portlets.

WSRP 1.0 is lacking any standard for transaction handling, there are some problems associated with security, reliability, and load balancing[2]. Furthermore, the response time could be unpredictably long. The portal pages are aggregated from multiple *producers* and portal must wait until all

Figure 11. WSDL definition for WSRP example

```
<?xml version=”1.0” encoding=”UTF-8”?>
<wsdl:definitions xmlns:urn=”urn:oasis:names:tc:wsrp:v1:bind”
    xmlns:wsdl=”http://schemas.xmlsoap.org/wsdl/”
    targetNamespace=”urn:myproducer:wsdl”>
  <wsdl:import namespace=”urn:oasis:names:tc:wsrp:v1:bind”
    location=”http://www.oasis-open.org/committees/wsrp/
      specifications/version1/wsrp _ v1 _ bindings.wsdl”/>
  <wsdl:service name=”WSRPService”>
    <wsdl:port name=”WSRPBaseService”
      binding=”urn:WSRP _ v1 _ Markup _ Binding _ SOAP”>
      <soap:address xmlns:soap=”http://schemas.xmlsoap.org/wsdl/soap/”
      location=”http://myproducer.com:9098/portal/producer”/>
  </wsdl:port>
  <wsdl:port name=”WSRPServiceDescriptionService”
      binding=”urn:WSRP _ v1 _ ServiceDescription _ Binding _ SOAP”>
      <soap:address xmlns:soap=”http://schemas.xmlsoap.org/wsdl/soap/”
      location=”http://myproducer.com:9098/portal/producer”/>
  </wsdl:port>
  <wsdl:port name=”WSRPRegistrationService”
    binding=”urn:WSRP _ v1 _ Registration _ Binding _ SOAP”>
      <soap:address xmlns:soap=”http://schemas.xmlsoap.org/wsdl/soap/”
        location=”http://myproducer.com:9098/portal/producer”/>
  </wsdl:port>
  <wsdl:port name=”WSRPPortletManagementService”
    binding=”urn:WSRP _ v1 _ PortletManagement _ Binding _ SOAP”>
      <soap:address xmlns:soap=”http://schemas.xmlsoap.org/wsdl/soap/”
      location=”http://myproducer.com:9098/portal/producer”/>
    </wsdl:port>
  </wsdl:service>
</wsdl:definitions>
```

fragments are ready for rendering. Any remote service may slow down the entire portal.

WSRP 2.0 is fully aligned with the portlet specification 286 and contains all additional features announced with JSR 286. Therefore, it supports building composite applications using coordination means. The event and public parameters support loose coupled event paradigm. Similar as JSR 286 it also allows for additional AJAX use cases utilizing resource serving through the portlet. The capability of setting HTTP headers and cookies, filters, request dispatching provides a framework for better integration with servlets.

Using WSRP and UDDI extension for remote portlets, makes the end-user completely shielded from the technical details of WSRP. In contrast to the standard use of data-oriented web services, any changes to web service structure are implemented within the remote portlet and the *consumer* is not affected by these changes.

UDDI version 1.1 allows the *producers* to describe its presence together with each of the services it offers. The most important feature planned for higher versions of UDDI specification (specifically version 2 and higher) is the provision of cross portlet communication. Portlets should be able to broadcast their event information to other portlets spread across multiple *producers* if necessary. This feature allows other portlets to tailor their generated content according to broadcasted events. This feature is being well supported by the WSRP 2 which enables the inter portlet communication on the consumer site.

So far, there is seemingly no need to publish remaining portlet metadata. However, we envisage that the concept of semantic web and web service matchmaking as outlined in R. Akkiraju, R. Goodwin, Prashant Doshi, Sascha Roeder (2003) will require better annotation of available remote portlets functionalities to be published in a public registry. In such case, searching for portlets defining certain metadata values in UDDI will become the necessity.

Comparing WSRP and RESTful Web service, the latter does not provide any presentation logic. However, RESTfull web services rely on standard HTTP protocol, utilizing the power of the resource URI to maintain the resource state. WSRP uses classic web service infrastructure (WSDL, UDDI and SOAP), which still requires the negotiation of various contracts between the provider and consumer. The burden of the implementation is leveraged by the presentation logic being provided by the producer.

With data oriented services, the portlet displaying web service's raw data arriving from `a UDDI businessService` structure (web service) reflects the infrastructure of the web service and needs to bind to the service.This is an undesirably tight coupling of user interface and service raw data which often cause problems to the *consumer* in time of any changes to web service raw data. This problem is typically resolved by the *producer* providing relevant libraries.

REFERENCES

JSR 168 (2005). Portlet Specification, http://www.jcp.org/en/jsr/detail?id=168

Servlets Specification 2.4 (2004). http://www.jcp.org/aboutJava/communityprocess/final/jsr154, last accessed November, 2005

JSR 286 (2008). Portlet Specification, http://jcp.org/en/jsr/detail?id=286

Danny Coward, Y. (2003). JSR-000154 Java™ Servlet 2.4 Specification (Final Release). Sun Microsystems Inc. http://www.jcp.org/aboutJava/communityprocess/final/jsr154/

Hepper, S and Hesmer, S. (2003). Introducing the Portlet Specification, JavaWorld, last accessed 2005, http://www-106.ibm.com/developerworks/websphere/library/techarticles/0312_hepper/hepper.html

Web Services Description Language (WSDL): An Intuitive View. developers.sun.com. http://java.sun.com/dev/evangcentral/totallytech/wsdl.html

WSRP specification version 1 (2003). Web Services for Remote Portlets, OASIS. http://www.oasis-open.org/committees/download.php/3343/oasis-200304-wsrp-specification-1.0.pdf. Last accessed 2005

WSRP Specification version 2.0 (2008). http://docs.oasis-open.org/wsrp/v2/wsrp-2.0-spec-os-01.html#_Toc04

Web Services Description Language (WSDL): An Intuitive View. developers.sun.com. http://java.sun.com/dev/evangcentral/totallytech/wsdl.html

R. Akkiraju, R. Goodwin, Prashant Doshi, Sascha Roeder (2003). A Method for Semantically Enhancing the Service Discovery Capabilities of UDDI. *In the Proceedings of IJCAI Information Integration on the Web Workshop*, Acapulco, Mexico, August 2003. www.isi.edu/info-agents/workshops/ijcai03/papers/Akkiraju-SemanticUDDI-IJCA%202003.pdf

Hugo Haas, P. L. H., Jean-Jacques Moreau, David Orchard, Jeffrey Schlimmer, Sanjiva Weerawarana (2004). Web Services Description Language (WSDL) Version 2.0 Part 3: Bindings. W3C. http://www.w3.org/TR/2004/WD-wsdl20-bindings-20040803

UDDI Specifications (2005). Universal Description, Discovery and Integration v2 and v3. http://www.uddi.org/specification.html, last accessed November, 2005

WSRP specification version 1 (2003). Web Services for Remote Portlets, OASIS. http://www.oasis-open.org/committees/download.php/3343/oasis-200304-wsrp-specification-1.0.pdf. Last accessed 2005.

ENDNOTES

1. In IBM WebSphere Portal 5.1, this activity is supported via the configuration portlets or XML configuration interface

This work was previously published in the International Journal of Web Portals, Vol. 1, Issue 2, edited by J. Polgar and G. Adamson, pp. 44-66, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 7.11
Hyperlink Structure Inspired by Web Usage

Pawan Lingras
Saint Mary's University, Canada

Rucha Lingras
Saint Mary's University, Canada

ABSTRACT

This chapter describes how Web usage patterns can be used to improve the navigational structure of a Web site. The discussion begins with an illustration of visualization tools that study aggregate and individual link traversals. The use of data mining techniques such as classification, association, and sequence analysis to discover knowledge about Web usage, such as navigational patterns, is also discussed. Finally, a graph theoretic algorithm to create an optimal navigational hyperlink structure, based on known navigation patterns, is presented. The discussion is supported by analysis of real-world datasets.

DOI: 10.4018/978-1-59904-990-8.ch024

INTRODUCTION

The structure of a Web site is usually based on how the designer envisions the site will be used. However once the Web site is put into use, the designer's theoretical approach may turn out to be not so practical. It is only the actual use of the Web site that will give the designer clues about how the users navigate through the site and in what content they are most interested. For example, if the users are flocking towards a particular type of content, then the designer could think about establishing the site as an authority on that type of content by providing more information on the topic. Conversely, if a type of content is not generating much attention, it may be because the users are missing the information. The designer may look into changing its location on the Web site.

The users' navigational patterns help reveal the users' interests, but they can also be used to adjust the hyperlink structure of the Web site for optimal navigation. This chapter describes how Web usage can be used to construct a hyperlink structure that is easier to navigate. The discussion includes data visualization, uncovering navigational patterns using conventional data mining techniques, as well as a graph theoretical algorithm to construct an improved navigational structure.

The second section of this paper provides a background on various tools to analyze Web navigation. The next section describes an application of data mining techniques to discover Web navigation patterns. An algorithm to create an optimal hyperlink structure is presented in afterwards. The chapter concludes by summarizing the content and identifying areas for future research and development.

BACKGROUND

Web usage mining applies data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. While Web content mining and Web structure mining utilize the information found in Web documents, Web usage mining uses secondary data generated by the users' interaction with the Web. Web access logs available on most servers are good examples of the datasets used in Web usage mining. Other Web usage data may include browser logs, user profiles, registration files, user sessions or transactions, user queries, bookmark folders, as well as mouse clicks and scrolls (Kosala and Blockeel, 2000). Web usage mining includes the creation of user profiles, as well as analysing user access patterns and navigation paths.

Prior to applying data mining techniques, it is essential to understand the dataset. This is typically done by creating multiple summary reports and, if possible, using visual representations. Before writing programs for analyzing Web access logs, one may want to consider one of the analysis tools already available. These analysis tools may provide answers to most questions regarding the usage of Web sites. The list below provides the freeware and open source Web access analysis tools listed on an Open Directory Web site (http://dmoz.org/). In addition to freeware and open source tools, the listing of commercial tools can also be found on the Open Directory site. This section provides a discussion on how to obtain summary reports, visualization of aggregate clickstream, as well as individual user sessions from Web access logs.

- Analog www.analog.cx
- AWStats awstats.sourceforge.net
- BBClone bbclone.de
- The Big Brother Log Analyzer bbla.sourceforge.net
- BlibbleBlobble LogAnalyser www.blibbleblobble.co.uk/Downloads/LogAnalyser
- Dailystats www.perlfect.com/freescripts/dailystats
- GeoIP www.maxmind.com/geoip
- High Speed Merging ww.whurst.net/programming/hHSM/index.php
- HitsLog Script www.irnis.net/soft/hitslog
- Http-Analyze www.http-analyze.org
- Kraken Reports www.krakenreports.com
- Logfile www.ratrobot.com/programming/shell
- LogFile Analyse www.jan-winkler.de/dev
- LogReport Foundation logreport.org
- MagicStats www.nondot.org/MagicStats
- Modlogan www.modlogan.org
- NedStat www.nedstat.com
- Pathalizer pathalizer.bzzt.net
- phpOpenTracker www.phpopentracker.de
- PowerPhlogger pphlogger.phpee.com
- RCounter rcounter.noonet.ru
- Realtracker Web site Statistics free.realtracker.com
- Relax ktmatu.com/software/relax

- Report Magic for Analog www.reportmagic.com
- RobotStats www.robotstats.com/en
- Sevink Internet Advertising www.sevink-2.demon.nl
- Sherlog sherlog.europeanservers.net
- Snowhare's Utilities www.nihongo.org/snowhare/utilities
- Superstat www.serversolved.com/superstat
- VISITaTOR- a free Web mining tool visitator.fh54.de
- Visitors www.hping.org/visitors
- WebLog awsd.com/scripts/Weblog
- Webtrax Help www.multicians.org/thvv/Webtrax-help.html
- W3Perl www.w3perl.com/softs
- Wwwstat www.ics.uci.edu/pub/Websoft/wwwstat
- ZoomStats zoomstats.sourceforge.net

One of the more popular analysis tools is called Analog (www.Analog.cx). Analog analyses Web access log files and produces summary reports that can be used to enhance the content of the Web site. The information provided by Analog includes the frequently accessed pages on a Web site and the search terms used to find these pages. Analog also provides temporal analysis that can be used to ensure sufficient bandwidth for the users. The summary reports provided by Analog, however, are of limited use for link analysis. In this section, we will review two software packages, Pathalizer and StatViz, which allow us to visualize the aggregate link usage and track individual user sessions.

We will illustrate the usage of these analysis tools using data obtained from the Web access logs of an introductory first year course in Computing Science at Saint Mary's University over a sixteen-week period. The initial number of students in the course was 180. Over the course of the semester, this number was reduced to 130-140 students. Certain areas of the Web site were protected, meaning users could only access them using their IDs and passwords. The users' activities in the restricted parts of the Web site consisted of submitting profiles, changing passwords, submitting assignments, viewing the submissions, accessing the discussion board, and viewing current class marks. The public portion of the Web site consisted of viewing the course information, lab manual, class-notes, class assignments, and lab assignments. If the users only accessed the public Web site, their IDs would be unknown. To protect user privacy, the usernames for the rest of the entries were changed to "user". The dataset and its analysis using Analog can be found in Akerkar and Lingras (2007). In this chapter, we will discuss visual representations of the students' link usage and navigation paths.

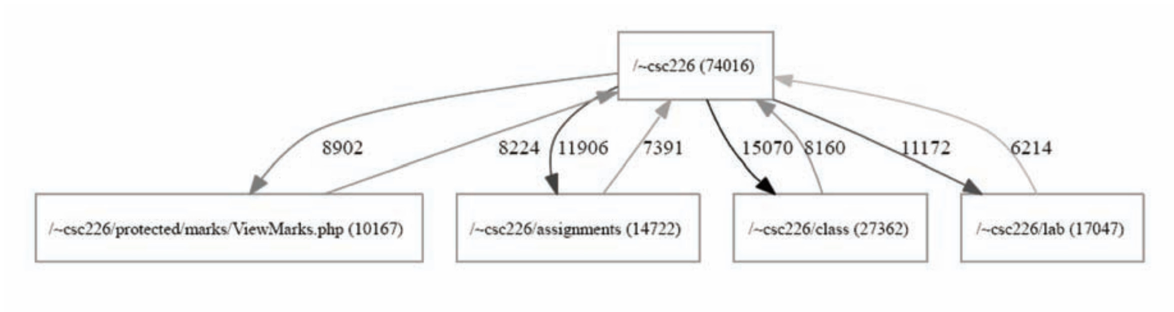
Analysis of Clickstreams: Visually Studying Navigation Paths

Pathalizer (2006) is a visualization tool that shows the most frequent paths taken by users when browsing a Web site. This information can be useful for improving the navigation within a Web site. In conjunction with the summarization of Web logs, the visual representation of navigation can also be used to determine which parts of the site need most attention.

Figure 1 shows a graph drawn by Pathalizer for the top eight traversed links in our sample Web log. Every node in Figure 1 is a page. The path/URL for the page and the number of hits on the page are listed within the boxes (or nodes). Every arrow represents a user visiting those two pages in succession. The width and the number associated with the arrow represent the number of times that path was taken (i.e., the thicker the arrow, the bigger the frequency).

Figure 1 tells us that one or more visitors directly went to all five pages shown. The homepage for the course was the most frequently visited page, and the link from the homepage to the class directory was the most frequently taken.

Figure 1. Top eight traversed links on the course Web site



The link from the homepage to the assignment page was the second most popular path, followed by the link from the homepage to the lab folder. This knowledge of popular pages and links provides a suggested hyperlink structure by telling us that:

- Every page should have a link to the homepage, and
- Links to the class folder, assignment, and lab folder should be prominently displayed on the homepage.

The specification of eight links to Pathalizer only gave us a bird eye view of the Web site. If we wish to have a more detailed analysis, we would need to increase the number of edges in the filter section of the application. However, there are problems with adding more edges in the analysis. As the number of edges increases, it is difficult to read the graph because of the large amount of information, as well as the smaller text size. More details about aggregate link usage can be found in Akerkar and Lingras (2007).

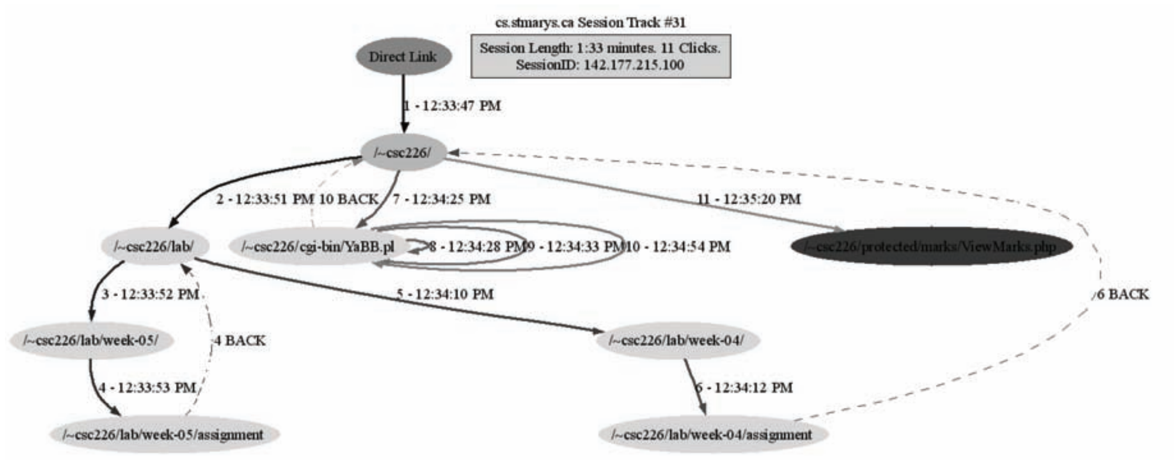
Visualizing Individual User Sessions

In the previous section, we looked at aggregate Web usage analysis. In this section, we will look at individual user interaction with a Web site. Since a Web site generally has a large number of visitors, it is practically impossible to analyze each one in great detail. However, one can randomly pick sessions of various sizes to study how individual visitors traverse through a Web site. Such an analysis should always be a precursor to the actual Web usage or data mining. Sometimes, such a visualization process may also be conducted after uncovering interesting navigational patterns from the data mining exercise. For example, if an e-Commerce Web site experiences a category of users aborting their sale, the site manager may pick a session from a list of such pre-emptive users and study the visual representation of their navigation pattern.

StatViz (2006) can be used to track the movements within an individual session stored in a logfile, from page to page. A session is simply associated with an IP address, which means all requests from a given computer are considered to be for the same session. This is not a good assumption, especially for a public computer. The session track reporter will produce one graph per session. Each graph is designed to give a good sampling of how visitors move around the site.

Tracking individual sessions is a computationally intensive activity. Moreover, it is difficult to study all the user sessions individually. Therefore, we selected a 1.25 hour snapshot from our classlog.txt. The time period that was chosen was from 11:30 am to 12:45 pm on a Tuesday (October 9, 2001). Usually, this time reports fairly intensive activity on the course Web site, as it follows the class and there are two labs

Figure 2. Tracking an individual Web session using StatViz



scheduled in parallel during that time. We picked three sessions of different lengths (Akerkar and Lingras, 2007). Due to space limitations, we will only show one session in this chapter, which is represented in Figure 2.

A graphical display of individual session tracks will allow us to understand how the students successfully or unsuccessfully navigate through the Web site. Studying these graphs will not only help us understand the information needs of our visitors, but also provide insight into how we could better present information on the Web site to facilitate easier navigation. Let us try to interpret the StatViz graph from Figure 2.

- The graph shows movement through the Web site as links from one page to another.
- Each node is a Web page; each solid line is a "click" from one page to the other as indicated by the arrow.
- Each line has a number next to it representing the number of that particular "click" in the session track, i.e. the sequence order. The time of the "click" is also shown next to the number.
- In some cases, there is also a dashed line with the same number as another click. These "BACK" links indicate that the visitor went back to that page using the back button before proceeding.
- Pages that are not on our site (external referrers) are shown as brown ovals.
- While Figure 2 is depicted in black and white in this chapter, the software colours the "entry" page green and the "exit" page red. If the entry and exit pages are the same, that page will be red.

Figure 2 shows an off-campus session which is relatively brief, lasting a little over 1.5 minutes. The IP number starting with 142.177 tells us that the user comes from an Internet service provided by the local phone company. Again, the user seemed to know what he or she was looking for. The user either used a bookmark or typed the URL directly to get to the entry page (green colored) /~csc226/. The user then proceeded with three clicks to look at the assignment for week-05, came back to the lab folder with the back button, and used two clicks to look at the assignment from previous week (week-04). The user then used the back button to get to the course homepage. Clicks 7-10 were used to browse the bulletin board. Finally, the user checked their marks and then exited the site. The user also knew the structure of the site very well.

This is why he or she managed to conduct three separate activities in a relatively short period of time: view two lab assignments, browse bulletin board, and view marks. This student's individual session track shows a successful navigation and suggests that the data on the Web site is displayed effectively. In order to come to this conclusion though, more individual sessions would have to be analysed.

In this section, we discussed data visualization techniques that can be used to understand the hyperlink usage from a Web log. Once we have some understanding of the overall navigation patterns on a Web site, we can use data mining techniques to look for interesting patterns.

DATA MINING TO DISCOVER NAVIGATION PATTERNS

The data in Web access logs are intrinsically sequential. We used data visualization techniques to look at aggregate navigation using Pathalizer, as well as navigation in individual sessions using StatViz. Applying data mining techniques to analyze sequences of Web requests is an important area of research (Cadez, *et al.*, 2000; Iváncsy and Vajk, 2006). Many of the techniques involved are at an experimental stage and contain sophisticated mathematical analysis.

Classification and Association Modeling of Web Navigation

Classification is one of the data mining techniques that can contribute to Web usage mining. Web personalization (Dai and Mobasher, 2003) is a common application of classification integrated in Web usage mining. The interest in Web personalization can be traced back to the Firefly system (obsolete URL: www.firefly.com), which was used to suggest music CDs that match the user's interests. The user's interests were determined by analyzing their navigation through the site. Similar attempts can also be seen on Amazon.com. When a user requests information about a book, the system provides a list of additional books. The list consists of books purchased by people who bought the same book that interests the user. Attempts at Web personalization are increasing at a rapid rate. The greater interest is also leading to more formal frameworks for Web personalization (Perkowitz and Etzioni, 1997). Joachim, *et al.* (1995) propose the use of a tour guide approach. Perkowitz and Etzioni (1997) attempt to formalize the concept of adaptive Web sites. Adaptive Web sites are defined as those that automatically improve their organization and presentation by learning from visitor access patterns. Perkowitz and Etzioni suggest that much of the earlier work has been focused on fairly simple adaptations such as automatically creating shortcuts on the Web site, and customizing the site to suit the needs of each individual user. They propose the use of sophisticated adaptations of Web sites to the users' needs, and aggregation of information gleaned from the user population to improve the navigation for a large number of users.

Baglioni, *et al.* (2003) used the classification algorithm C4.5 to develop a model to predict whether a user might be interested in visiting a section of a Web site based on the sections the user has already visited. The knowledge of the sections of interest will make it possible to create a personalized view of the Web site by using an on-the-fly menu or page reorganization.

There is a subtle difference between the classification model used by Baglioni, *et al.* (2003) and association mining. If one were to apply association mining, we would find pairs of channels that are accessed during the same session. The model developed by Baglioni, *et al.*, on the other hand, takes a list of channels of user interest and predicts the likelihood of the user accessing a given channel. The association mining conducted by Batista and Silva (2002) addressed essentially the same question as the one answered by the classification exercise reported in Baglioni, *et* *al.*, i.e. which category of articles are requested by the same visitor. Batista and Silva's approach falls into the classical *Market Basket Analysis* problem (Berry and Linoff, 1997).

The aim of market basket analysis is to find groups of items that are frequently referred together. In this case involving an online news site, a transaction is the Web request and the item is the news section from which the article can be found. Groups of items that occur frequently together in the same visit are called frequent itemsets (Agrawal and Srikant, 1994). A typical association mining process needs guidance for restricting the search space, usually in terms of a support threshold. The algorithms then find the itemsets that satisfy this minimum support threshold. Batista and Silva defined weak associations as those below 5% of the total number of occurrences, and heavy associations as those above 10%. The association mining results for the online news site showed strong associations between pairs such as (Politics, Society), (Politics, International News), and (Society, Local News).

Sequence Pattern Analysis of Web Logs

Cadez, *et al.* (2000) presented msnbc.com (2000) anonymous Web data that can be downloaded from: http://kdd.ics.uci.edu/databases/msnbc/msnbc.html. The data comes from msnbc.com's Web access logs and msn.com's news portions for the twenty-four hour period of September, 28, 1999. There were a total of 989,818 user sessions. The data is anonymized, so we have no knowledge of the login details of the users. We are also limited to knowing the category of the Web page that was requested by the user, and not the name of the actual page. The reporting of categories of pages as opposed to the actual pages in fact simplifies our job. There are anywhere from 10 to 5,000 pages per category and it would be difficult to keep track of each one of these pages. As with any other Web access logs, any page request served via a caching mechanism could not be recorded in the data.

One of the most useful pieces of information in the Web access logs is the sequences in which pages are accessed by users. Each sequence in the dataset corresponds to a user's Web request. The average length of the sequences is 5.7. Sequences can be used to provide appropriate links to simplify the navigation. One can do a frequency analysis of all the category pairs, such as (1,1), (1,2), (1,3),…,(17,1), (17,2),…,(17,3). In total, there are $17 \times 17 = 289$ pairs. The sequence of category numbers in a pair is important for two reasons:

- The links are always between a pair of pages. Thus, knowing which pages are requested from a given page is the most relevant information needed in order to determine the navigational links.
- The pairs of sequences will have the highest frequency. For example, a sequence (i,j,k) cannot have a higher frequency than either of the pairs, (i,j) or (j,k).

Analysis of the paired sequences reveals interesting information. For example, 138,441 users first accessed a page from category 1 followed by another page in the same category, i.e. 1. Conversely, only two users first read an article from category 17 and then moved to a page from category 16. The pairs that stay within the same category – such as the top four sequences, (1,1), (2,2), (14,14), (8,8) – tend to have higher frequencies. This seems reasonable because once a user starts reading an article from one category he or she is likely to access another article from the same category. Usually, such sequences would be well served by the Web structure, since the Web site is likely to provide reasonable navigation between articles within the same category. The pairs where one of the categories is 1 (corresponding to front page) tends to have higher frequencies. (The fifth highest frequency is for the pair (1,2).)

This sequence probably corresponds to users first coming to the front page and then following links to other categories. These sequences will also be better served by the Web site, because the front page will have links to other categories.

The pair (6,7) with a frequency of 25,106 is the first pair with two distinct categories, neither of which is front page. Category 6 is “on-air” and category 7 is “misc”. The high frequency for this pair suggests that the Web site should provide easy navigation from “on-air” to “misc”. On the other end of the spectrum, analysis shows very little movement between some of the categories. There are three possible explanations for such a lack of migration:

- These categories have limited contents, and hence very few visitors. For example, only 2,032 users accessed a page from category 16 (bbs).
- These categories attract different types of readers.
- The link structure does not make it possible to easily navigate between these categories. The Web administrator may consider improving the navigational structure.

The aforementioned discussion shows how one can interpret paired sequences of Web usage to study the Web structure. One can extend such an analysis to longer sequences. There are a total of 17 × 17 × 17 = 4913 triplets. Similar to the paired sequences, the triplets where all three pages are from the same category tend to have higher frequencies. The triplets where the three categories are not distinct do not provide any additional knowledge about user behaviour than the one gathered from paired sequences. Therefore, we will focus only on triplets where all three categories are distinct. The first such pair is (1,7,4), which corresponds to frontpage-misc-local. This information could probably be used to put the links to categories 7 and 4 next to each other on pages from categories 1.

One could extend the analysis to longer sequences. However, it should be noted that 635 of the 4913 triplets had a frequency of 0. The longer sequences with distinct categories would have smaller and smaller frequencies, since the average length of sequences is only 5.7. Another possible analysis could include compressing the sequences by replacing multiple sequential articles from the same categories, and then analyzing the resulting sequences. For example, a sequence “3 2 2 4 2 2 2 3 3” could be represented as “3 2 4 2 3” that shows how a user migrates from one category to next.

The msnbc.com data (2000) has been subject to more sophisticated analysis by a number of researchers. Cadez, *et al.* (2000) partition users from msnbc.com into clusters based on similarity in their navigation paths through the site. The clustering approach is model-based (as opposed to sequence of length based approaches we have seen thus far). It partitions users according to the order in which they request Web pages. Cadez, *et al.* use a mixture of first-order Markov models using the Expectation-Maximization algorithm. They display aggregate paths for each cluster. As expected from our earlier analysis, the larger clusters tend to have a navigation path that seems to navigate within a given category.

Iváncsy and Vajk (2006) show how the automata theory can be used for discovering frequent Web navigation patterns from the msnbc.com data. The SM-Tree algorithm discovers the frequent page sequences using finite state machines. The navigation patterns are rarely sequential. Sometimes users use the back button to go back one level and then traverse to the next page. Such navigation is best represented using a tree. Iváncsy and Vajk show how the PDTree algorithm based on pushdown automaton can be used to identify the tree-like Web navigation patterns.

The following section describes how Lingras and Lingras (2007) use the sequence analysis of Web page accesses to create weighted directed graphs. The information contained in such graphs can be used to improve the Web surfing experience.

CREATING AN OPTIMAL HYPERLINK STRUCTURE USING GRAPH THEORY

In this section, we will use the MSNBC data of Web access patterns (2000) to develop a graph theoretic model of Web usage sequences for optimal navigation within the Web site (Lingras and Lingras, 2007). An ideal navigational structure would limit the number of outgoing links from a page, provide an adequate number of incoming links to a page, and also ensure that most transfers between pages can happen within two to three clicks. Two important granular concepts, namely neighborhood systems (Krishnan and Raghavachari, 2001) and rough set theory (Lin, 2003) are used to propose concepts of primary, secondary, and tertiary neighborhood. These neighborhoods are used to formalize the aforementioned requirements of an ideal navigational structure. A modification of the minimal spanning tree algorithm is also shown to be useful for developing an optimal hyperlink structure.

Graph of Web Usage Sequences

A graph G is defined using a pair of set (V, E), where V is a set of vertices and E is a set of edges. For directed graphs, edges are ordered pairs, i.e. $E = \{(u, v) \mid u, v \in V\}$. For undirected graphs, the edges are sets of vertices with a cardinality of two, i.e. $E = \{\{u, v\} \mid u, v \in V\}$. Let us look at a more specific example of a Web graph. Let $V = \{home, cat_1, cat_2, cat_3, page_1, page_2, \bullet\bullet\bullet, page_9\}$ be the Web pages on a Web site. If we have links from every page to every other page in V, we can reach any page from any other page with a single click. However, such a fully connected graph will mean that every page must have twelve links. A large number of links on a page may overwhelm a user, especially on a site that has several thousand pages. The user would also be more likely to miss certain links, as a large number of links would lead to clutter. However, if a page has reasonably small number of links, the user is more likely to notice all of them. Reducing the number of links, though, will make it necessary for a user to follow multiple links in order to reach a desired page. These two competing goals can be defined using the concepts of primary, secondary, and tertiary neighborhoods as follows.

The primary neighborhood for a Web page p, $N_1(p)$, is defined as the set of Web pages that can be reached using a single click. Mathematically, we can write it as: $N_1(p) = \{v \in V \mid (p, v) \in E\}$. The secondary neighborhood, $N_2(p)$, for a Web page is defined as the set of Web pages that can be reached using two or fewer clicks. The tertiary neighborhood, $N_3(p)$, for a Web page is defined as the set of Web pages that can be reached using three or fewer clicks. It is possible to define neighborhood for higher degrees such as $N_4(p)$ or $N_5(p)$ as well.

The neighborhood concept used here is borrowed from the neighborhood systems proposed by Lin (2003). The primary neighborhood is similar to the lower bound in rough set theory (Pawlak, 1992). The higher degree neighborhoods can be viewed as a generalization of the upper bound.

It can be easily seen that secondary neighborhood of a page is a union of the primary neighborhood of the page with primary neighborhoods of all the pages in its primary neighborhood, such as:

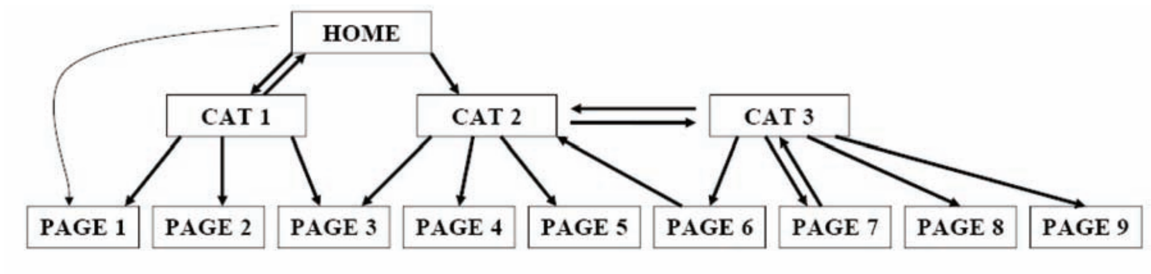
$$N_2(p) = N_1(p) \cup \bigcup_{v \in N_1(p)} N_1(v)$$

Similarly,

$$N_3(p) = N_2(p) \cup \bigcup_{v \in N_2(p)} N_1(v)$$

Our goal is to:

Figure 3. Example of an optimal navigation structure



i. Restrict the size of $N_1(p)$ to make sure that the page is not cluttered with too many hyperlinks.
ii. Ensure that for every Web page access (p, v),$v \in N_3(p)$. (We can use higher degree neighborhoods such as $N_4(p)$ or $N_5(p)$ if it is appropriate.)

Figure 3 shows a possibly optimal navigational structure for our example set of Web pages. As we can see, the cardinality of primary neighborhood is less than or equal to three. Moreover, each page has at least two incoming hyperlinks. The following are some of the primary and secondary neighborhoods:

$N_1(home) = \{cat_1, cat_2, page_1\}$

$N_1(cat_1) = \{page_1, page_2, page_3, home\}$

$N_1(cat_2) = \{page_3, page_4, page_5, cat_3\}$

$N_1(page_1) = \varnothing$.

$$\begin{aligned} N_2(home) &= N_1(home) \cup N_1(cat_1) \cup N_1(cat_2) \cup N_1(page_1) \\ &= \{cat_1,\ cat_2\ \ page_1\ \ page_2\ \ page_3\ home,\ cat_3,\ page_4,\ page_5\} \end{aligned}$$

Since we do not know the Web page access sequences, we cannot verify that the tertiary neighborhoods will include all the Web page access sequences.

In order to develop an algorithm that achieves the aforementioned goal, we will be using a weighted directed graph, where a weight is associated with each edge. Figure 1 shows an example of a weighted Web graph for a course offered at Saint Mary's University.

Usually, the weights represent a cost associated with the edge. Therefore, most graph theoretical algorithms are designed to minimize the sums of weights. Thus, we will have to take an inverse of the number of Web page accesses as our weights, as described below, for the MSNBC data. There will be an edge from every category to every other category. Since the weights represent a cost associated with that edge, we use the following formula:

$$w_{(u,v)} = \frac{\text{Average frequency of a paired sequence}}{frequency_{(u,v)}}$$

It is easily seen that the smaller the weight the more popular the edge.

We can use the concept of a spanning tree to find the smallest number of links for a Web site. A spanning tree for a connected undirected graph is defined as a subgraph with the same number of vertices as the original graph, with the minimum number of edges to keep all the vertices connected to each other. A minimum spanning tree is a spanning tree with smallest sum of edge weights.

We will modify the notion of the spanning tree for our Web graph based on the assumption that every well-designed Web site has a link to "Home" from all the pages. Therefore, we do not need to consider any edges to the root in our analysis. That means, as long as there is a path from the root to

Figure 4. Algorithm to find an optimal hyperlink structure

1) Definitions:
   - Let $G = (V, E)$ be the original Web graph and $G_t = (V_t, E_t)$ be its minimum spanning Web graph.
   - Let V_{temp} be a temporary set of vertices.
   - Let C be the set of candidate edges.
   - Let $unused$ be the set of unused edges.
   - Let $desiredInDegree$ be the desired number of incoming edges for each vertex.
   - Let $inDegree$ be the current number of incoming edges for each vertex.
2) Initialization:
   - $V_{temp} = \{root\}$.
   - $C = \{(root, x) | (root, x) \in E\}$.
   - $unused = \emptyset$.
   - $inDegree = 0$.
3) While New Vertex is added to V_{temp} and $V_{temp} \neq V$
   - Remove the edge (u, v) with smallest weight from C
   - If $v \notin V_{temp}$ Then
     - Add v to V_{temp}.
     - Add all the edges (v, y) from E to C.
     - Add (u, v) to E_t.
     - Add a penalty δ to all the edges (u, x) that are in C and $unused$.
   - Else
     - Add (u, v) to $unused$.
4) If $inDegree == 1$ Then $V_t = V_{temp}$
5) If $inDegree < desiredInDegree$ Then
   - Increment $inDegree$
   - $V_t = \{root\}$
   - Add all the edges from $unused$ to C.
   - $unused = \emptyset$.
   - go to Step 3.

every page, all the pages are connected to each other. Hence, we can define a spanning Web tree as a subgraph of a Web graph such that:

- There is a designated vertex called root that corresponds to the homepage or frontpage of the site, which has no incoming edges,
- all the vertices other than the root have exactly one incoming edge and the weights of the edges are the minimum.

Preliminary analysis of the Web page access sequences from MSNBC suggests that one will end up with the Web tree that is in the form of a star with most of the Web pages connected directly to the homepage. This means that there will be too many links on the homepage, which, as discussed before, would not be advantageous. To avoid too many links on a page, we can add a penalty to the weights of the remaining edges from a vertex, whenever an edge from the vertex is added to the spanning Web tree. The penalty helps in restricting the size of the primary neighborhood for each page. In order to increase the chances of including all the Web page access sequences in tertiary neighborhood, we deviate from the minimum spanning Web tree by allowing multiple incoming edges for each node. The resulting Web graphs will result in larger tertiary neighborhoods. The algorithm is shown in Figure 4.

If *desiredInDegree* = 1, the algorithm given in

Figure 5. Suggested hyperlink structure for MSNBC

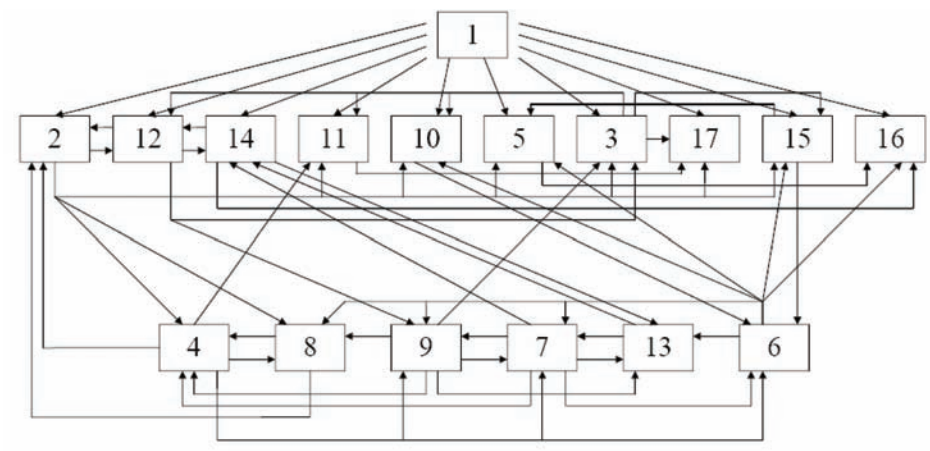


Figure 4 will result in a minimum spanning Web tree. While a minimum spanning Web tree provides navigation from each vertex to every other vertex by going through the root using the link to "Home", one may want to provide additional navigation paths to these nodes. We can add these additional paths to the hyperlink structure by specifying larger values for *desiredInDegree*. The algorithm will add more incoming links to each vertex, if they exist in the original graph.

Analysis of the Graph of Web Usage Sequences

We applied the algorithm shown in Figure 4 to the frequency of transitions between two categories from msnbc.com. We experimented with different values of *desiredInDegree* and *delta* (the penalty). We found that *desiredInDegree* = 4 and *delta* = 0.1 gave us a reasonable hyperlink structure, as shown in Figure 5.

The *desiredInDegree* of 4 did result in all vertices having four incoming edges. However, the algorithm does allow for less than four incoming edges if four incoming edges did not in fact exist.

In the secondary neighborhood (with the root or "Home" being the primary neighborhood), the amount of outgoing edges varies. For example, Category 2 has nine outgoing edges, while Category 11 only has one. One can conclude that even greater importance should be placed on Category 2.

The tertiary neighborhood (with the root or "Home" being the primary neighborhood) also contains an important vertex, Category 6, which has outgoing edges and points to eight vertices. The large amount of outgoing edges in Category 2 and Category 6 may be a reason why a greater penalty should be put into place. A greater penalty to the weights would decrease the amount of links and thus clutter on a page.

The graph contains many symmetric links, i.e. it contains many pairs having an incoming and outgoing edge to each other. Examples include: (7, 13) (13, 7); (2, 12) (12, 2); (12, 14) (14, 12); (9, 7) (7, 9); and (4, 8) (8, 4). Clearly, these categories belong to the same logical group.

It is interesting to note that while most vertices from the tertiary neighborhood have at least one incoming edge from the secondary neighborhood, Category 7 only has incoming edges from other tertiary neighborhood members. This shows that Category 7 may generate a highly specific target audience. But, it also runs the risk of easily being "lost" among all the other categories.

CONCLUSION AND FUTURE RESEARCH AND DEVELOPMENT

In this chapter, we explored various ways to analyse and improve a Web site's hyperlink structure. Web access logs were used to study user navigation patterns, which can be used to create a hyperlink structure that provides better navigation. We looked at various tools, such as Pathalizer and StatViz, to visualize the Web navigation. We also discovered the various ways data mining techniques could be applied, including gathering more knowledge about Web usage and creating a more personalized Web experience. Finally, we discussed how the minimum spanning tree algorithm can be extended to create a Web graph that provides a possibly optimal navigational structure based on Web usage patterns. An ideal navigational structure would provide easier navigation by limiting the number of outgoing links from a page while still providing an adequate number of incoming links, which would ensure most transfers between pages happening within two or three clicks.

Creating dynamic Web pages based on predicted user needs is one of today's most popular areas of research. These attempts are usually referred to as the adaptive Web; it assumes that one can quickly understand what the user wants and quickly create an appropriate navigation page. In some cases, it is possible to categorize a Web user with their initial navigational patterns. However, this first judgement does not always reflect the user's developed navigation as they become more comfortable with the Web site or if their interests change. This chapter describes a restructuring of the current static Web structure on a regular basis. Such a restructured Web site will cater to majority of Web visitors, without having to make a quick judgement about their navigations intentions. There is currently limited research in this area. The authors believe that more Web site designers will start bringing the knowledge of Web usage in determining Web structure. The algorithm presented in this chapter is an initial step in this direction. A theoretical framework such as the one described in (Lingras, 2007) will play a key role in further development.

REFERENCES

Akerkar, R., & Lingras, P. (2007). *Building an Intelligent Web. Theory and Practice*, Jones and Bartlett, Boston.

Baglioni, M., Ferrara, U., Romei, A., Ruggieri, S., & Turini, F. (2003). Preprocessing and Mining Web Log Data for Web Personalization, *Proceedings of 8th Natl' Conf. of the Italian Association for Artificial Intelligence* (AI-IA 2003).

Batista, P., & Silva, M. J. (2002). Mining Web Access Logs of an On-line Newspaper, *Proceedings of 2nd International Workshop on Recommendation and Personalization in eCommerce* (RPeC'02).

Cadez, I., Heckerman, D., Meek, C., Smyth, P., & White, S. (2000). Model-Based Clustering and Visualization of Navigation Patterns on a Web Site, *Technical Report, MSR-TR-00-18, Microsoft Research*, Microsoft Corporation, One Microsoft Way, Redmond, Washington.

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining. Information and pattern discovery on the World Wide Web, *Proceedings of International Conference on Tools with Artificial Intelligence*, 558–567.

Dai, H., & Mobasher, B. (2003). A road map to more effective Web personalization. Integrating domain knowledge with Web usage mining . *Proceedings of the International Conference on Internet Computing*, *2003*(IC03).

Gabow, H., Galil, Z., Spencer, T., & Tarjan, R. (1986). Efficient algorithms for finding minimum spanning trees in undirected and directed graphs . *COMBINATORICA*, *6*, 109–122. doi:10.1007/BF02579168

Iváncsy, R., & Vajk, I. (2006). Frequent Pattern Mining in Web Log Data. *Acta Polytechnica Hungarica*, *3*(1), 77–90.

Joachims, T., Armstrong, R., Freitag, D., & Mitchell, T. (1995). Webwatcher. A learning apprentice for the world wide Web, *Proceedings of AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*.
msnbc.com (2000). msnbc.com Anonymous Web Data. Retrieved from *http.//kdd.ics.uci.edu/databases/msnbc/msnbc.html*

Kosla, R., & Blockeel, H. (2000). Web mining research. A survey . *SIG KDD Explorations*, *2*(15), 1–15. doi:10.1145/360402.360406

Krishnan, R., & Raghavachari, B. (2001). The Directed Minimum-Degree Spanning Tree Problem, *Proc. FSTTCS 2001*, *LNCS 2245*, Springer-Verlag Berlin Heidelberg, 232-243.

Lin, T. Y. (2003). Neighborhood Systems. Mathematical Models of Information Granulations, *Proc. 2003 IEEE International Conference on Systems*, *Man and Cybernetics*, Washington, D.C., USA.

Lingras, P. (2007). Neighbourhood Sets based on Web Usage Mining, *Proceedings of 2007 North American Fuzzy Information Processing Society Conference*, 659-664.

Lingras, P., & Lingras, R. (2007). Adaptive hyperlinks using page access sequences and minimum spanning trees, *Proceedings of 2007 IEEE International Conference on Fuzzy Systems*, 529-534.

Malkevitch, J. (2006). Trees: A Mathematical Tool for All Seasons. Retrieved from http://www.ams.org/featurecolumn/archive/trees.html.

Pathalizer (2006a). Project details for Pathalizer. Retrieved from http://freshmeat.net/projects/pathalizer/

Pathalizer (2006b). Pathalizer download. Retrieved from http://pathalizer.sourceforge.net/

Pawlak, Z. (1992). Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers.

Perkowitz, M., & Etzioni, O. (1997). Adaptive Web sites: an AI challenge, *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*.

Ramadhan, H., Hatem, M., Al-Khanjri, Z., & Kutti, S. (2005). A Classification of Techniques for Web Usage Analysis . *Journal of Computer Science*, *1*(3), 413–418.

StatViz. (2006). StatViz - Graphical Clickstream/Path Analysis of Web Traffic, http://statviz.sourceforge.net/

Turner, S. (2006). Analog, http://www.Analog.cx

KEY TERMS AND DEFINITIONS

Web Usage Mining: Extracting useful information from the Web usage statistics.

Web Structure Mining: Extracting useful information from the hyperlinked Web structure.

Web Graph: Graph theoretical representation

This work was previously published in the Handbook of Research on Text and Web Mining Technologies, edited by M. Song and Y. Wu, pp. 386-400, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 7.12
Search Engine-Based Web Information Extraction

Gijs Geleijnse
Philips Research, The Netherlands

Jan Korst
Philips Research, The Netherlands

ABSTRACT

In this chapter we discuss approaches to find, extract, and structure information from natural language texts on the Web. Such structured information can be expressed and shared using the standard Semantic Web languages and hence be machine interpreted. In this chapter we focus on two tasks in Web information extraction. The first part focuses on mining facts from the Web, while in the second part, we present an approach to collect community-based meta-data. A search engine is used to retrieve potentially relevant texts. From these texts, instances and relations are extracted. The proposed approaches are illustrated using various case-studies, showing that we can reliably extract information from the Web using simple techniques.

INTRODUCTION

Suppose we are interested in *'the countries where Burger King can be found'*, *'the Dutch cities with a university of technology'* or perhaps *'the genre of the music of Miles Davis'*. For such diverse factual information needs, the World Wide Web in general and a search engine in particular can provide a solution. Experienced users of search engines are able to construct queries that are likely to access documents containing the desired information. However, current search engines retrieve Web pages, not the information itself[1]. We have to search within the search results in order to acquire the information. Moreover, we make implicit use of our knowledge (e.g. of the language and the domain), to interpret the Web pages.

DOI: 10.4018/978-1-60566-112-4.ch009

Table 1. Comparison between the Web as a corpus and 'traditional' corpora

Web Corpus	Newspaper Corpus
Redundancy. Because of the size of the Web, we can expect information to be duplicated, or formulated in various ways. If we are interested in a fact, we have to be able to identify just one of the formulations to extract it.	**No or fewer redundancy.** Especially for smaller corpora, we cannot expect that information is redundantly present.
Temporal and unreliable. The content of the Web is created over several years by numerous contributors. The data is thus unreliable and may be out-dated.	**Constant and reliable**. In corpus-based IE, it is assumed that the information in the corpus is correct and up-to-date.
Multilingual and heterogeneous. The Web is not restricted to a single language and the texts are produced by numerous authors for diverse audiences.	**Often monolingual and homogeneous.** If the author or nature (e.g. articles from the Wall Street Journal) of the corpus is known beforehand, it is easier to develop heuristics or to train named entity recognizers.
No representative annotated corpora. As no representative annotated texts are available, the Web as a corpus is currently less suited for supervised machine learning approaches.	**Annotated test corpora available.** In order to train supervised learning based named entity recognizers (NERs), test corpora are available where instances of a limited number of classes are marked within the text.
Dynamic. The contents of the Web changes continuously, results of experiments may thus also change over time.	**Static.** Experimental results are independent of time and place as the corpora are static.
Facts and opinions. As a multitude of users contribute to the Web, its contents are also suited for opinion mining.	**Facts only.** Information Extraction tasks on Newspaper corpora mainly focus on the identification of facts.

Apart from factual information, the Web is the de-facto source to gather community-based data as people with numerous backgrounds, interests and ideas contribute to the content of the Web. Hence the Web is a valuable source to extract opinions, characterizations and perceived relatedness between items.

In this chapter, the focus is on gathering and structuring information from the 'traditional' Web. This structured information can be represented (and shared) using the standard Semantic Web (SW) languages. Hence, this chapter focuses on the automatic creation of content for the SW. For simplicity, we abstract from the SW standards RDF(S)/OWL.

The Web-as-a-Corpus vs. Traditional Text Corpora

Information extraction (IE) is the task of identifying instances (or *named entities*) and relations between those instances in a collection of texts, called a text corpus.

In the nineties, the Message Understanding Conferences (MUC) focused on the recognition of named entities (such as names of persons and organizations) in a collection of texts (Chinchor, 1998). Initially, this work was mostly based on rules on the syntax and context of such named entities. For example, two capitalized words preceded by *mr.* will denote the name of a male person. As the creation of such rules is a laborious task, approaches became popular where named entities were recognized using machine learning (Mitchell, 1997), for example in (Zhou & Su, 2002; Brothwick, 1999; Finkel, Grenager, & Manning, 2005). However, such approaches typically make use of annotated training sets where instances (e.g. '*Microsoft'*) are labeled with their class *('Organization'*).

Traditional information extraction tasks focus on the identification of named entities in large text corpora such as collections newspaper articles or biomedical texts. In this chapter however, we focus on the Web as a corpus. In Table 1 the most important differences between the two can be found.

Suppose that we are interested in a list of all countries in the world with their capitals. When we extract information from a collection of newspaper articles (e.g. 3 months of the New York Times), we cannot expect all information to be present. At best, we can try to discover every country-capital combination that is expressed within the corpus. When we use the Web however, all the required information can be expected to be present. Moreover each of the combinations is likely to be expressed on various pages with multiple formulations. For example, *'Amsterdam is the capital of the Netherlands'* and *'The Netherlands and its capital Amsterdam (...)'* are different formulations of the same fact. In principle, we have to be able to interpret only one of the formulations to extract the country-capital combination. Hence, in comparison with a 'traditional' newspaper corpus, we can both set different objectives and apply different methods to extract information from the Web.

Heterogeneous vs. Homogeneous Sources

In this work we focus on unstructured natural language texts. Information extraction from structured sources is thoroughly described in for example (Chang, Kayed, Girgis, & Shaalan, 2006) and (Crescenzi & Mecca, 2004).These 'wrappers' make use of the homogeneous lay-out of large Web sites. Large Web sites such as *amazon.com* and *imdb.com* make use of a database and present automatically generated Web pages. The lay-out is uniform over the whole site, but the information changes from page to page. The performing artist, the title of the album and other catalogue data can be found on the exact same place on the page. The HTML-source of the two pages will also only differ at these places. For pages within a large Web site, a wrapper algorithm can be created the information of interest from an arbitrary page within the site. Wrappers are relatively simple and time efficient. However, they are Web site and thus domain dependent. Instead we focus on information from arbitrary Web sites in this chapter.

Fact Mining

The first part of this chapter discusses a method to extract factual information from the Web. To formalize the concept *information* we define an initial ontology on the domain of interest (e.g. movies, literature, hamburger restaurants). Given this initial ontology, we populate it with extracted knowledge. For example, if we consider an ontology with the classes *Movie* and *Actor*, we can populate the ontology by finding instances of these classes. Hence, if we identify the terms *'Top Gun'*, *'The Godfather'* as movie titles and the terms *'Tom Cruise'* and *'Marlon Brando'* as actors we can add these terms as instances to the ontology. Moreover if acts-in is a relation between Movie and Actor, then the challenge is to discover the instances of that relation from texts on the Web, e.g. (*Marlon Brando, The Godfather*).

In Section 2 of this chapter, we present an algorithm that—given a domain of interest—extracts, structures and combines information from the Web. With structured information available, we can easily find the information we are interested in. The extracted information can, e.g. be used by recommender systems to acquire additional metadata. This metadata can be used to make meaningful recommendations for music or TV programs. For example, suppose a user has expressed a preference for TV programs relating to Italy. The recommender system will be able to recognize regions as *Tuscany* and *Veneto* and cities as *Milano* and *Florence* using extracted information. Occurrences of such terms in a program guide description will mark a program as relevant. Likewise, if the user has expressed a preference for TV programs relating to photography the system will be able to recognize the names of famous photographers as *Cartier-Bresson* and *Moholy-Nagy*.

Community-Based Knowledge Mining

The Web is not only a well-suited text corpus to mine factual information. As a large community of users contributes to the contents of the Web, it can also be used to mine more subjective knowledge. For example, we call Paul Gauguin a post-impressionist and to be similar to Vincent van Gogh, *Christina Aguilera* a pop artist similar to *Britney Spears*. Such qualifications may not all be facts, but are thoughts shared by a large community.

In the second part of the chapter we focus on methods to automatically find such internet community-based meta-data. On the one hand we classify instances (e.g. pop artists) into categories and on the other hand identifying a distance matrix of related instances. The information found can be used to create an automated folksonomy: a knowledge base where items are tagged using implicit input from multiple users.

In restricted domains (e.g. Movies) in fact mining, the use of wrappers may be well usable. The Internet Movie Database[2] for example is a reliable, semi-structured source to extract data on movies. When we are interested in subjective data based on opinions of the Web community however, we cannot restrict ourselves to a single source. We combine data from multiple Web sites, and thus multiple contributors, to characterize instances. We can however use semi-structured data from social Websites such as *last.fm* as a benchmark on restricted domains like music (Geleijnse, Schedl, & Knees, 2007).

Related Work

Information extraction and ontology constructing are two closely related fields. For reliable information extraction, we need background information, e.g. an ontology. On the other hand, we need information extraction to generate broad and highly usable ontologies. An good overview on state-of-the-art ontology learning from text can be found in (Buitelaar, Cimiano, & Magnini, 2005).

Early work on relation identification from the Web can be found in (Brin, 1998). Brin describes a Web site-dependent system to identify hypertext patterns that express a given relation. For each Web site, such patterns are learned and explored to identify instances that are similarly related. SnowBall (Agichtein & Gravano, 2000) is a successor of Brin's system with an embedded named-entity recognizer. The idea of extracting relations using patterns is similar to one of the methods presented here. However, in Snowball the relations gathered are not evaluated.

KnowItAll is a hybrid named-entity extraction system (Etzioni et al., 2005) that finds lists of instances of a given class from the Web using a search engine. It combines hyponym patterns (Hearst, 1992) and learned patterns for instances of the class to identify and extract named-entities. Moreover, it uses adaptive wrapper algorithms (Crescenzi & Mecca, 2004) to extract information from html markup such as tables. Contrary to our method, it does not use instances to formulate queries. In (Downey, Etzioni, & Soderland, 2005) the information extracted by KnowItAll is evaluated using a combinatorial model based on the redundancy of information on the Web.

Cimiano and Staab (2004) describe a method to use a search engine to verify a hypothesis relation. For example, if we are interested in the 'is a' or hyponym relation and we have the instance Nile, we can use a search engine to query phrases expressing this relation (e.g. ''rivers such as the Nile'' and ''cities such as the Nile''). The number of hits to such queries is used to determine the validity of the hypothesis. Per instance, the number of queries is linear in the number of classes (e.g. city and river) considered.

In (de Boer, van Someren, & Wielinga, 2007) a number of documents on art styles are collected. Names of painters are identified within these documents. The documents are evaluated by counting the number of painters in a training set (of e.g. expressionists) that appear in the document. Painters appearing on the best ranked documents are then mapped to the style. De Boer et al. use a training set and page evaluation, where we simply observe co-occurrences.

A document based technique in artist clustering is described in (Knees, Pampalk, & Widmer, 2004). For all music artists in a given set, a number of documents is collected using a search engine. For sets of related artists a number of discriminative terms is learned. These terms are used to cluster the artists using support vector machines. The documents are obtained in a similar way in our document-based method. However, we restrict ourselves to identifying names of artists and categories in the documents.

The number of search engine *hits* for pairs of instances can be used to compute a semantic distance between the instances (Cilibrasi & Vitanyi, 2007). The nature of the relation is not identified, but the technique can for example be used to cluster related instances. In (Zadel & Fujinaga, 2004) a similar method is used to cluster artists using search engine counts. In (Schedl, Knees, & Widmer, 2005), the number of search engine hits of combinations of artists is used in clustering artists. However, the total number of hits provided by the search engine is an estimate and not always reliable (Véronis, 2006). In (Geleijnse, Korst, & de Boer, 2006) an approach is presented where one instance is queried and the resulting texts are mined for occurrences of other instances. Such an approach is not only more efficient in the number of queries, but also gives better results.

In (Pang, Lee, & Vaithyanathan, 2002; Dave & Lawrence, 2003; Kim & Hovy, 2004; Pang & Lee, 2005) methods are discussed to identify opinions on reviewed products. For example, given is a set of reviews of some flat screen television mined from the Web. The task is to assign a grade to the product or its specific features (e.g. the quality of the speakers).

The extraction of social networks using Web data is a frequently addressed topic. For example, Mori et al. (2006) use tf·idf (see (Manning & Schütze, 1999)) to identify relations between politicians and locations and (Jin, Matsuo, & Ishizuka, 2006) use inner-sentence co-occurrences of company names to identify a network of related companies.

FACT MINING

In this first part of the chapter, we will describe a method to populate an ontology, given a domain of interest. Here we only consider factual information, i.e. opinions and the like are here not taken into account but will be considered in the second part of this chapter.

Problem Definition

The Semantic Web community (Berners-Lee, Hendler, & Lassila, 2001) is providing standards for machine readable and interpretable information on the Web. The languages RDF(S) and OWL are developed for this purpose by the World Wide Web Consortium[3]. Dedicated reasoners are created for the semantic Web languages for ontology-based question-answering services. As such, these reasoners are able to provide answers to information demands like the above, given a sufficiently populated ontology.

For our purposes we define an ontology as follows:

Definitions. A reference ontology O is a 4-tuple (C, I, P,T), with:

C= $(c_0, c_1, \ldots, c_{N-1})$, an ordered set of N classes,
I = $(I_0, I_1, \ldots, I_{N-1})$, with I_j, $0 \le j < N$, the set of instances of class $c_j \in C$,
P = $(p_0, p_1, \ldots, p_{M-1})$, a set of M binary relations on the classes, with p_i: $c_{i,0} \times c_{i,1}$ $0 \le i < M$, and $c_{i,0}, c_{i,1} \in C$, and
T = $(T_0, T_1, \ldots, T_{M-1})$, is a set of instances of the

relations in P with $T_i = \{(s,o) \mid p_i(s,o)\}$ for each i, $0 \le i < M$

An *initial ontology* of O is defined as O' = (C, I', P,T'), where:

$I'_j \subseteq I_j$ for all j, $0 \le j < N$,
$T'_i \subseteq T_i$ for all i, $0 \le i < M$, and
$(s,o) \in T'_k \Rightarrow s \in I'_{k,0}\ o \in I'_{k,1}$ for some k

Popular search engines currently only give access to a limited list of possibly interesting Web pages. A user can get an idea of relevance of the pages presented by analyzing the title and the snippet presented. When a user has sent an accurate query to the search engine, the actual information required by the user can already be contained in the snippet.

We are interested whether the data in the snippets presented by a search engine is sufficient to extract information. With the definitions presented above, we formulate the information extraction problem as an ontology population problem:

Problem. Given an initial ontology O', extend O' to some O'' that maximizes the precision and/or recall.

We define precision and recall as measures of a class $c_j \in C$:

$$precision(c_j) = \frac{|I_j \cap I''_j|}{|I''_j|}$$

and

$$recall(c_j) = \frac{|I_j \cap I''_j|}{|I''_j|}$$

Similar measures can be formulated for relations p_i.

Global Outline

We choose to extract information from arbitrary Web sites. To find relevant Web sites—and thus relevant information—we use a state of the art search engine. Currently, both Google and Yahoo! allow a limited amount of automatic queries per day.

When using a search engine, we have to deal with the following restrictions.

1. The search engines return a limited number of search results per query (at most 5,000 per day using the *Yahoo! API*) .
2. We want to perform as few queries to a search engine as possible to limit the use of its services.

We therefore need accurate queries, for which we can expect the search engine to return relevant snippets. We therefore choose to use known instances in our queries to simultaneously find instances of classes and instances of relations. For example, given the instance '*Johan Cruijff'* in the initial ontology, we can use this instance in the query 'Johan Cruijff was born in' in order to retrieve a place in general and Cruijff's place of birth in particular. The place of birth, Amsterdam, can be extracted from the retrieved documents. Now, *'Amsterdam'* can be used in the query 'was born in Amsterdam' to discover other (Amsterdam-born) persons, like Willem Frederik Hermans and Paul Verhoeven.

For a given relation p_k, we thus use both an instance $I'_{k,0}$ and a natural language formulation of the relation in our queries. Subsequently, the snippets are scanned for instances of $I'_{k,1}$ and $(I'_{k,0}, I'_{k,1}) \in T'_k$. In Section 2.3 we focus on the identification of relation patterns. Section 2.4 handles the identification of instances of a class from the snippets. We combine these strategies into the ontology population algorithm as described in Section 2.5.

Identifying Effective Relation Patterns

Given is a relation p_k and a set T'_k of instances of p_k. For relation p_k, defined on $c_{k,0} \times c_{k,1}$, in the partial ontology O', we have to identify natural language formulations of this relation. We are thus interested in patterns P_k of the form "[$c_{k,0}$] expression [$c_{k,1}$]"[4] that express the relation p_k in natural language.

Say, for example, we consider the classes *'author'* and *'book title'* and the relation *'has written'*. We assume that we know some related instance-pairs, e.g. *('Leo Tolstoy', 'War and Peace'*) and *('Günter Grass', 'Die Blechtrommel'*). We then want to find natural language phrases that relate authors to the titles of the books they wrote. If we find phrases that typically connect two related terms (i.e. patterns), we can expect them to also connect other related instance-pairs. Thus, if we query a pattern in combination with the name of an author (e.g. *'Umberto Eco wrote'*), we want the search results of this query to contain the books by this author.

Such patterns have to meet two criteria:

- **Precision:** Preferably, the phrase is unambiguous, i.e. the probability that the terms found do not belong to the intended class must be small. For example, consider the relation place of *birth(Person, City)*. The pattern *[Person] was born in [City]* is an ambiguous representation of this relation, since *[Person] was born in* can precede a date or the name of a country as well.
- **Recall:** The pattern must frequently occur on the Web. Rare patterns are not likely to give much search results when querying such a pattern in combination with an instance.

Identifying Relation Patterns

We focus on the relation p_k in our ontology O. For easy reference, we assume $p_k = c_q \times c_a$. The problem is to identify a set of patterns that lead to highly relevant search results when queried in combination with instances of c_q. We first generate a list of relation patterns with the use of the following algorithm. For evaluation purposes, we also compute the frequency of each pattern found.

- **Step 1:** Formulate queries using an instance-pair $(x,y) \in T'_k$. Since we are interested in phrases within sentences rather than in keywords or expressions in telegram style that often appear in titles of Web pages, we use the allintext: option. This gives us only search results with the queried expression in the bodies of the documents rather than in the titles. We query both allintext:" x * y " and allintext:" y * x ". The * is a regular expression operator accepted by both Google and Yahoo!. It is a placeholder for zero or more words.
- **Step 2:** Send the queries to the search engine and collect the snippets of the pages it returns for each query.
- **Step 3:** Extract all phrases matching the queried expressions and replace both x and y by the names of their classes.
- **Step 4:** Remove all phrases that are not within one sentence.
- **Step 5:** Normalize all phrases by removing all mark-up that is ignored by the search engine. Since the search engines are case-insensitive and ignore punctuation, double spaces and the like, we translate all phrases found to a normal form: the simplest expression that we can query that leads to the document retrieved.
- **Step 6:** Update the frequencies of all normalized phrases found.

- **Step 7:** Repeat the procedure for any unqueried pair $(x', y') \in T'_k$

We now have generated a list with relation patterns and their frequencies within the retrieved snippets.

Selecting Relation Patterns

From the list of relation patterns found, we are interested in the most effective ones. We are not only interested in the most precise ones. For example, the retrieved pattern ''född 30 mars 1853 i'' proved to be a 100% precise pattern expressing the relation between a person ('Vincent van Gogh') and his place of birth ('Zundert'). Clearly, this rare phrase is unsuited to mine instance-pairs of this relation in general. On the other hand, high frequency of some pattern is no guarantee for effectiveness either. The frequently occurring pattern ''was born in London'' (found when querying for Thomas Bayes * England) is well-suited to be used to find London-born persons, but in general the pattern is unsuited—since too narrow—to express the relation between a person and his or her country of origin. Taking these observations into account, we formulate three criteria for selecting effective relation patterns.

1. The patterns should *frequently* occur on the Web, to increase the probability of getting any results when querying the pattern in combination with an instance.
2. The pattern should be *precise*. When we query a pattern in combination with an instance in I_q, we want to have many search results containing instances from c_a.
3. If relation R is not functional, the pattern should be *broad*, i.e. among the search results when querying a combination of the pattern and an instance in I_q there must be as many distinct R-related instances from c_a as possible.

To measure these criteria, we use the following scoring functions for a relation pattern s.

1. **Frequency**. (shown in Box 1)
2. **Precision**.

$$f_{prec} = \sum_{x \in I'q}^{n} \frac{P(s,x)}{|I'_q|}$$

For instances $x \in I'_q$, $I'_q \subseteq I_q$ we calculate P(s,x) as follows.

Where $F_I(s,x)$ and $F_O(s,x)$ are given as such in Box 2.

Box 1.

$f_{freq}(s)$ = "number of occurrences of s in the snippets as found by the algorithm described in the previous subsection"

Box 2.

$F_I(s, x)$ = "the number of snippets after querying s in combination with x containing instances of c_a" $F_0(s, x)$ = "the number of snippets found (at most 1,000)"

Box 3.

$B(s, x)$ = "the number of distinct instances of class c_a found after querying s in combination with x"

3. **Breadth**.

$$f_{spr} = \sum_{x \in I'q} B\left(s, x\right)$$

Where B(s,x) is defined in Box 3.

$B\left(s,x\right) =$ "the number of distinct instances of class c_a found after querying s in combination with x"

The larger we choose the test set, the subset I'_q of I_q, the more reliable the measures for precision and breadth. However, the number of queries increases with the number of patterns found for each instance we add to I'_q. We finally calculate the score of the patterns by multiplying the individual scores: $score(s)=f_{freq} \cdot f_{prec} \cdot f_{spr}$

For efficiency reasons, we only compute the scores of the patterns with the highest frequencies. The problem remains how to recognize a (possible multi-word) instance in the snippets. For an ontology alignment setting—where the sets I_q and I_a are not to be expanded—these problems are trivial: we determine whether $t \in I_a$ is accompanied by the queried expression. For a setting where the instances of c_a are not all known (e.g. it is not likely that we have a complete list of all books written in the world), we solve this problem in two stages. First we identify rules per class to extract candidate instances. Thereafter we use an additional query to verify if a candidate is indeed an instance of class c_a.

Instance Identification

A separate problem is the identification of terms in the text. An advantage is that we know the place in the text by construction (i.e. either preceding or following the queried expression). A disadvantage is that each class requires a different technique to identify its instances. Especially terms with a less determined format, such as movie titles, are hard to identify. We therefore design recognition functions f_i for each class.

For these functions f_i, we can adopt various techniques from the fields of (statistical) natural language processing in general and information extraction in particular. A regular expression that describes the instances of class c_i can be a part of the function f_i. The reader may also think of the use of part of speech tagging (Brill, 1992), N-gram models (Manning & Schütze, 1999; Downey, Broadhead, & Etzioni, 2007), off-the-shell named entity recognizers (Zhou & Su, 2002) or shallow parsers (Lin, 1998). We note that the HTML-markup can be of use as well, since terms tend to be emphasized, or made ‘clickable’.

After extracting a term, we can perform an additional check to find out whether the extracted term is really an instance of the concerning class. We perform this check with the use of a search engine. We query phrases that express the term-class relation. Again, these phrases can be constructed semi-automatically. Hyponym patterns are candidates as well for this purpose (Hearst, 1992, 1998; Cimiano & Staab, 2004). A term is to be accepted as instance, when the number of hits of the queried phrase is at least a certain threshold. For example, we query the phrase ‘Cities such as Eindhoven and’ to check whether ‘Eindhoven’ is indeed an instance of the class City. When we use such a check function, we can allow ourselves to formulate less strict recognition functions f_i. That is, false instances that are accepted by f_i, are still rejected as an instance by the use of the check function.

Sketch of Algorithm

Per relation, we maintain a list of instances that already have been used in a query in combination with the patterns expressing this relation. Initially, these lists are thus empty. The following steps of the algorithm are performed until either some stop criterion is reached, or until no new instances or instance-pairs can be found.

- **Step 1:** Select a relation p_k on $c_{k,0} \times c_{k,1}$ and an instance v from either $I_{k,0}$ or $I_{k,1}$ we have not yet used in a query.
- **Step 2:** Combine the patterns expressing p_k (e.g. 'starring') with v (e.g. '*David Hasselhoff'*) and send these queries ('*starring David Hasselhoff'*) to the search engine.
- **Step 3:** Extract instances from the snippets using the instance identification rules for the class applicable.
- **Step 4:** Add the newly found instances to the corresponding instance set and add the instance-pairs found (thus with v) to $T'_{(k,0),(k,1)}$.
- **Step 5:** If there exists an instance that we can use to formulate new queries, then repeat the procedure.

Note that instances of class $c_{k,0} = c_{l,1}$ learned using the algorithm applied on relation p_k on $c_{k,0} \times c_{k,1}$ can be used as input for the algorithm applied to some relation p_l on $c_{l,0} \times c_{l,1}$ to populate the sets $I'_{l,0}$ and $T'_{(l,0),(l,1)}$.

Experimental Results

In this section, we discuss some experiments we conducted with this method. In Section 2.6, we investigate whether the learned patterns are indeed intuitive, by learning a list of relation patterns expressing the broader-narrower or hyponym relation. Section 2.6 handles the use of learned patterns to populate an ontology on restaurant chains and the countries where they can be found, while we populate a movie ontology in Section 2.6. Finally, we discuss the identification of a ranked list of historical persons and their biographical information in Section 2.6.

Learning Effective Hyponym Patterns

We are interested whether the effective surface text patterns are indeed intuitive formulations of some relation p_k: $c_q \times c_a$. As a test-case, we compute the most effective patterns for the hyponym relation using a test set with names of all countries.

Our experiment was set up as follows. We collected the complete list of countries in the world from the CIA World Factbook (footnote: http://www.cia.gov/cia/publications/factbook). Let I_a be this set of countries, and let I_q be the set { 'countries', 'country' }. The set T_k consists of all pairs (a,'countries') and (a, 'country'), for a $\in I_a$. We apply the surface text pattern learning algorithm on this set T_k.

The algorithm identified almost 40,000 patterns. We computed f_{spr} and f_{prec} for the 1,000 most frequently found patterns. In Table2, we give the 25 most effective patterns found by the algorithm. Focusing on these patterns, we observe two groups: 'is a' and hyponym patterns as identified by Hearst (1992).

The hyponym patterns 'like' and 'such as' show to be the most effective. This observation is useful, when we want to minimize the amount of queries for hyponym patterns.

Expressions of properties that hold for each country and only for countries, for example the existence of a country code for dialing, are not trivially identified manually but are useful and reliable patterns.

The combination of 'is a', 'is an' or 'is the' with an adjective is a common pattern, occurring 2,400 times in the list. In future work, such adjectives can be identified in the snippets using a Part of Speech tagger (Brill, 1992) or a shallow parser (Lin, 1998 ; Marneffe, MacCartney, & Manning, 2006).

A Restaurant Ontology

The Text Retrieval Conference (TREC) question answering track in 2004 contains list question, for example 'Who are Nirvana's band members?' (Voorhees, 2004). We illustrate the use of our ontology population algorithm in the context of such list-question answering with a small case-study.

Table 2. Learned hyponym patterns and their scores given in descending order of effectiveness

pattern	freq	prec	spr
(countries) like	645	0.66	134
(countries) such as	537	0.54	126
is a small (country)	142	0.69	110
(country) code for	342	0.36	84
(country) map of	345	0.34	78
(countries) including	430	0.21	93
is the only (country)	138	0.55	102
is a (country)	339	0.22	99
(country) flag of	251	0.63	46
and other (countries)	279	0.34	72
and neighboring (countries)	164	0.43	92
(country) name republic of	83	0.93	76
(country) book of	59	0.77	118
is a poor (country)	63	0.73	106
is the first (country)	53	0.70	112
(countries) except	146	0.37	76
(country) code for calling	157	0.95	26
is an independent (country)	62	0.55	114
and surrounding (countries)	84	0.40	107
is one of the poorest (countries)	61	0.75	78
and several other (countries)	65	0.59	90
among other (countries)	84	0.38	97
is a sovereign (country)	48	0.69	89
or any other (countries)	87	0.58	58
(countries) namely	58	0.44	109

Note that we do not consider the processing of the question itself in this research.

Inspired by one of the questions (‘What countries is Burger King located in?’), we are interested in populating an ontology with restaurants and the countries in which they operate. We identify the classes ‘country’ and ‘restaurant’ and the relation ‘located in’ between the classes.

We hand the algorithm the instances of ‘country’, as well as two instances of ‘restaurant’: ‘McDonald’s’ and ‘KFC’. Moreover, we add three instance-pairs of the relation to the algorithm. We use these pairs and a subset $I'_{country}$ of size eight to compute a ranked list of the patterns. We extract terms consisting of one up to four capitalized words. In this test we set the threshold for the number of hits for the queries with the extracted terms to 50. After a small test with names of international restaurant branches, this seemed an appropriate threshold.

The algorithm learned, besides a ranked list of 170 surface text patterns (Table 3), a list of 54 instances of restaurant (Table 4). Among these instances are indeed the names of large international chains, Burger King being one of them. Less expected are the names of geographic locations

Table 3. Top learned patterns for the restaurant-country relation

pattern	prec	spr	freq
c_a restaurants of c_q	0.24	15	21
c_a restaurants in c_q	0.07	19	9
c_a hamburger chain that occupies			
villages throughout modern day c_q	1.0	1	7
c_a restaurant in c_q	0.06	16	6
c_a restaurants in the c_q	0.13	16	2
c_a hamburger restaurant in southern c_q	1.0	1	4

and names of famous cuisines such as ‘Chinese’ and ‘French’. The last category of false instances found that have not be filtered out, are a number of very common words (e.g. ‘It’ and ‘There’).

We populate the ontology with relations found between Burger King and instances from country using the 20 most effective patterns. The algorithm returned 69 instance-pairs with countries related to ‘Burger King’. On the Burger King Web site[5] a list of the 65 countries can be found in which the hamburger chain operates. Of these 65 countries, we identified 55. This implies that our results have a precision of 55/69 = 80% and recall of 55/65 = 85%. Many of the falsely related countries—mostly in Eastern Europe—are locations where Burger King is said to have plans to expand its ‘empire’.

Populating a Movie Ontology

For this case study, we have constructed a small partial ontology on the movie domain. It is defined as O'_{movie}= ((Director, Actor, Movie),({‘Steven Spielberg’,‘Francis Ford Coppola’}, {}, {}),(‘acts-in(Movie,Actor),director-of(Movie,Director)),({}, {})).

We thus only identify three classes, of which only the class Director has instances. Using our method, we want to find movies directed by these directors. The movies found are used to find starring actors, where those actors are the basis of the search for other movies in which they played, etc. The process continues until no new instances can be found.

Relation patterns. This small ontology contains two relations, acts in and director of. For these relations, we have manually selected the sets of patterns: {“[Movie] starring [Actor]”, “[Actor] and [Actor]”} and {“[Director]’s [Movie]”, “[Movie], director: [Director]”}.

Instance identification. We identify a term as a Movie title, if it is placed in a text between quotation marks (Geleijnse & Korst, 2005; Sumida, Torisawa, & Shinzato, 2006). Although this may seem a severe restriction, in practice we can permit to loose information contained in other formulations since each query-result gives much redundant information. So, if a movie title is placed between quotation marks just once in the search results, we are able to recognize it.

A person’s name (instances of the classes Director and Movie) is to be recognized as either two or three words each starting with a capital. Another feature of the recognition function is the use of lists with taboo words. If a taboo word is contained in an expression, we ignore it. We use a list of about 90 taboo words for the person names (containing words like ‘DVD’ and ‘Biography’). For the movie titles we use a much shorter list, since movie titles can be much more diverse. We have constructed the taboo word lists based on the output of a first run of the algorithm.

We check each o the extracted candidate instances with the use of one of the following que-

Table 4. Learned instances for restaurant

Chinese	Bank	Outback Steakhouse
Denny's	Pizza Hut	Kentucky Fried Chicken
Subway	Taco Bell	Continental
Holywood	Wendy's	Long John Silver's
HOTEL OR	This	**Burger King**
Japanese	West	Keg Steakhouse
You	BP	Outback
World	Brazil	San Francisco
Leo	Victoria	New York
These	Lyons	Starbucks
FELIX	Roy	California Pizza Kitchen
Marks	Cities	Emperor
Friendly	Harvest	Friday
New York	Vienna	Montana
Louis XV	Greens	Red Lobster
Good	It	There
That	Mark	Dunkin Donuts
Italia	French	Tim Hortons

ries: ''The movie [*Movie*]'', ''[*Actor*] plays'', or ''[*Director*] directed''. A candidate is accepted, if the number of search results to the query exceeds a threshold. After some tests we choose 5 as a threshold value, since this threshold filtered out not only false instances but most of the common spelling errors in true instances as well.

Formulation of queries. The relation patterns lead to the following set of queries: {''[Director]'s'', ''[Movie] starring'', ''[Movie] director'', ''starring [Actor]''}. We have analyzed the first 100 snippets returned by the search engine after querying a pattern in combination with an instance.

Results

We first ran the algorithm with the names of two (well-known) directors as input: Francis Ford Coppola and Steven Spielberg. Afterwards, we experimented with larger sets of directors and small sets of beginning directors as input.

An interesting observation is that for this case study the output can be independent of the input sets. That is, when we take a subset of the output of an experiment as the input of another experiment, the outputs are the same, modulo some small differences due to the changes in the search results over time.

We have found 7,000 instances of the class Actor, 3,300 of Director and 12,000 of Movie. The number of retrieved instances increases, about 7%, when 500 query results are used instead of 100.

Precision. When we analyze the precision of the results, we use the data from the Internet Movie Database (IMDb) as a reference. An entry in our ontology is accepted as a correct one, if it can be found in IMDb. We have manually checked three sequences of 100 instances (at the beginning, middle and end of the generated file) of each class. We estimate a precision of 78%. Most misclassified instances were misspellings or different formulations of the same entity (e.g. ''Leo DiCaprio'' and ''Leonardo DiCaprio'').

In the future, we plan to add post processing to recognize these flaws. We can analyze the context (e.g. when 2 actors act in the same set of movies) and use approximate string matching techniques to match these cases.

Likewise, we have also analyzed the precision of the relations. We estimate the precision of the relation between movie and director around 85%, and between movie and actor around 90%.

Recall. The number of entries in IMDb exceeds our ontology by far. Although our algorithm performs especially well on recent productions, we are interested how well it performs on classic movies, actors and directors. First, we made lists of all Academy Award winners (1927-2005) in a number of relevant categories, and checked the recall (Table 5).

IMDb has a top 250 of best movies ever. The algorithm found 85% of them. We observe that results are strongly oriented towards Hollywood productions. We also made a list of all winners of the Cannes Film Festival, the 'Palme d'Or'. Alas, our algorithm only found 26 of the 58 winning movies in this category.

Extracting Information on Historical Figures

The second case study aims at extracting a long list of historical persons and in addition extracting for each of them biographical information such as nationality, period of life, and profession. Using this additional information, we can create sublists of e.g. 17th-century Dutch painters. The information extraction is carried out in two phases. First a long list of historical persons is extracted, and secondly, additional information on these persons is gathered.

Relation Patterns and Query Formulation

It has been observed by e.g. (Ravichandran & Hovy, 2002) that a surface pattern as ''Wolfgang Amadeus Mozart ('' is very successful to determine the year of birth of in this case Mozart, as the open bracket will be often followed by the period of life of the person (in this case: 1756-1791). We decided to use this observation but in a different fashion (Korst, Geleijnse, De Jong, & Verschoor, 2006). Instead of looking for the year of birth of a given person, we use year intervals that possibly relate to the lifetime of a person to find historical persons. More precisely, we issued all year intervals ''($y_1 - y_2$)'' as queries, with $y_1 \leq [1000 \ldots 1990]$, $y_2 - y_1 \leq [15 \ldots 110]$, and $y_2 \leq 2005$. In other words, we search for persons who were born during the last millennium and who died at an age between 15 and 110. Note that, in this way, we will only find persons that already passed away.

Instance Identification

For each of these issued queries, we scan the at most 1000 snippets that the search engine returned. In each of these snippets, we determine the first occurrence of the queried pair of numbers. Since search engines ignore non-alphanumeric characters, the queried pair of numbers may also occur as

Table 5. Recall of academy award winners

Category	Recall
Best Actor	96%
Best Actress	94%
Best Director	98%
Best Picture	87%

y_1, y_2 or as y_1 / y_2. If the queried pair of numbers is in the intended context ($y_1 – y_2$), i.e. if they are surrounded by brackets and separated by a hyphen, then the words directly preceding this first occurrence are stored for later analysis, to a maximum of six words. In this way, we obtain for each queried pair of numbers up to 1000 short text fragments that potentially contain person names. In addition, for each of the stored text fragments, we remove potential pre- and suffixes that normally cannot be part of a name. For example, we delete all words that precede a full stop (except when preceded by a single capital letter), a colon, or a semicolon. In addition, of words consisting of upper-case letters only we transform the upper-case into lower-case letters, except for the first one (with some specific exceptions concerning ordinal numbers of kings, queens, etc., composite names including hyphens or apostrophes, and Scottish and Irish names). This results in a set of candidate names.

The *check* phase consists of two filtering steps: one to filter out non-person names and one to filter out multiple variants of a single person name. These steps are next discussed in more detail. Not all text fragments we have found in the extraction phase will be person names. Typically, historic periods, art styles, geographic names, etc. can also directly precede a time interval. Table 6 illustrates the difficulties in discriminating between person names and other text fragments. We note that West Mae probably refers to the person name *Mae West* and that *Napoleon Hill* refers to a person as well as to a geographic location in the state Idaho (USA).

To filter out non-person names, we first constructed from dedicated Web sites a long list of the most common first names (boy's and girl's names). If a text fragment starts with such a name, then this is a strong indication that the text fragment is a person name. In addition, we constructed a long list of suspect words that typically do not occur in person names, as follows. From the many snippets that we gathered with the year interval queries we extracted all words, counting how often they occur with a capital and without a capital. If a word occurs most often without a capital, and it is not a special word as 'van', 'de', or 'la', then it is added to the long list of suspect words. We next apply a rule-based approach using these lists of first names and suspect words to filter out text fragments that probably do not relate to person names. In addition to filtering out non-person names, we also want to filter out multiple occurrences of the same person name. These occurrences are caused by variations in spelling of names and errors in the lifetimes. To this end, we carried out the following filtering steps.

1. **Keeping only the last name/lifetime variants that occur most often.** For each last name/lifetime combination, we often find different variants of first names preceding it. For example, Bach (1685–1750) is preceded by, e.g., Johann Sebastian, JS, and Johann S. Of all these variants we only store the

Table 6. Some examples to illustrate the difficulties in discriminating between persons names and other text fragments

Person Name	Non-Person Names
Art Blakey	Art Deco
West, Mae	West Virginia
Amy Beach	Miami Beach
HP Lovecraft	HP Inkjet
Napoleon Hill	Napoleon Hill

one that is found most often, i.e., the variant that occurs most often in the text fragments we found in the 1000 snippets returned on query ‘’(1685–1750)”.

2. **Filtering out small variations in name.** If two names have exactly the same lifetime and the edit distance (Gusfield, 1997) between these full names is less than a given threshold, then only the variant that is found most often is kept. As threshold we use an edit distance of two.
3. **Filtering out single errors in lifetimes.** If two names are completely identical but their lifetimes differ in only the year of birth *or* the year of death, then only the variant that is found most often is kept.

Experiments indicate that in this step we reduce the candidate set of names by approximately 25%.

Ordering Persons by Fame

To order the persons by fame, we use the number of hits (the ‘Google Page Count’ GPC) as our measure of fame. Now, the question is which query we should issue to the search engine to determine the GPC of a person. The query should be neither too general nor too specific. A single person is often identified in different ways, e.g. Johann Sebastian Bach, JS Bach, JOHANN SEBASTIAN BACH and Bach, Johann Sebastian all refer to the same person. The last variant is called an *inversion*. The latter two variants can be transformed into the first variant by substituting upper-case characters by lower-case ones and by adjusting the order of first and last names. Complicating factors in the identification of inversions are *(i)* that a comma between last name and first names is sometimes omitted and *(ii)* that many first names also occur as last names. An additional complication is that the first names sometimes vary per language (e.g. Charles vs. Karel). To achieve that we are less sensitive to these variants, we use the following query to determine the GPC:‘’[last name] ([year of birth]–[year of death])”

For kings, queens, popes, etc., we use the Latin ordinal number as last name. In this way Charles V (1500–1558), Carlos V (1500–1558), and Karel V (1500–1558) are all covered by query ‘’V(1500–1558)”. Note that we assume the combination of last name and lifetime to be specific enough to uniquely identify historical persons.

Extracting Additional Information

The first phase, described above, resulted in a large list of historical persons that was ordered using GPC as measure of fame. For further details on this list we refer to (Korst, Geleijnse, De Jong, & Verschoor, 2006). In the next phase, we extracted additional information, such as gender, nationality, and professions. Also, we tried to retrieve related images and a few one-liners that already give a brief impression of how the person gathered fame. We extracted additional information for the top 10,000 of the list of historical persons that we obtained in the first phase. We next briefly describe how we gathered this additional material.

To acquire additional information, we again issued queries of the type ‘’Albert Einstein was’’, i.e., we used the full name of a person followed by the word was, where we restrict ourselves to English language pages. From the snippets returned, we extracted complete sentences that contain the query. Hence, if only a fraction of a sentence was given in a snippet, then this fraction was simply ignored. These sentences were next used to identify specific words that indicate gender, nationality and professions.

Determining gender. We simply counted words that refer to the male gender, namely the words he, his, son of, brother of, father of, man and men. Likewise, we counted words that refer to the female gender, namely the words she, her, daughter of, sister of, mother of, woman, and women. We simply assigned the gender with the highest count.

Table 7. The 20 persons born between 1880 and 1889 with the highest GPC

Born In [1880–1889]		
James Joyce (1882-1941)	Ireland	author
Bela Bartok (1881-1945)	Hungary	composer
Pablo Picasso (1881-1973)	Spain	artist
Anton Webern (1883-1945)	Austria	musician, composer
HL Mencken (1880-1956)	United States	author, journalist
Niels Bohr (1885-1962)	Denmark	scientist, physicist
Adolf Hitler (1889-1945)	Germany	leader
Amedeo Modigliani (1884-1920)	Italy	artist, painter
Agustin Barrios (1885-1944)	Paraguay	musician, composer
Le Corbusier (1887-1965)	Switzerland	architect
John Maynard Keynes (1883-1946)	United Kingdom	economist
Ludwig Wittgenstein (1889-1951)	Austria	philosopher
Igor Stravinsky (1882-1971)	Russia	composer
TS Eliot (1888-1965)	United Kingdom	poet
Franz Kafka (1883-1924)	Czech Republic	author
Franklin D. Roosevelt (1882-1945)	United States	president
Marc Chagall (1887-1985)	Russia	painter, artist
Martin Heidegger (1889-1976)	Germany	philosopher
Kahlil Gibran (1883-1931)	Lebanon	poet, philosopher,...
Heitor Villa-Lobos (1887-1959)	Brazil	composer

Determining nationality. We extracted for each country from the CIA World Factbook Web site the country name (in conventional short form) and the corresponding adjective that indicates nationality, e.g. ‘Belgium’ and ‘Belgian’. In addition, for some countries we added a number of additional terms relating to parts of the country, such as ‘Flemish’ for Belgium and ‘English’, ‘Scottish’, and ‘Welsh’ for the United Kingdom. To determine the nationality, we count for each country the number of word occurrences in the set of sentences, and simply assign the nationality with the highest count. So far, we did not consider country names of countries that do no longer exist, such as Prussia.

Determining professions. As for gender and nationality, we now simply count how often each of these profession names occur in the sentences. However, instead of only selecting the one with the highest count, we here want to be able to retain multiple professions. For that reason, we select the ones that have at least a count of $0.5 \cdot c_{max}$, where c_{max} is the score of the highest scoring profession, ordered by decreasing count.

Results

To give an impression of the results that we obtained in this case study, we present three tables. Table 7 gives the top of the persons born in the period 1880—1889, Table 8 gives the top of the persons that has as their highest scoring profession either artist or painter. For more details and examples we refer to (Korst, Geleijnse, De Jong, & Verschoor, 2006) and (Geleijnse and Korst, 2007).

Considering Table 9, we observe that Thomas Young is given the American nationality, while

Table 8. The 25 artists/painters with the highest GPC

Artists/Painters		
Leonardo da Vinci (1452–1519)	Italy	artist, scientist,...
Pablo Picasso (1881–1973)	Spain	artist
Vincent van Gogh (1853–1890)	Netherlands	artist, painter
Claude Monet (1840–1926)	France	artist, painter,...
Pierre-Auguste Renoir (1841–1919)	France	painter
Paul Gauguin (1848–1903)	France	painter
Edgar Degas (1834–1917)	France	artist, painter,...
Paul Cezanne (1839–1906)	France	painter, artist
Salvador Dali (1904–1989)	Spain	artist
Henri Michaux (1899–1984)	Belgium	artist, poet
Gustav Klimt (1862–1918)	Austria	painter, artist
Peter Paul Rubens (1577–1640)	Belgium	artist, painter
Katsushika Hokusai (1760–1849)	Japan	painter
Amedeo Modigliani (1884–1920)	Italy	artist, painter
JMW Turner (1775–1851)	United Kingdom	artist, painter
James McNeill Whistler (1834–1903)	United States	artist
Rene Magritte (1898–1967)	Belgium	artist, painter
Henri Matisse (1869–1954)	France	artist
Rembrandt van Rijn (1606–1669)	Netherlands	artist, painter
Edouard Manet (1832–1883)	France	artist, painter
Herm Albright (1876–1944)	—	artist, engraver,...
Marc Chagall (1887–1985)	Russia	painter, artist
Edvard Munch (1863–1944)	Norway	painter, artist
Wassily Kandinsky (1866–1944)	Russia	artist, painter
Francisco Goya (1746–1828)	Spain	artist, painter

he is from the United Kingdom. This is due to the fact that Thomas Young is a common name, thus many of the sentences starting with ‘Thomas Young was’ actually refer to different persons. Unfortunately, Marie Curie is not in this sublist. The only reason is that her only profession that is retained was ‘scientist’. Otherwise, she would have ranked between Rutherford and Feynman. For the same reason, Robert Hooke would have ranked between Heisenberg and Volta, and Robert Wilhelm Bunsen would have ranked between Hahn and Curie. The first profession of Sir David Brewster is judge, resulting from the phrase “Sir David Brewster was afraid that the members could scarcely judge ...”.

Recall. To get an impression of the performance of our algorithm, we estimate the recall by choosing a diverse set of six books containing short biographies of persons whom we would expect to find in our list. For each of these books, we determined for the persons that could potentially be found by our algorithm (i.e., the persons who are born in the intended time period and have died). Of these 1049 persons, 1033 were present in our list, which is a fraction of 0.98. For further details on the chosen books we refer to Korst, Geleijnse,

Table 9. The 40 physicists with the highest GPC

Albert Einstein (1879–1955)	scientist, physicist
Isaac Newton (1642–1727)	scientist, mathematician, physicist
Galileo Galilei (1564–1642)	astronomer, physicist, scientist
Niels Bohr (1885–1962)	scientist, physicist
Lord Kelvin (1824–1907)	physicist, scientist, mathematician
Christiaan Huygens (1629–1695)	astronomer, scientist, mathematician, physicist
Michael Faraday (1791–1867)	chemist, scientist, physicist
Georg Christoph Lichtenberg (1742–1799)	physicist, author, mathematician, astronomer
James Clerk Maxwell (1831–1879)	physicist
Max Planck (1858–1947)	physicist
Ernest Rutherford (1871–1937)	scientist, physicist
Richard Feynman (1918–1988)	physicist
Hermann von Helmholtz (1821–1894)	scientist, physicist
Werner Heisenberg (1901–1976)	physicist
Alessandro Volta (1745–1827)	physicist
Wolfgang Pauli (1900–1958)	physicist
Sir Arthur Eddington (1882–1944)	scientist, physicist, director, author
Heinrich Hertz (1857–1894)	scientist, physicist, explorer, author, researcher
Wilhelm Conrad Roentgen (1845–1923)	scientist, physicist, director
Paul Dirac (1902–1984)	physicist
Andre Marie Ampere (1775–1836)	physicist, mathematician
Joseph John Thomson (1856–1940)	scientist, physicist, explorer
James Prescott Joule (1818–1889)	scientist, physicist
Ludwig Boltzmann (1844–1906)	physicist
Thomas Young (1773–1829)	physicist, physician, scientist
Ernst Mach (1838–1916)	philosopher, physicist
Evangelista Torricelli (1608–1647)	physcist, mathematician
Charles Augustin Coulomb (1736–1806)	engineer, physicist, scientist
Otto Hahn (1879–1968)	chemist, physicist
Pierre Curie (1859–1906)	physicist
Otto von Guericke (1602–1686)	scientist, inventor, physicist
Sir David Brewster (1781–1868)	judge, physicist
William Henry Bragg (1862–1942)	physicist
Thomas Kuhn (1922–1996)	philosopher, physicist, historian
George Gamow (1904–1968)	physicist
Sadi Carnot (1796–1832)	physicist, engineer
Gustav Theodor Fechner (1801–1887)	physicist, psychologist, philosopher
Joseph Swan (1828–1914)	chemist, inventor, physicist
Louis de Broglie (1892–1987)	physicist
Augustin Fresnel (1788–1827)	inventor, physicist, engineer

De Jong, & Verschoor, 2006). We observe that the recall is close to one, for each of the six books, even for a more specialized topic as 17th century Dutch painters. Of the total 108 of these painters mentioned in one of the books, 106 were found. We note that of the 16 persons that did not appear in our list, there were 4 persons for which the books could not provide the lifetime. For the recall of the additional information, we observe that for the 10,000 persons that we considered all were given a gender, 77% were given a nationality, and 95% were given one or more professions.

Precision. All kinds of imperfections can still be observed in our list of historical persons, such as remaining inversions, missing parts of a name, and errors in lifetimes, although each of these occurs relatively infrequently. We concentrate on estimating the fraction of names that do not relate to persons. The corresponding precision that is obtained by the algorithm has been estimated as follows. We selected three decennia, namely 1220-1229, 1550-1559 and 1880-1889, and analyzed for each the candidate persons that were 'born' in this decennium. For the first two decennia we analyzed the complete list, for decennium 1880-1889 we analyzed only the first 1000 as well as the last 1000 names. This resulted in a precision of 0.94, 0.95, and 0.98, respectively. As the decennium of 1880-1889 resulted in considerably more names, we take a weighted average of these results. This yields an estimated precision for the complete list of 0.98 (Korst, Geleijnse, De Jong, & Verschoor, 2006). Regarding the precision of the property *lived-in*, we make the following observations. Considering the list of 450,000 potential instances that our algorithm found for this property, we observe that 235 were found with a GPC of at least 10,000 and 2450 were found with a GPC of at least 1000. Clearly, the probability that instances with a high GPC contain spelling errors in person name or lifetime is quite low, since accidental spelling errors in the last name or in the lifetime will result in a low GPC. Indeed, we found that the accuracy of our results was better than that of the information in some of the books.

Biographical entries. To get a better impression of the quality of the biographical entries, we manually checked 50 persons, evenly distributed in the top-2500. Of these 50 persons, we observed that gender, nationality and professions were all correct for 38 persons. No errors in gender were detected in any of the 50 persons. For three persons the nationality was not found. All nationalities found proved to be correct. For two persons, all given professions were wrong. For eight others, one or more given professions were incorrect, but usually the professions with the highest count were correct.

COMMUNITY-BASED KNOWLEDGE MINING

In this second part of the chapter, we focus on more subjective knowledge. For example, we are interested the most appropriate genre for a musical artist, using the collective knowledge of the Web community. As such knowledge is not only expressed (directly) within sentences, we discuss two alternative methods to find relations, apart from the pattern-based method described in the previous part. Since we focus on subjective knowledge, multiple relations are possible, e.g. Madonna may both be a Rock and a Pop artist. We therefore assign a score for each of the relations found. For example, the aim is to find a higher score for the relation between ABBA and Disco than between ABBA and Dead Metal. As the focus is on scoring relations, we assume the classes to be complete. This second part of the chapter is organized as follows. In Section3.1 we formally define the problem, while Section3.2 discusses the three alternative methods to categorize instances. In Section3.3 we discuss a similar method to find relatedness between the instances. The found relatedness scores can be used to improve the categorization (Section3.4). Experiments in two domains are discussed in Section3.5. Finally we conclude this part of the chapter in Section4.

Problem Description

We want to map a complete set of instances of class c_a- artists or items such as movies or paintings—to a set of categories: instances of class c_s. Given are two sets of instances I_a of size N and I_s of size M. Hence, I_a is a set with instances such as paintings, or artists. The set I_s contains categories like movements in art or genres. We are interested in a mapping m: $I_a \rightarrow I_s$.

Definition. We call a category m(b) *most appropriate* for b if a domain expert would select m(b) from the set I_s as the category best applicable to b.

Problem 1. Find for each b $\in I_a$ the most appropriate m(b) $\in I_s$.

Problem 2. Find for each pair of instances of c_a, *(b,c)* a score Γ*(b,c)* expressing the extent of relatedness between b and c.

We use co-occurrences of instances in I_a and categories in I_s (e.g. Johnny Cash and country) on the Web to extract such information to compute a preliminary mapping. Additionally, we assume that related instances in I_a often share a category in I_s. We compute distances between instances using Web co-occurrences in order to identify the nearest neighbors for each instance. The preliminary mapping of each artist and its nearest neighbors are combined into a final mapping m.

Three Classification Methods

In this section, we present three methods to classify instances in I_a using Web data. The first method is based on analyzing the total numbers of co-occurrences of instances in I_a and categories in I_s on the Web. To retrieve this data we again use a state of the art search engine. An important drawback of this page count method is that it has a high *Google Complexity*, i.e. it requires many queries to a search engine, namely $O(N \cdot M)$. For large sets this can be problematic, since search engines currently allow only a limited amount of automated queries per day (Cafarella, Downey, Soderland, & Etzioni, 2005). Moreover, the number of hits can fluctuate over time (Véronis, 2006), which hampers the reuse of old hit counts. We present two alternative methods that do not suffer from these drawbacks.

We are interested in a mapping m', based on co-occurrences of elements in I_a and I_s. In Section3.4 we combine this mapping m' with a distance matrix between related instances in I_a to find a definite mapping *m*.

Page-Count-Based Mapping (PCM)

To obtain the mapping m' we perform a query "b","g" for each pair (b,g) $\in I_a \times I_s$. Per query, we extract the estimated number of *hits* co(*b*,*g*).

co(*b*,*g*)='the number of hits for query "b", "g"'

We assume that the order of the terms *b* and *g* in the query does not effect the number of hits, thus we assume co(*b*,*g*)= co(*g*,*b*).

This Page-Count-based Mapping (PCM) is simple and intuitive. If we are for example interested in categorizing music artists into genres, we analyze the number of hits to queries for combinations of the names of the artist and each genre. Assuming Johnny Cash to be a country artist, we expect that more documents contain both the terms *Country* and *Johnny Cash* than *Reggae* and *Johnny Cash*.

Per b $\in I_a$ we could map the g $\in I_s$ with the most hits. However, we observe that frequently occurring categories in I_s have a larger probability to be mapped to any instance in I_a. For example, the query '*Pop*' results in 8 times more hits than the query '*Disco*'. Although we consider *Boney M* as a *disco-act*, the query *'Boney M, pop'* gives twice the amount of hits as *'Boney M, disco'*. This observation leads to a normalized approach, inspired by the theory of pointwise mutual information (Manning & Schütze, 1999; Downey et al., 2005). For $b \in I_a$ and $g \in I_s$, we define a scoring function S(b,g) as follows.

$$S(b,g) = \frac{co(b,g)}{1 + \sum_{c \in Ia} co(c,g)} \tag{1}$$

In the denominator we add 1 to the sum of all co-occurrences with g to avoid dividing by 0.

Having computed the scores for all pairs, we select a preliminary mapping *m'* for each $b \in I_a$. Per instance we select the category $g \in I_s$ with the highest score S.

$$m'(b) = \arg\max_{h \in 1g} S(b,h) \tag{2}$$

Using PCM we thus need to perform N · M queries.

Pattern-Based Mapping (PM)

The Pattern-based Mapping (PM) is based on occurrences of terms in phrases on the Web. We observe combinations of terms in phrases that express the relation we are interested in. For example, if we are interested in the relation between music artists (in I_a) and their genres (in I_s), an appropriate phrase that links terms of the two could be *'[artist] is one of the biggest [genre] artists'*. We can identify these patterns automatically by using a training set of related instances and categories; see (Ravichandran & Hovy, 2002; Geleijnse & Korst, 2006) and Section2.3 of this chapter. Learning patterns can be done with *O*(N) queries.

We use combinations of a pattern and an instance or a category as a query to the search engine. For example, if we have the pattern *''[genre] artist such as [artist]''*, we use *''artist such as''* in queries in combinations with all names of genres and artists. We use this pattern e.g. both for the query *''Country artists such as''* and for the query *''artists such as Prince''*. In the snippets found with the first query, we identify instances in I_a, while in the results for the second query we search for categories in I_s related to Prince. These queries provide access to relevant data. From the snippets returned by the search engine, we thus identify the elements of either I_a or I_s to measure the number of co-occurrences of the pairs. Hence, using PM co(b,g) is defined as follows.

co(b,g) = 'number of occurrences of *b* by querying patterns with *g*' + 'number of occurrences of *g* by querying patterns with *b*'

Using PM we only need *O*(M+N) queries. We use the same scoring function S(b,g) as given in (1) to obtain a preliminary mapping as given in (2).

Document-Based Mapping (DM)

In the Document-based Mapping (DM) approach we collect the first *k* URLs of the documents returned by the search engine for a given query. These *k* URLs are the most relevant for the query submitted based on the ranking used by the search engine (Brin & Page, 1998).

In the first phase of the algorithm, we query all instances in both I_a and I_s and collect the top *k* documents for each of the queries. For instances in I_a, we retrieve each document using the URLs found by the search engine. We count the occurrences of the categories in I_s (thus the names of the categories) in the retrieved documents for the intermediate mapping *m'*. From the documents retrieved with a category $g \in I_s$, we similarly extract the occurrences of instances in I_a.

The documents obtained using DM are the most relevant for each element $b \in I_a$. For the artists queried we expect biographies, fan pages, pages of museums, entries in database sites and so on. The categories in I_s (e.g. the genres or styles) mentioned in these pages will most probably reflect the genre of the artist queried.

Thus the co-occurrences function is here defined as follows.

co(b,g) = 'number of occurrences of *b* in documents found with *g*' + 'number of occurrences of *g* in documents found with *b*'

The co-occurrences of elements in I_a and I_s again are used for an intermediate mapping using the same scoring function. This method also requires only *O*(M+N) queries. However, additional data communication is required since for each query up to *k* documents have to be downloaded instead of using only the data provided by the search engine.

Finding Related Instances

We use the same three co-occurrence-based methods to compute the relatedness between elements in I_a. We consider instances in I_a to be related, when they frequently co-occur in the same context. In Section3.4, we use this information in a final mapping *m* between elements in I_a and I_s. Per pair (b,c) $\in I_a \times I_a$ we compute the score $\Gamma(b,c)$, similar to the score S in (1).

$$r\left(b,c\right) = \frac{co\left(b,c\right)}{\sum_{x,x\neq} co\left(x,c\right)} \tag{3}$$

Again, we do not use a majority voting to prevent frequently occurring instances to be strongly related to many other instances.

In PCM we combine the names of two artists into a query and extract the number of *hits*. Using this method this phase requires N^2 queries.

If we use PM to obtain the numbers of co-occurrences of instances in I_a, we can specify the nature of the relatedness. For example, for instances of the class pop artist, we can solely be interested in artists who have played together. A pattern such as *''[pop artist] recorded a duet with [pop artist]''* could be suitable for this purpose. This phase of the method consists of $k \cdot$ N queries (with *k* the number of patterns). In the documents obtained with the DM method we only expect occurrences of other terms in I_a that are strongly connected with the b $\in I_a$ queried. For DM no additional queries have to be performed in this phase, since we can reuse the documents obtained in the first phase.

Combining Results in Final Mapping

We use the assumption that related instances in I_a often share the same category. We investigate whether the use of relatedness between instances in I_a helps to improve to the precision of the mapping *m'*. We combine the scores Γ with the preliminary mapping *m*' as follows. Per b $\in I_a$, we inspect *m*' to determine the category that is assigned most often to b and its *n* closest related instances. We thus expect that the most appropriate category *g* for b is most often mapped by *m*' among b and its nearest neighbors.

Per instance b $\in I_a$, we construct an ordered list B(n) with b and its *n* nearest neighbors

$B(n) = (b_0, b_1, \ldots, b_n)$
with $b = b_0$ as its first element and
$\Gamma(b,b_i) \geq \Gamma(b,b_{i+1})$, for i < 0

For a final mapping *m* of instances b $\in I_a$ to a category in I_s, we inspect the most occurring category mapped by *m'* to b and its *n* nearest neighbors.

$$m\left(b,n\right) = \arg\max_{h \in Is} \sum_{c \in B\left(n\right)} \tau\left(c,h\right)$$

with

$$\tau(c,h) = 1 \quad \text{if } m'(b_i) = h$$
$$= 0 \quad \text{otherwise}$$

Table 10. patterns for artist—artist relation

"like I_{224} and I_{224}"
"such as I_{224} and I_{224} "
"including I_{224} and I_{224} "
"for example I_{224} and I_{224} "
"namely I_{224} and I_{224}"
" I_{224} and I_{224} "
" I_{224} I_{224} and other"

If two categories have an equal score, we select the first occurring one. That is, the category that is mapped by *m'* to b or to the artist most related to b.

Experimental Results

We evaluate the methods discussed in this part of the chapter as follows. First, we evaluate the instance-similarity scoring using the three methods PM, DM and PCM. These similarities are used in the subsequent part to classify musical artists into genres. Finally, we repeat this experiment on a different domain to classify painters into art movements.

Musical Artist Similarity

We use the common test set I_{224} of 224 artists, equally divided over 14 genres as defined by Knees et al. (2004)[6] to evaluate the computed artist similarities Γ(b,c). We consider two artists to be similar, if they share a genre in the test set. In these experiments, we only evaluate precision. If for an b $\in I_a$ no mapping or related instance could be found, we consider the result to be incorrect.

For PCM we added the extra term music for finding co-occurrences of the artists. For example the terms Bush and Inner Circle co-occurred a lot on the Web, due to American politics. By adding the term music we restrict ourselves to documents handling music.

Figure 1. Precision for classification of the 224 musical artists

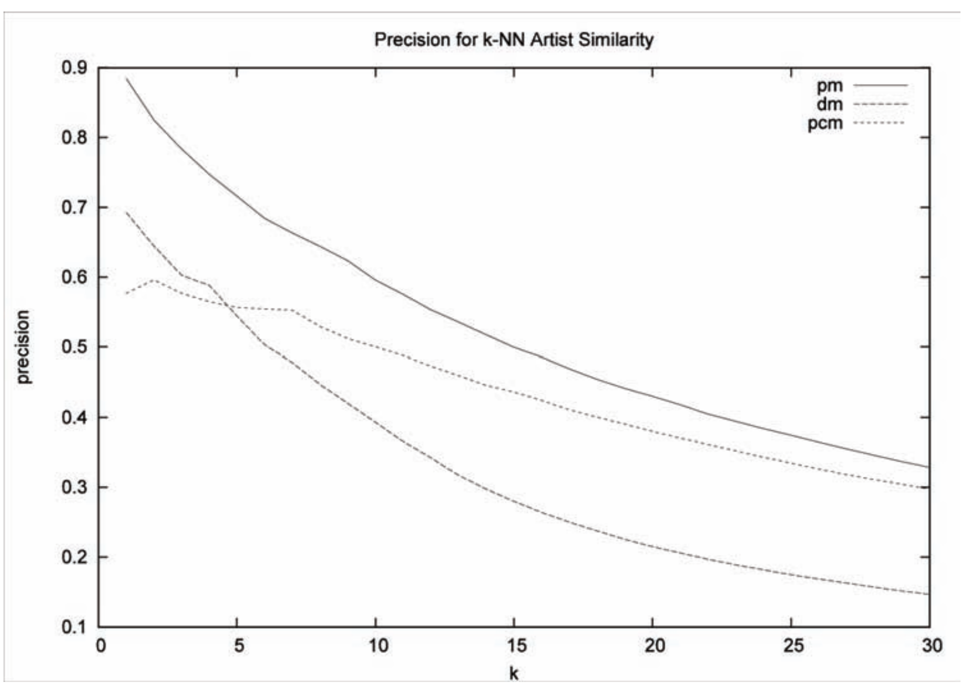


Table 11. The patterns for the artist-genre relation

"I_g (artists OR bands OR acts OR musicians) like I_{224}"
"I_g (artists OR bands OR acts OR musicians) such as I_{224}"
"I_g (artists OR bands OR acts OR musicians) for example I_{224}"
"I_{224} and other I_g (artists OR bands OR acts OR musicians)"

Since we are not interested in the nature of the relatedness between artists, for PM we selected general enumeration patterns (Table 10) to obtain co-occurrences.

Figure 1 shows the average precision of the similarity of the artists and their *k*-NN for the sets of 224 artists. We can conclude that the pattern based method PM gives good results and outperforms both DM and PCM. For smaller values of *k* the method most inefficient in the number of queries is outperformed by both DM and PM. The performance of DM drops quickly due to the fact that only few related artists are mentioned among the best ranked pages for the queried instances.

Musical Artist Genre Tagging

In this experiment, I_{224} is again the set of all artist names in the list composed by Knees, Pampalk & Widmer (2004). This list consists of 14 genres, each with 16 artists.

To find a mapping between I_{224} and the set of genres I_g, the genres mentioned in the list are not all suitable for finding co-occurrences. For example, the term *classical* is ambiguous and *Alternative Rock/Indie* is an infrequent term. We therefore manually rewrote the names of the genres into unambiguous ones (such as *classical music*) and added some synonyms. After collecting the numbers of co-occurrences of artists and genres, we summed up the scores of the co-occurrences for synonyms. Thus, for each artist *b* the number of co-occurrences with the terms *Indie* and *Alternative Rock* are added to the co-occurrences of *b* with the genre *Alternative Rock/Indie*. Motivated by the results in (Schedl et. al, 2005), for PCM we used the allintitle option in the artist classification experiment.

For PM we selected for the genre-artist relations the patterns in Table 11 from a list of patterns found expressing this relation.

For all three methods, we reuse the computed artist similarities.

In Table 12 the performance of the preliminary mapping can be found for the three methods (n = 0). We were able to map all artists to a genre. Co-occurrences between genres and artists thus could be found using PCM, PM as well as DM. The latter performs best. With respect to the preliminary mapping, the method with the smallest amount of Google queries performs best. The data found on the best ranked documents is thus reliable.

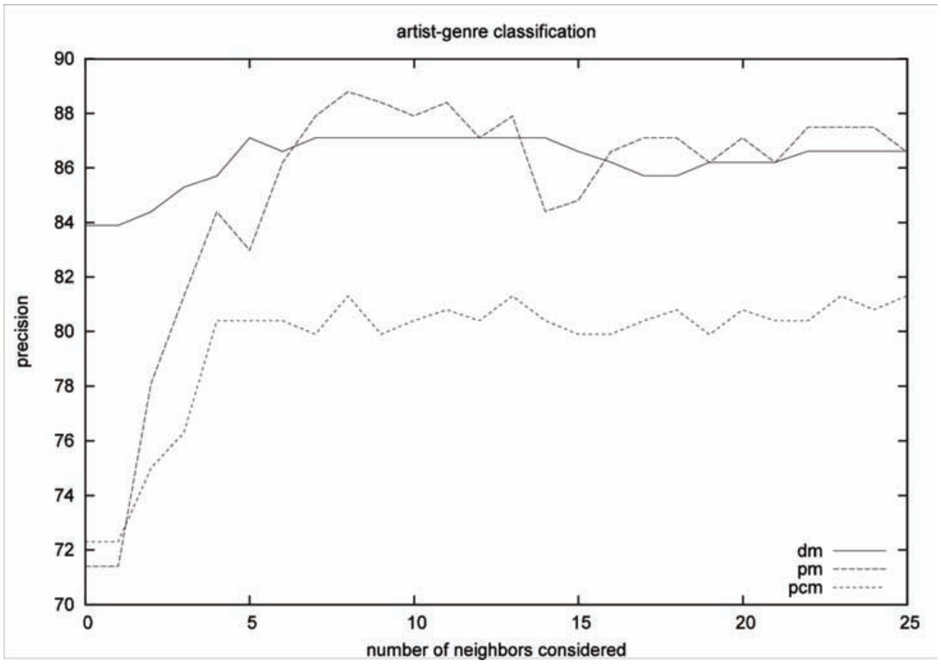
Using DM only few related artists can be found on the documents visited. This leads to a stable performance for the final mapping when expanding the list of related artists (Figure 2). That is, we only consider artists that co-occur at least once. Contrary to especially PCM, large numbers of *n* do not deteriorate the precision.

The performance of the pattern-based method strongly improves by considering related artists, the best performance is obtained for n = 8. All methods perform best for values of n between 5 and 13. The *Rock n' Roll* artists proved to be the

Table 12. Precision (%) without related artists and best precision per method

method	n=0	best	(corresponding *n*)
PCM	71.4	81.3	(13)
PM	72.2	88.8	(8)
DM	83.9	87.1	(5)

Figure 2. Precision (%) for classification of the musical artists



most problematic to categorize. The artists in the genres *classical*, *blues* and *jazz* were all correctly categorized with the best scoring settings. With the supervised music artist clustering method discussed in (Knees et al., 2004) a precision of 87% was obtained using complex machine learning techniques and a relatively large training set. In (Schedl et al., 2005) a precision of up to 85% was obtained using $O(N^2)$ queries. We can conclude that our simple and unsupervised method produces similar results. Moreover, we compute a classification of artists into genres instead of clusters of artists.

Painter Movement Classification

For this experiment, we constructed a list of painters I_a and a list of movements I_s in art using Wikipedia and map the two. From Wikipedia we extracted a set I_a of 1,280 well-known painters from the article *List of painters* and a set I_s of 77 movements in art from *List of art movements*[7]. We tested the performance of the algorithm on the subset of 160 painters who could be extracted from the Wikipedia pages describing movements (e.g. from the page on Abstract Expressionism). The other 1,120 painters are either not mentioned on the pages describing styles or are mentioned on more than one page. However, when computing similarities between the painters, we take all 1,280 painters into account.For the elements of I_s in this test no synonyms were added. For fairness, we excluded pages from the domain wikipedia.org in the search queries.

For PM, we selected learned patterns for the mapping between the elements in I_a and I_s. For learning, we used instance-pairs outside the test set. The best scoring patterns can be found in Table 13. For the relation between the instances in I_a, these patterns found were mostly enumeration patterns, e.g. *''including b and''*. The complete details of both experiments and the patterns used in PM can be found on the Web (**footnote**http://gijsg.dse.nl/Webconmine/). Due to the rareness of

Table 13. Best scoring learned patterns for painter—movement relation

“I_a I_s”
“I_s I_a”
“I_a and other I_s”
“I_a and I_s”
“I_a tog initiativ til I_s”
“I_a and the I_s”
“I_a surrealism I_s”
“I_a synthetic I_s”
“I_s artist I_a”
“I_a express I_s”
“I_a of the I_s”
“I_a uit de I_s”
“I_a experimenting with I_s”
“I_a arte I_s”
“I_s painter I_a”

some of the painters and names of movements, we did not use any additional terms in the queries for DM or PCM.

In Table 14 the performance of the preliminary mapping *m’* can be found for the three methods (n = 0). The experiments show that in general the use of related instances improves the classification (see Table 14 and Figure 3). It shows again that the methods with the lowest Google Complexity thus PM and DM perform better than PCM.

Although in the painter-movement experiment the number of categories identified (77) is much larger than in the previous experiment (16), the performance of PM and especially DM is still good. The results of PCM indicate that when the precision of the intermediate mapping is low (35%), the use of related instances does not improve the results. In this experiment we even observe a deterioration of the performance. Here DM clearly outperforms PM. This can be explained by the fact that using PM considerably less painter-movement pairs could be extracted. We expected the recall of PM to increase when applying stemming on the names of movements and the texts extracted (Porter, 1997). Although the number of pairs extracted slightly increases, the precision does not improve (Table 14).

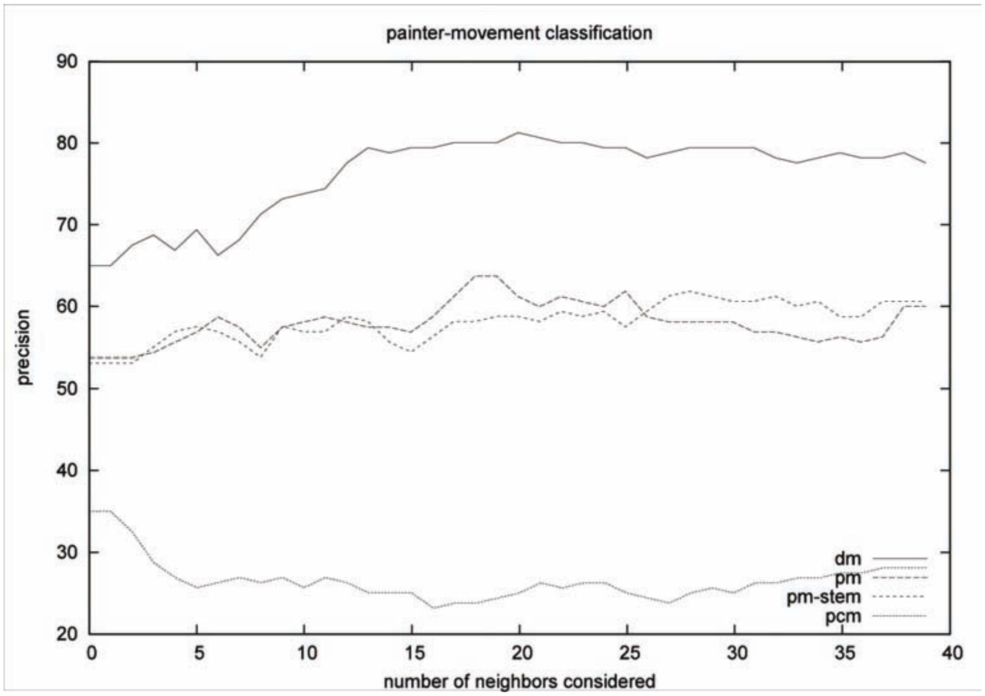
CONCLUSION

In Section 2 we have presented a framework algorithm for ontology population using queried expressions to mine factual information. We combine patterns expressing relations and an instance of a class into queries to generate highly usable search engine snippets. From these snippets we simultaneously extract instances of the classes and instance pairs of the relations. The results of the experiments are convincing. When submitting effective queries, the snippets provide enough data to populate an ontology with good recall and precision. The method is simple and easily adaptable to different domains. In the second part, we have discussed three alternative methods PCM, PM and DM to obtain co-occurrences of terms using a search engine. These methods are applied to gain a preliminary mapping between instances such as artists or painters and categories such as genres or art-movements. The distance between related instances is used to obtain a more reliable mapping. The three alternatives used have a different complexity with respect to the number

Table 14. Precision (%) without related instances and best precision per method

	Painter-Movement		
method	N = 0	best	(corresp. n)
PCM	35.0	35.0	(0)
PM	53.8	63.8	(18)
DM	65.0	81.2	(20)
PM-STEMMING	53.2	61.9	(28)

Figure 3. Precision (%) for classification of the painters



of queries to a search engine. The method using patterns and the one using complete documents are linear in the number of items in the sets of instances and categories, where the page-count-based mapping is quadratic. This distinction is important for classifying large sets of instances, since search engines allow only a limited amount of automated queries per day. We can precisely classify artists to genres, where the most efficient methods with respect to the Google complexity perform best. A second experiment consisted of the mapping of painters to their movements. This experiment underlines that the document-based and pattern-based method outperform the query-expensive method based on the number of search engine hits. We showed that simple and unsupervised methods can be used for a reliable classification. Using related instances indeed helps to improve the classification of instances. The experiments show an increase of the performance in both experiments. However, the Google count based method in painter classification shows that this additional step deteriorates the precision, if the classification is very unreliable.

FUTURE RESEARCH DIRECTIONS

We have shown that the pattern based approach in general gives access to highly relevant snippets. For future work, it is the process of identifying the instances within the snippets that needs further research.

The creation of precise rules is a laborious task, especially since each class requires its specific set of rules. The experiments in Part I showed that the challenge lies in the improvement of the precision of the recognition rules and the check functions.

The use of check functions based on enumeration of candidate instance has potential. Part II showed us that enumeration patterns are reliable to identify related instances. Now, if our ontology contains the instances *KFC* and *McDonald's,* we can formulate enumeration queries containing these instances and candidate instances. Fore example, the enumeration *'KFC, Chinese and McDonald's'* is not found by Google, where *'KFC, Burger King and McDonald's'* gives 31 hits.

With the known instances and the instances learned using the rules, we can automatically create an annotated corpus. This annotated corpus can be used to train a named entity recognizer. Such a named entity recognizer should be robust and consider only the local context of the named entities, as the snippets contain incomplete sentences. Approaches using only fully parsed sentences are therefore less usable. Memory-based learning (Daelemans & van den Bosch, 2005) is a technique that does not abstract from the learned instances. Moreover, only a small window of words is taken into account. If the information is redundantly available within the snippets, this technique may lead to satisfactory results.

The use of named entity recognizers based on machine learning also gives the possibility to compare the various approaches. When we collect the downloaded snippets, this comparison can be made on the same data.

With respect to the community-based data, in this chapter we assumed that the mapping was functional, i.e. only one category could be assigned to an instance. However, in some tasks multiple categories can apply to an instance. We therefore want to adapt the system such that multiple categories accompanied with a confidence mark can be assigned to an instance, analogue to the Web 2.0 concept of social tagging (O'Reilly, 2005). Moreover, methods can be exploited to learn other words related to some category, e.g. with the $tf \cdot idf$-approach (Knees et al., 2004; Manning & Schütze, 1999).

REFERENCES

Agichtein, E., & Gravano, L. (2000). Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM international conference on digital libraries*.

Balog, K., Mishne, G., & De Rijke, M. (2006). Why are they excited? Identifying and explaining spikes in blog mood levels. In *Conference companion of the 11th meeting of the European chapter of the association for computational linguistics (eacl 2006)* (pp. 207-210). Trento, Italy.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May). The Semantic Web. In *Scientific American.*

Blackburn, P., & Bos, J. (2005). *Representation and inference for natural language. A first course in computational semantics*. CSLI.

Brill, E. (1992). A simple rule-based part-of-speech tagger. In *Proceedings of the third conference on applied natural language processing (ANLP'92)* (pp. 152-155). Trento, Italy.

Brin, S. (1998). Extracting patterns and relations from the World Wide Web. *In Webdb workshop at sixth international conference on extending database technology (EDBT'98).*

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1-7), 107-117.

Brooks, C. H., & Montanez, N. (2006). Improved annotation of the blogosphere via autotagging and hierarchical clustering. *In Proceedings of the 15th international conference on world wide Web (www2006)* (pp. 625-632). Edinburgh, UK.

Brothwick, A. (1999). *A maximum entropy approach to named entity recognition*. Unpublished doctoral dissertation, New York University.

Buitelaar, P., Cimiano, P., & Magnini, B. (Eds.). (2005). Ontology learning from text: Methods, evaluation and applications,*123*. IOS Press.

Bunt, H., & Muskens, R. (Eds.). (2000). Computing meaning, *1*(73) Kluwer. Cafarella, M. J., Downey, D., Soderland, S., & Etzioni, O. (2005). Knowitnow: Fast, scalable information extraction from the Web. In *Proceedings of human language technology conference and conference on empirical methods in nlp* (pp. 563-570). Vancouver, Canada.

Chang, C.-H., Kayed, M., Girgis, M. R., & Shaalan, K. F. (2006). A survey of Web information systems. *IEEE Transactions on Knowledge and Data Engineering*, *18*(10), 1411–1428. doi:10.1109/TKDE.2006.152

Chinchor, N. A. (Ed.). (1998). *Proceedings of the seventh message understanding conference (muc-7).* Fairfax, Virginia: Morgan Kaufmann.

Cilibrasi, R., & Vitanyi, P. (2007). The Google similarity distance. *IEEE Transactions on Knowledge and Data Management*, *19*(3), 370–383. doi:10.1109/TKDE.2007.48

Cimiano, P., & Staab, S. (2004). Learning by Googling. *SIGKDD Explorations Newsletter*, *6*(2), 24–33. doi:10.1145/1046456.1046460

Crescenzi, V., & Mecca, G. (2004). Automatic information extraction from large Web sites. *Journal of the ACM*, *51*(5), 731–779. doi:10.1145/1017460.1017462

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th annual meeting of the association for computational linguistics (acl 2002).* Philadelphia, PA.

Daelemans, W., & van den Bosch, A. (2005). *Memory-based language processing*. Cambridge University Press.

Dave, D., & Lawrence, S. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the twelfth international World Wide Web conference (www2003)* (pp. 519-528). Budapest, Hungary: ACM Press.

de Boer, V., van Someren, M., & Wielinga, B. J. (2007). A redundancy-based method for the extraction of relation instances from the Web. *International Journal of Human-Computer Studies*, *65*(9), 816–831. doi:10.1016/j.ijhcs.2007.05.002

de Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the IEEE / ACL 2006 Workshop on Spoken Language Technology.*

Downey, D., Broadhead, M., & Etzioni, O. (2007). Locating complex named entities in Web text. In *Proceedings of the twentieth international joint conference on artificial intelligence (ijcai'07)*. Hyderabad, India.

Downey, D., Etzioni, O., & Soderland, S. (2005). A probabilistic model of redundancy in information extraction. In *19th international joint conference on arti_cial intelligence (ijcai'05)* (pp. 1034-1041). Edinburgh, UK.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification*. Wiley-Interscience Publication.

Dumais, S., Banko, M., Brill, E., Lin, J., & Ng, A. (2002). Web question answering: Is more always better? *In Sigir '02: Proceedings of the 25th annual international acm sigir conference on research and development in information retrieval* (pp. 291-298). New York, NY, USA: ACM Press.

Etzioni, O., Cafarella, M. J., Downey, D., Popescu, A., Shaked, T., & Soderland, S. (2005). Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, *165*(1), 91–134. doi:10.1016/j.artint.2005.03.001

Finkel, J. R., Grenager, T., & Manning, C. D. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl 2005).* Ann Arbor, MI.

Geleijnse, G., & Korst, J. (2005). *Automatic ontology population by googling*. In K. Verbeeck, Geleijnse, G., & Korst, J. (2006). Learning effective surface text patterns for information extraction. In *Proceedings of the eacl 2006 workshop on adaptive text extraction and mining (atem 2006*) (pp. 1-8). Trento, Italy: Association for Computational Linguistics.

Geleijnse, G., & Korst, J. (2007). Creating a dead poets society: Extracting a social network of historical persons from the Web. In K. Aberer et al. (Eds.), *Proceedings of the Sixth International Semantic Web Conference and the Second Asian Semantic Web Conference (iswc + aswc 2007), Busan, Korea, 4825 of Lecture Notes in Computer Science (LNCS),* pp 156-168). Heidelberg, Germany: Springer.

Geleijnse, G., Korst, J., & De Boer, V. (2006). Instance classification using co-occurrences on the Web. In *Proceedings of the iISWC 2006 Workshop on Web Content Mining with Human Language Technologies (Webconmine).* Athens, GA. (http://orestes.ii.uam.es/workshop/3.pdf)

Geleijnse, G., Schedl, M., & Knees, P. (2007). The quest for ground truth in musical artist tagging in the social Web era. In S. Dixon, D. Bainbridge, & R. Typke (Eds.), *Proceedings of the Eighth International Conference on Music Information Retrieval (ismir'07)* (pp. 525-530).Vienna, Austria: Austrian Computer Society.

Gusfield, D. (1997). Algorithms on strings, trees, and sequences: Computer science and computational biology. Cambridge, UK: Cambridge University Press.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics* (pp. 539-545). Nantes, France.

Hearst, M. (1998). Automated discovery of wordnet relations. In C. Fellbaum (Ed.), *Wordnet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Jin, Y., Matsuo, Y., & Ishizuka, M. (2006). Extracting a social network among entities by Web mining. In *Proceedings of the ISWC 2006 Workshop on Web Content Mining with Human Language Technologies (Webconmine).* Athens, GA.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing*. Prentice Hall.

Kim, S.-M., & Hovy, E. (2004). Determining the sentiment of opinions. In *Coling '04: Proceedings of the 20th International Conference on Computational Linguistics* (p. 1367). Morristown, NJ, USA: Association for Computational Linguistics.

Knees, P., Pampalk, E., & Widmer, G. (2004, October). Artist classification with Web-based data. In *Proceedings of 5th International Conference on Music Information Retrieval (ismir'04)* (pp. 517-524). Barcelona, Spain.

Korst, J., Geleijnse, G., De Jong, N., & Verschoor, M. (2006). Ontology-based extraction of information from the World Wide Web. In W. Verhaegh, E. Aarts, & J. Korst (Eds.), Intelligent Algorithms in Ambient and Biomedical Computing (pp. 149-167). Heidelberg, Germany: Springer.

Lin, D. (1998). Dependency based evaluation of minipar. In *Proceedings of the Workshop on the Evaluation of Parsing Systems at the First International Conference on Language Resources and Evaluation*. Granada, Spain.

Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. Cambridge, Massachusetts: The MIT Press.

McCallum, A. (2005). Information extraction: Distilling structured data from unstructured text. *ACM Queue; Tomorrow's Computing Today*, *3*(9), 48–57. doi:10.1145/1105664.1105679

Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Journal of Web Semantics*, *5*(1), 5–15. doi:10.1016/j.websem.2006.11.002

Mishne, G. (2007). *Applied text analysis for blogs*. Unpublished doctoral dissertation, University of Amsterdam

Mitchell, T. (1997). *Machine learning*. McGraw Hill.

Mori, J., Tsujishita, T., Matsuo, Y., & Ishizuka, M. (2006). Extracting relations in social networks from the Web using similarity between collective contexts. In *Proceedings of the 5th International Semantic Web Conference (iswc 2006)*, 4273, 487-500. Athens, GA: Springer.

O'Reilly, T. (2005). What is Web 2.0. (http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-Web-20.html)

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (acl 2005)* (pp. 115-124). Ann Arbor, MI.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (emnlp)* (pp. 79-86).

Porter, M. F. (1997). An *algorithm for suffix stripping. In Readings in information retrieval* (pp. 313-316). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Ravichandran, D. (2005). *Terascale knowledge acquisition*. Unpublished doctoral dissertation, University of Southern California.

Ravichandran, D., & Hovy, E. (2002). Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (acl 2002)* (pp. 41-47). Philadelphia, PA.

Schedl, M., Knees, P., & Widmer, G. (2005, June). A Web-Based approach to assessing artist similarity using co-occurrences. In *Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing (CBMI'05).* Riga, Latvia.

Sumida, A., Torisawa, K., & Shinzato, K. (2006). Concept-instance relation extraction from simple noun sequences using a full-text search engine. In *Proceedings of the ISWC 2006 Workshop on Web Content Mining With Human Language Technologies (Webconmine).* Athens, GA.

Tuyls, K., Nowe, A., Kuijpers, B., & Manderick, B. (Eds.). *Proceedings of the seventeenthbelgium-netherlands conference on artificial intelligence (bnaic2005)* (pp. 120-126). Brussels, Belgium: Koninklijke Vlaamse Academie van Belge voor Wetenschappen en Kunsten.

Véronis, J. (2006). Weblog. (http://aixtal.blogspot.com)

Voorhees, E. (2004). Overview of the trec 2004 question answering track. In *Proceedings of the 13th Text Retrieval Conference (trec 2004).* Gaithersburg, Maryland.

Zadel, M., & Fujinaga, I. (2004, October). Web services for music information retrieval. In *Proceedings of 5th International Conference on Music Information Retrieval (ismir'04)*. Barcelona, Spain.

Zhou, G., & Su, J. (2002). Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (acl 2002)* (pp. 473-480). Philadelphia, PA.

ADDITIONAL READING

A solid introduction into the field of statistical natural language processing is the book by Manning and Schütze (1999). It is well readable as it contains both chapters discussing elementary mathematical notions and the linguistic essentials. Sentence parsing, word sense disambiguation and term identification are some of the introduced topics that are relevant for Web Information Extraction. An alterative is (Jurafsky & Martin, 2000).

In (2005) McCallum introduces the problem of information extraction to a broad audience. A list of relevant papers is provided for further reading as well as a number of URLs of tools for information extraction, for instance the GATE framework (Cunningham, Maynard, Bontcheva, & Tablan, 2002). Not named by McCallum, but nevertheless interesting is the KnowItAll project by the group of Oren Etzioni (2005). In one of the KnowItAll papers, (Downey et al., 2005), the redundancy of information on the Web is exploited to identify instances on the Web without supervision. The same paradigm is the basis of (De Boer et al., 2007) where relations between painters and art movements are learned.

In the second part of this chapter, the tagging of items is discussed. Previous work addresses the 'auto tagging' of Weblogs (Brooks & Montanez, 2006; Balog, Mishne, & De Rijke, 2006; Mishne, 2007). Given a post on a Weblog, the question is which tags are appropriate labels for this post.

In this work, we used a set of rules to identify the instances within the texts. However, when an annotated training set is available (or can be created), instances can also be recognized using machine learning. In (Manning and Schütze,1999) a thorough introduction to Hidden Markov Models is given and their applications in natural language processing. Daelemans and Van den Bosch use Memory-based learning to identify named-entities (2005). A broad overview of machine learning and pattern classification techniques can be found in (Duda, Hart, & Stork, 2000).

The above mentioned approaches are based on statistical techniques. In (Blackburn & Bos, 2005) and (Bunt & Muskens, 2000) models are discussed where the meaning of a texts can be represented in a formal language. A field closely related to Information Extraction is Question Answering. Given a corpus, answers to questions like 'Who is the Mayor of San Francisco?' need to be answered. In (Ravichandran, 2005; Dumais, Banko, Brill, Lin, & Ng, 2002) question answering systems are presented that use the Web as a corpus.

ENDNOTES

1. The question-answering services of http://www.google.com or http://www.ask.com do not provide answers to these (simple) questions.
2. http://imdb.com
3. http://w3c.org
4. We use the $[c_k]$ notation to denote a variable instance of class c_k
5. http://www.whopper.com
6. http://www.cp.jku.at/people/knees/publications/artistlist224.html
7. www.wikipedia.org Both pages visited April 2006

APPENDIX: QUESTIONS FOR DISCUSSION

1. Why is it hard to compare two different approaches in ontology population from the Web? (The obtained query results differ.)
2. Alike the SW languages (introduced in chapter X by Cardoso), we consider binary relations between instances. Show that these can also be used to express associations with more than two instances.
3. Suppose that you have created an ontology with the class City. How would you identify instances of this class? How about instances of the class Texan Cities, cities within the state Texas?
4. What is the Google Complexity of the ontology population method applied on the movie ontology in Section 2.6?
5. In the second part of the Chapter, relations are given a score. How would you model this scoring using the SW languages, cf. (Mika, 2007)?
6. Suppose you are interested in a ordered list of all American presidents from George Washington to George W. Bush. Formulate classes and relations that describe this information demand. What would be appropriate patterns to populate this ontology?
7. KnowItAll is a system that extracts information from the Web from both structured sources (e.g. tables and lists) and unstructured Web texts. Can you think of information demands where the extraction of data from tables is less usable?

This work was previously published in Semantic Web Engineering in the Knowledge Society, edited by J. Cardoso and M. Lytras, pp. 208-241, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 7.13
Profiling of Web Services to Measure and Verify their Non-Functional Properties

Witold Abramowicz
Poznań University of Economics, Poland

Monika Kaczmarek
Poznań University of Economics, Poland

Dominik Zyskowski
Poznań University of Economics, Poland

ABSTRACT

The following chapter focuses on the problem of the proper definition of non-functional properties and methods that may be applied in order to estimate their values. First of all, a reader is familiarized with the concept of non-functional properties and different views on the quality of Web services. Then, selected approaches to obtain values of non-functional properties are presented. The focus of attention is Web services profiling that seems to be one of the most promising methods to perform this task. The framework presented in this chapter was implemented and tested within the EU Adaptive Services Grid project.

DOI: 10.4018/978-1-60566-042-4.ch005

INTRODUCTION

The paradigm of service-oriented architecture (SOA) is currently one of the most popular approaches followed by modellers and IT developers in order to build IT systems. The potential of SOA is being strengthened by the growing popularity of Web services technology. Web services allow for encapsulation of business functionalities provided using IT infrastructure and easy integration with other systems via standard communication protocols. Globalization, along with a tendency to outsource some activities, boosts the exploitation of Web services in enterprise applications.

One of the essential elements needed to ensure the success of Web services (as well as Semantic

Web services) technology is a proper Web service description to be used not only in order to invoke a Web service, but also to discover it and perform composition. The common agreement is that a Web service should be represented by its surrogate, describing its functional, non-functional, and behavioural characteristics. The functional features focus on what a Web service does, the non-functional ones on how it does it, and the behavioural ones inform us which parties are involved in the process of service provisioning. Whereas, there seems to exist a common agreement on how the description of the functional and behavioural properties of a service should look like and its role in Web services' interactions; there is still an ongoing discussion about the scope and the methods that should be used to express the non-functional side of a Web service.

Within the last few years, a number of different approaches to define non-functional properties and quality of service (QoS) models for Web services have been proposed. Each of them defines a different set and understanding of non-functional properties as well as QoS features. Yet, as shown in this chapter, these initiatives are still not mature enough as they focus mostly on the technical aspects of a Web service and, in most cases, disregard the business ones. Another problem that appears, once the model of non-functional properties is defined, relates to the methods that can be utilized in order to obtain values of defined properties. Various techniques to carry out this task were proposed. They differ in terms of reliability, trustworthiness, as well as the issue of continuous provisioning of up-to-date values of parameters.

The following chapter focuses on the problem of definition of non-functional properties and methods to estimate their values. The special focus is assigned to the Web services profiling, being, in our opinion, one of the most promising methods to perform this task. The chapter is organized as follows. First, we present our understanding of non-functional properties and quality of service. In the following section, the methods to compute values of non-functional properties are discussed. The following section outlines current methods to compute values of non-functional properties. Then, the concepts of service profiling, service profile, and its elements are presented. Moreover, in this section we also describe the technical details of service profiling system implemented within the Adaptive Services Grid project. Finally, the summary follows.

NON-FUNCTIONAL PROPERTIES OF WEB SERVICES

To fully exploit the advantages of Web services technology, as indicated in the introduction section, their proper description is required. The common agreement is that a Web service should be represented by its surrogate, providing information on its functional (what a Web service does), non-functional (how it does it), and behavioural characteristics (which parties are involved in the process of service provisioning). However, before the initiatives in the area of Web services description are presented, differences between a service and a Web service, being crucial for our further discussion, need to be mentioned.

A service is usually defined as a provision of value in some domain (Preist, 2004) or seen as a business activity that often results in intangible outcomes or benefits (Baida, Gordijn, Omelayenko, & Akkermans, 2004). Let us consider an example of a person who wants to buy a book on knowledge representation published by Springer. The service this person is looking for is the provisioning of books with the specified constraints. A provision is in fact independent on how the supplier and provider interact (Lara & Olmedilla, 2005), that is, it does not really matter at this point whether the requester goes to a bookshop or uses the Amazon.com portal to buy the book of interest.

A Web service in turn may be defined as a computational entity accessible over the Internet (using particular standards and protocols) (Preist, 2004). The focus is assigned here to the way that the requester and provider interact with each other (Lara & Olmedilla, 2005). In the considered example with a book, a bookshop (e.g., Amazon.com) may provide a component accessible via Web service standards (i.e., a Web service) to request the book. Therefore, a Web service is a technical implementation, an interface to a real-world service defined as a certain activity undertaken on behalf of a certain entity. Final users are in fact more interested in the real service they get, rather than in the interface itself. Hence, Web services may be considered as two inseparable parts: a technical interface (described using Web service description language (WSDL) and a real (business) functionality (described by other means) the interface provides access to.

The above perception of a service and Web services entails a question what kind of a Web services description is required for the needs of interactions with and between Web services. A Web service is an interface used to request the actual provisioning of a real-world service fulfilling the requester needs. Therefore, for example, in order to discover a Web service and use it to request a required service, the technical description of a Web service (the interface) is of course crucial but not sufficient. What is also indispensable is the description of a real world service and non-functional properties of both a service and a Web service. The way consumers interact with traditional services and their requirements regarding their description are a result of social and economic interactions that have been taking place for many years. If Web service providers do not consider this fact, they will fail. Therefore, a Web service description should adhere to the well-established requirements of the consumers and cover not only functional, but also non-functional properties (NFP) of a service (Abramowicz, Kaczmarek, & Zyskowski, 2006), (Abramowicz, Kaczmarek, Kowalkiewicz, & Zyskowski, 2005).

Reviewing the Web service description initiatives at least two things should to be taken into account, namely the scope of such a description and formalism used to express it. There are many initiatives in the area of service description. The earliest ones, like WSDL (W3C, 2007), focused mainly on purely technical details needed in order to invoke a Web service (such as its interface, ports, bindings, etc.) and were expressed using XML notation. Then the other initiatives, like UDDI registry (UDDI, 2004), adding few non-functional properties to the Web service description followed. Finally, the semantic initiatives like OWL-S (W3C, 2004), WSMO (Roman et al., 2006), or SAWSDL (Farrel & Lausen, 2006) expressed using the logic-based language, such as, for example, Web services modeling language (WSML) or resource description framework (RDF) and trying to also capture information on the real-world service standing behind the Web service interface were undertaken.

Having a look at those initiatives, one thing may be quite easily noticed. There seems to exist a common agreement on how the description of the functional properties of a service should look like and its role in Web services interactions. However, there is still an ongoing discussion on the scope and methods that should be used to express the non-functional side of a service.

Functional properties represent the capability of a Web service. These properties are mainly related to the input and output parameters as well as constraints/a state of the world before and after service execution. Therefore, in most cases either the functionality is expressed as only information on inputs and outputs (like in WSDL where input and output parameters required by a service are defined) or as the semantically annotated quadruple IOPE (inputs, outputs, preconditions, and effects) in OWL-S or pre- and post-conditions defined within WSMO. The functional properties are used mainly for the needs of discovery and composition of Web services. The mechanisms operating on all of the above mentioned formalisms are already implemented and work more or

Table 1. A few exemplary non-functional parameters

Parameter name	Definition
execution price	an amount of money that needs to be paid to the service provider for service execution
latency time	a round-trip time between sending a request and receiving a response
average (maximum) response time	an average (maximum) time needed for the packet of control data to get to the provider's server (where the service is executed) and then return to the requester
robustness	ability of a service to act properly if some of the input parameters are missing or incorrect
availability	probability whether a service is capable of processing the client's request or not at a certain time
charging method	by execution unit, subscription, or by data chunk size and so forth.
payment method	information on method of payment (wire transfer, etc.)

less efficiently in various projects (Abramowicz, Haniewicz, Kaczmarek, & Zyskowski, 2006a; Kuster, Koenig-Ries, Stern, & Klein, 2007; Liu, Peng, & Chen, 2006).

In turn, the non-functional properties play a crucial role in almost all service interactions (to mention only selection, discovery, and filtering). Non-functional properties of a service may be defined as anything that exhibits a constraint over the functionality (O'Sullivan, Edmond, & Hofstede, 2002). In fact, non-functional parameters are distinctive criteria for the success of businesses offering their services using Web services technology. They allow differentiating between Web services offering the same (or quite similar) functionality, as, in most cases, service substitutes differ when it comes to the values of specific non-functional properties. Their role became even more important, as nowadays Web services are not only used internally, but also support collaboration between various organizations. In consequence, final users (especially business users) desire to know in advance the real quality and non-functional properties of external services they are to use.

The non-functional parameters may be represented as qualitative and quantitative measures of a Web service (or a service). The nonquantitative ones include security or transactions, whereas quantitative ones include such attributes as cost or time. NFP should of course include business constraints and interservice dependencies, if possible. However, different types of services require different properties describing them and which properties are necessary depends on the domain, intended use, and users' requirements. If services are to be used to automate B2B and B2C models, they have to be described in a proper manner and meet specific business requirements. The table below presents a few exemplary non-functional parameters.

The non-functional model for Web services is still under development. Each of the already mentioned service description initiatives or standards like WSDL, UDDI, OWL-S, WSMO, or SAWSDL treats non-functional properties in different ways. No non-functional properties can be expressed using WSDL. A list of non-functional parameters provided by UDDIs includes only some attributes such as, for example,: provider name, service name, and category. In turn, OWL-S and WSMO take into account a wider range of NFP (than, for example, UDDIs), including not only information on service providers, but also some performance-related information, such as execution time and so forth. The short overview of non-functional aspects supported by the selected Web services description approaches is presented in the following table.

The lack of a real support (i.e, languages, methodologies, tools) for non-functional properties may result from the following issues (Eenoo,

Table 2. Overview of the support of the selected Web services description approaches to NFP

Web services description	Approach to NFP
WSDL	Nonfunctional properties are neglected
UDDI	Defines a set of non-functional properties of a service provider (included in BusinessEntity) such as: address, phone number, e-mail address, and some meta data about a service as, for example, service category
OWL-S	Includes the following non-functional properties: service name, text description, quality rating, and service category; all are stored in the ServiceProfile class. The list may be extended using the ServiceParameter from the ServiceProfile class.
WSMO	It recommends a set of NFP for each element of a Web service description (e.g., contributor, creator, date, and so forth) provided by the Dublin Core Metadata Initiative. WSMO does not provide a model for the non-functional properties of a service (Toma, 2006), but there is an on-going work in this direction.
O'Sullivan's approach	(O'Sullivan et al. (2002) describe a set of the most relevant non-functional properties for Web services and their modelling. Exemplary concepts considered are as follows: service provider, locative model, temporal model, service availability, obligations, price, payment, discounts, trust, security, and so forth.

Hylooz, & Khan, 2005; Rosa, Cunha, Freire, & Justo, 2002; Toma, 2006):

- Non-functional properties are usually too abstract and most of the time they are stated informally;
- In some cases there is no clear distinction between the functional and non-functional aspects of a service;
- Non-functional properties are often considered to be represented only after the functional and behavioural have been described;
- Non-functional properties very often conflict and compete with each other (e.g., availability and performance);
- Complexity of modelling non-functional properties (difficult to formalize).

DUALITY IN WEB SERVICES QUALITY

Quality of a Web service may be defined as an extent to which a Web service provisioning process as well as delivered results meet expectations of a user. It is a subset of non-functional properties of a Web service. When considering the quality concept of Web services their dual character (the relation to the real world services) should be taken into account. Therefore, in defining quality of service one needs to consider two aspects: the quality of a Web service implementation (the interface), and the quality of a real service (available through a Web service interface). That is why a QoS concept for Web services should be divided into two separate groups, namely, quality of execution (QoE) and quality of result (QoR) as proposed by Abramowicz, Filipowska, Kaczmarek, Kaczmarek, Kowalkiewicz, Rutkowski et al. (2006). Let us consider a simple route planning service which is a part of the Adaptive Services Grid project's demonstration scenario (Noll, 2004).

QoR of this service may be defined as:

- A service provider's statement on the overall quality of the result provided by the service (low/high, etc.). It should be understood as follows. To what extent should a user trust the provided route description? Is it adequate and reliable? Does it show exactly what a user wanted? Will a user following the obtained route reach the desired destination point? Other parameters may be also considered (e.g., the route description, the resolution and colours of the image, and accuracy). These characteristics have a crucial impact on the satisfaction of

Table 3. RoutePlanning service example (Noll, 2004)

Service name	**Route Planning Service**
Description	Creates a route description for the customer's coordinates and the given attraction. The route description consists of a coloured high-resolution picture and a textual description.
Nonfunctional Properties	
Service Name	Map24RoutePlanningService
Provider Name	Map24.de
Information Quality	High
Functional Properties	
Preconditions	Location ls, Location lg
Positive Effects	RouteDescription rd, hasRoute (rd, r)

a user and are a part of the QoR concept.

- Users' feedback (assessment) understood as their satisfaction from the returned result (not from the interface through which they communicate with a service) expressed in a defined scale. However, it would be very difficult, if not impossible, to collect such information from users. They would rather provide an overall evaluation of both a service implementation and real service effects.

The QoR concept is domain specific. In fact, it is very difficult, if not impossible, to define a measure that would hold for all possible services. It is not the case with QoE, which is independent of the domain and rather easy to compute. The quality of execution relates to the underlying technology (i.e., technical and network-related aspects). The following properties may be a part of the QoE model:

- **Response latency:** Time needed for the control data to get to the service and back to the client.
- **Maximal throughput:** How many requests a provider is able to process in a given time period.
- **Execution duration:** Time needed to fulfil a user request (time between sending a request and receiving an answer).
- **Execution price:** Amount of money a user needs to pay in order to use an interface to the service.
- **Service robustness:** The ability of a service to act properly if some of the input parameters are missing or incorrect (e.g., the wrong coordinates or incorrect data types, etc.).

The following table summarizes our short discussion on the differences between the QoR and QoE concepts.

Another aspect that needs to be mentioned is

Table 4. Comparison of QoR and QoE

Quality of Result	**Quality of Execution**
Quality of a real service	Quality of an interface (WS implementation)
Domain specific	Domain independent
Very hard to measure and monitor	Rather easy to measure and monitor
In most cases has no impact on QoE	May have an impact on QoR

the difference between an execution price and a service price. A service price is the amount of money a user has to pay for a real service; for example, when using a route planning it is a price of the attraction ticket (e.g., ticket to the cinema) (it influences QoR). In this case, an execution price is the amount of money we have to pay for using the interface to book tickets, not a price of the ticket itself (it influences QoE). When buying a book at Amazon.com, the execution price is 0 (using the Amazon Web page to search and order is free), but the service price is the price of the book and the delivery costs. In case of information services (such services where output returned by a service is equal to the effect we wanted to obtain) it is rather unclear whether the price we have to pay for the information is a service price or execution price, and the classification may depend on many factors.

Most of the current initiatives aiming at providing definitions and descriptions of quality dimension address only some generic parameters (mostly network related), such as execution price and duration, availability and reliability, and so forth (Liu, Ngu, & Zeng, 2004; Menasce, 2002; Zeng, Benatallah, Dumas, Kalagnanam, & Sheng, 2003), and do not differentiate between the QoR and QoE concepts. More parameters, considering also QoR, are presented by O'Sullivan et al. (2002), but they are not widely used in practice. Moreover, QoR properties are not considered in most of the methods trying to compute the values of non-functional properties. Therefore, in the remaining part of this chapter, whenever a reference to QoS is made, it refers to those quality parameters of a service that are computable (therefore, in most cases they exclude QoR parameters). Whenever a clear differentiation needs to be made between quality of result and quality of execution, respective terms are used.

METHODS AND APPROACHES TO DERIVE VALUES OF NON-FUNCTIONAL PROPERTIES

The simplest way to derive values of NFP is to rely on service providers advertising this information. However, taking directly the values advertised by a service provider is not advisable. It requires users to trust the accuracy of the values declared by service providers. However, service providers do have an interest in overestimating NFP of their services, so a solution allowing measurement of (programmatically) the values of NFP for verification purposes is needed. Moreover, values of non functional parameters are often assumed to be constant in time and space (service location), but they may change, depending on the details of the service request, execution environment, and so forth. For example, the response time of a Web service may be less than 5 minutes during the working days, but during the weekends, it may be less than 1 minute as the interest in the particular service decreases.

To avoid the problems of accuracy of non-functional properties' values given by service providers, some other methods to derive (or verify) their values are needed (Abramowicz et al., 2005). Ran (2003) proposes a QoS model using a QoS certifier to verify published QoS criteria. The approach requires all Web services providers to advertise their services with the QoS certifier. However, this approach does not take into account the dynamism of the environment and the fact that the values of a Web service change in time. The approach does not provide, for example, methods to update the QoS values automatically and it lacks the details regarding the verification process.

Sheth, Cordoso, Miller, and Kochut (2002) propose a QoS middleware infrastructure that requires a built-in tool to monitor metrics of NFP automatically. Such an approach requires the willingness of service providers to give up some of their autonomy. It may also require service providers to cover execution costs. Moreover, if the polling interval is set to too long, the QoS will not be up-to-date. If the polling interval is set to too of a short time, it might incur a high performance overhead. A similar approach emphasizing a service reputation, is proposed by Maximilien and Singh (2002a, 2002b).

Another approach obtains information on values of QoS parameters from the users themselves. When collecting quality information from the users feedback, each user is required to evaluate QoS (and at the same time QoR) of the consumed service. The main advantage of this approach is that QoS values can be computed based on the real user experience (up-to-date runtime execution data). The main disadvantage is the fact that a user judgment is not objective; users use different definitions of quality, have different past experiences, and so forth.

In other approaches called "*a'posteriori approach*" (Casati, Castellanos, Dayal, & Shan, 2004) QoS values are solely collected through an active monitoring. The monitoring can be performed by a user, service broker or platform, dedicated QoS registry (Kuropka & Weske, 2006; Liu et al., 2004), or an already mentioned QoS certifier (Ran, 2003). The data are collected from the actual consumption of a service and therefore are accurate and objective. One avoids the necessity to install rather expensive middleware in order to constantly check large numbers of service providers. However, there is a high overhead since QoS must be constantly checked for a large number of Web services. On the other hand, the approach that relies on a third party to rate or endorse a particular service provider is expensive and static in nature.

When the service related data collection is envisioned through, for example, workflow monitoring or user feedback, another important issue is how to compute the values of quality-related parameters from the collected data. There are a few initiatives to solve the problem. One of them (Maximilien & Singh, 2004) suggests performing an analysis of past executions of atomic and composite services by using data mining and workflow log mining techniques. Moreover, some statistical methods can be applied as well (Liu et al., 2004).

Workflow management systems are a very important infrastructure for complex applications. They usually register the start and completion of activities as well as other events that occur during execution. This information is stored as workflow log files (Aalst, Zhang, Shanahas, & et al., 2003) that further are processed using workflow and process mining techniques. The goal of workflow mining is to find a workflow model on a basis of a workflow log (Aalst et al., 2003). In turn, process mining is a method of distilling a structured process description from a set of real executions (Aalst et al., 2003). Many methods to perform these tasks were developed (e.g., probabilistic workflow mining, or Petri nets [Aalst et al., 2003]) and may be successfully applied also to the Web services area.

In the next section, the Web services profiling, being an alternative method to derive the values of non-functional properties of a Web service, is presented.

WEB SERVICE PROFILING, SERVICE PROFILE, AND ITS ELEMENTS

Service profiling is a process of computation of values of non-functional properties. The main goal of service profiling is to create service profiles of atomic and composite services. A service profile may be defined as an up-to-date description of a selected subset of non-functional properties of a service. It not only characterizes a service but also allows for services comparison based on aggregated values of non-functional parameters and, in consequence, selection of a service most suited to the requirements of a user.

In order to compute the values of non-functional properties, service profiling needs first to collect information on services executions, aggregate it, and then derive required information. The raw data may come from multiple data sources. Every source has its own specific purpose and provides different information. The following possible sources of information that further feed the profiling system with appropriate data may be distinguished: service registries, monitoring data, data coming from service level agreements (SLA) storing information on contracted QoS values, feedback from service consumers about obtained service quality, and so forth.

The aim of the Web services profiling is to perform fair and open NFP computation. Therefore, as the service execution history data are the most objective and reliable source of information on the service, they are in fact the primary source of information. The Web services profiling does not perform only the core workflow mining. It analyses log files in order to obtain data needed for the profiling process, but, in addition, it takes advantage of the raw data collected from service properties defined in SLA, published by service providers, and obtained from users' feedback. For instance, it compares contracted values from SLA against these from execution. In consequence, it is possible to check to what extent the agreement between a provider and a consumer is fulfilled. Moreover, appropriate algorithms may discover which values of particular parameters are, for example, likely to be guaranteed by providers.

Service profiling is, in our opinion, a trustworthy method of service quality measurement. It does not rely on providers' declarations about quality of their services. Statistical procedures used to compute values, data coming from execution logs, and so forth, assure high reliability of results of service profiling. The information declared initially by a service provider might be verified by what is stated in SLA, being approved by its provider and then by the results of the analysis of execution data. This kind of verification increases the reliability of our mechanism and we do not need a third party to verify the correctness of the values of profile parameters as procedures are transparent and parameters precisely defined. In addition, a service profiling mechanism is generic (a number of parameters it operates on may be easy modified) and independent of the service description provided by a service provider.

Service Profile

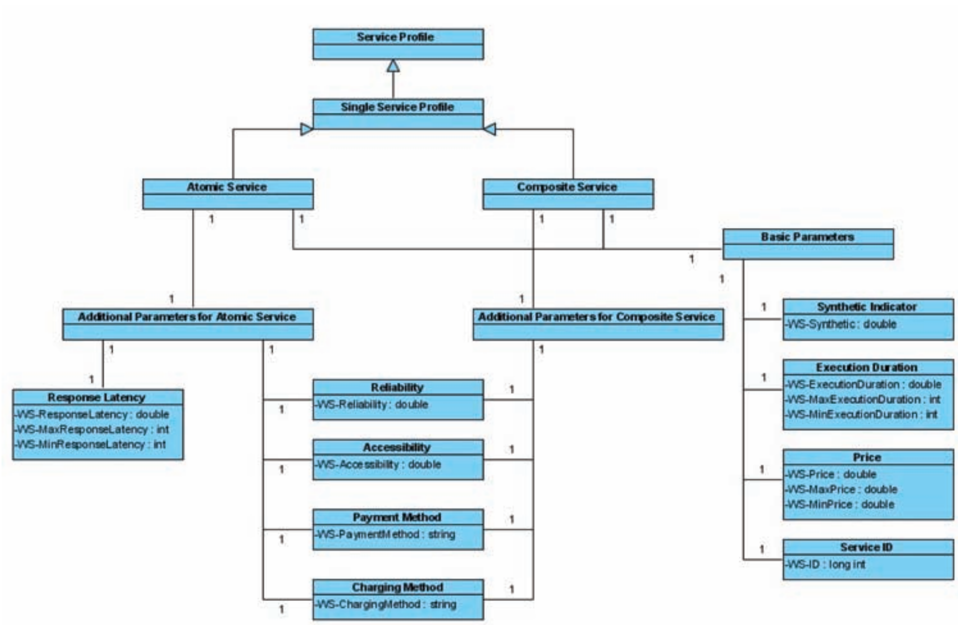
As already stated, a service profile may be defined as an up-to-date description of a subset of non-functional properties of a service. It allows for services comparison based on non-functional parameters and selection of the service most suited to the requirements of a user.

In order to create an adequate service description one needs to consider that the collected or derived data, taken into account by a service profiling mechanism, may differ in terms of its stability in time. Regarding the type of information on services, we can distinguish three main categories:

- **Static Information:** Values of service properties that do not change over time, such as name of the service, and are provided by a service provider.
- **Semistatic information:** Values of service properties that may change over time, such as quality of service and price. This information changes periodically, but not very often.
- **Dynamic Information:** Values of service properties that may be (and usually are) different in every execution of the service. It relates mainly to the network related quality of service.

From the profiling point of view, the most interesting parameters are the dynamic and semistatic ones. In addition, parameters that are estimated and finally included in a service profile may be simple reflections of service behaviour or adequately aggregated to show an overall quality of a service. Therefore, we consider two groups of non-functional properties:

Figure 1. Service profile structure - class diagram



- **Simple Properties:** Values of service properties that can be monitored on an individual level. This is mostly information presented in service level agreements. Such properties may include, for example, latency time, execution cos,t and so on.
- **Derived Properties:** Where additional manipulation is needed (performed by a service profiling system). Such properties may include reliability, availability, or, in our case, a synthetic indicator.

Our belief is that a service profile should be easily interchanged between building blocks of SOA systems. In order to allow for simple messaging and processing of profiles, we decided to represent them as XML documents. The greatest advantage of this solution is that XML schema is easily verifiable and interpretable by machines. A standardized form of a service profile makes it easy to be adapted in industrial applications.

Because of flexibility of service profiling, the set of parameters included in a profile may vary due to different quality parameters considered in different IT systems. The exemplary structure of a profile (as seen in Figure 1) was derived based on the requirements defined in the already mentioned ASG project.

The excerpt of a service profile schema is presented in the Listing 1. Please note that for some parameters, average, minimal, and maximal values are determined. These values may be helpful when a user precisely expresses the user's needs on quality parameters. Therefore, a user may specify that the user is looking for a service where parameters meet accurately expressed criteria.

Additionally, a service profiling system may offer provider profiles that show how, in general, services of a given provider behave. They may be useful to represent the overall quality of services provided by a concrete provider. These profiles are more quality-oriented, whereas service profiles are more performance-oriented. In this case, quality orientation means that time-related QoS parameters are less important than the fact whether a given service was accessible or produced expected results.

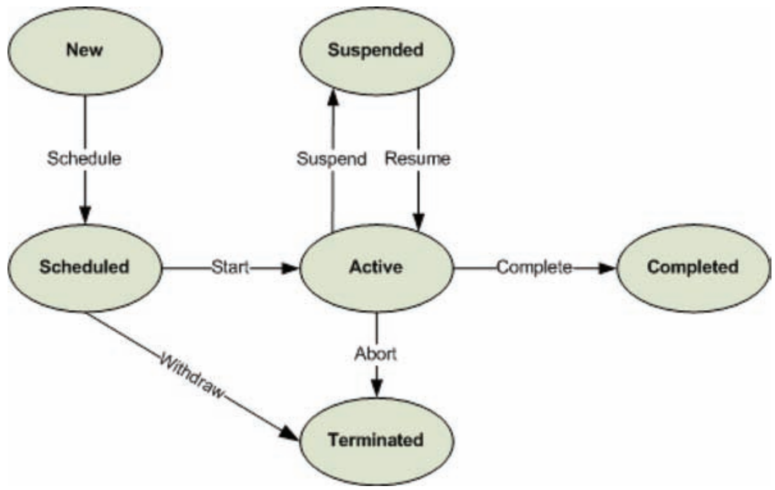
Figure 2. Listing 1: Excerpt of exemplary service profile schema

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="ProfileData">
<xs:complexType>
<xs:element name="ServiceProfile" use="required">
<xs:complexType>
<xs:element name="BasicData" use="required">
<xs:complexType>
<xs:sequence>
<xs:element name = "WS-ID" type = "xs:string" use="required"/>
<xs:element name = "WS-Price" type = "xs:float" use="required"/>
<xs:element name = "WS-MinPrice" type = "xs:float" use="required"/>
<xs:element name = "WS-MaxPrice" type = "xs:float" use="required"/>
<xs:element name = "WS-ExecutionDuration" type = "xs:float" use="required"/>
<xs:element name = "WS-ExecutionDurationFulfilment" type = "xs:float" use="required"/>
<xs:element name = "WS-MinExecutionDuration" type = "xs:positiveInteger" use="required"/>
<xs:element name = "WS-MaxExecutionDuration" type = "xs:positiveInteger" use="required"/>
<xs:element name = "WS-Synthetic" type = "xs:float" use="required"/>
<xs:element name = "WS-SlaFulfilmentIndicator" type = "xs:float" use="required"/>
</xs:sequence>
</xs:complexType>
</xs:element>
<xs:element name="AdditionalData" use="required">
<xs:complexType>
<xs:sequence>
<xs:element name = "WS-PaymentMethod" type = "xs:string" use="required"/>
<xs:element name = "WS-ChargingMethod" type = "xs:string" use="required"/>
<xs:element name = "WS-Accessibility" type = "xs:float" use="required"/>
<xs:element name = "WS-Reliability" type = "xs:float" use="required"/>
<xs:element name = "WS-ResponseLatency" type = "xs:float" use="optional"/>
<xs:element name = "WS-ResponseLatencyFulfilment" type = "xs:float" use="optional"/>
<xs:element name = "WS-MinResponseLatency" type = "xs:positiveInteger" use="optional"/>
<xs:element name = "WS-MaxResponseLatency" type = "xs:positiveInteger" use="optional"/>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:schema>
```

Service Profile Computation

The most popular information sources for service profiling are execution logs. These log files usually have a strictly defined structure (Aalst et al., 2003), so the automated processing of them is feasible and algorithms are rather straightforward. For example, the execution duration may be easily counted as a difference between end time and start time of a service execution (these values are stored in the log file). Of course, to compute the values of other parameters other methods may be required. For instance, in order to compute the value of a reliability parameter, a profiling system needs to keep track of service execution states. In our approach, we consider the finite state machine of Web service transitions as shown in the Figure 3.

Figure 3. Types of Web services events. Based on Aalst et al. (2003)



Therefore, it is possible to determine the number of started services that were completed. Thus, the assessment of reliability parameter is not a problem. A similar approach is used for accessibility parameter computation. For more details please refer Kowalkiewicz, Ludwig, Kaczmarek, and Zyskowski (2005). In the Table 5 we present an exemplary set of non-functional properties and outline methods of their computation.

When creating a service profile the time horizon is taken into account. A user may need a particular instance of a service only once in a given point of time or may need to use the service a few times in a given time period. Therefore, the horizon of the prognosis should be considered. In the first case, short-time information about a service is important, and in the second case, more attention should be paid to the long-term behaviour of a service, taking into account also more historical data.

Another challenging issue is the set of non-functional parameters that should be used to describe composite services and the way to compute values of these parameters. The possible solutions may be found presented by Liu et al. (2004), Maximilien and Singh (2004), and Zeng et al. (2003). They suggest using a similar set of attributes, as for atomic services and computing their values using statistical methods.

Composite service profiles are the aggregations of atomic service profiles. A description of a composite service profile is very similar to a service profile, because it treats a composite service like an atomic one. That is why the structure of its profile does not differ significantly from the profile of an atomic service. However, the values of some

Table 5. Some parameters of service profile and their computation methods

Parameter name	Computation method
Execution duration	Difference between end and start time of service execution
Accessibility	Number of successful invocations divided by all the invocations in a given time period
Reliability	Number of successful executions divided by all of the executions in a given time period
Price	Average price in a given period of time
Synthetic indicator	Statistical aggregation of all considered parameters denoting an overall quality of a service

parameters are computed as statistical measures based on characteristics of atomic services included in the composed service. Moreover, not all parameters that are computed for an atomic service profile are included in composite service profiles. For example, the response latency value is only computable for atomic services.

In order to compute a value of quality parameters of a composite service we can proceed twofold:

- The execution log mining may be performed in order to compute values of parameters using methods similar to these for atomic services;
- A composite service execution plan may be used to compute *hypothetical value* of quality parameter. Such plans are usually described using business process execution language for Web services (BPEL4WS) language. First, the average values for each atomic service included in the composition are computed, then the plan is analysed, the critical path is identified, and the hypothetical value is computed. For instance, the execution duration of the composite service is computed as a sum of execution durations of services being on the critical path. Other calculations include analysis of workflow patterns, determination of how many times services were executed (in case of loops), and so forth. Details about such computation are given by Kowalkiewicz et al. (2005).

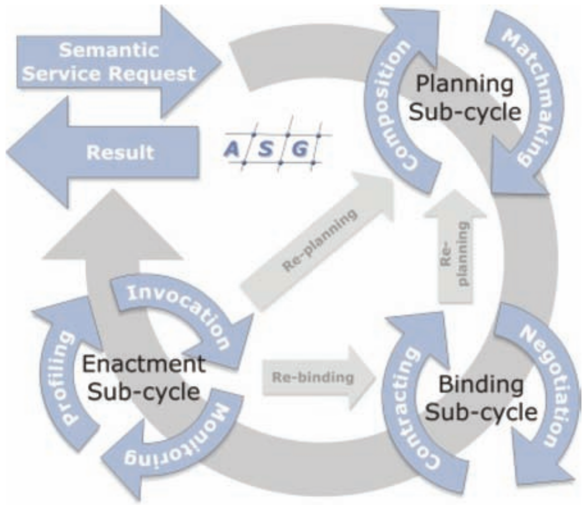
It can be very interesting to rank services according their quality. In order to do that, a method that would allow one to compare objects (in our case, services) with regard to different properties that describe these objects was defined. Our decision was to take advantage of the multiple criteria analysis (MCA) that ideally fitted to our needs. We used the MCA method to rank services based on their quality attributes. This ranking was created by computing a synthetic indicator reflecting the overall service quality. Then, it was possible to compare the values of synthetic indicators of several services and make a choice between them. The detailed description of MCA and the procedure to compute the value of a synthetic indicator is described by Abramowicz, Haniewicz, Kaczmarek, and Zyskowski (2006b).

Dynamic Service Profiling in the Adaptive Services Grid Project

Taking into account the issues discussed in the previous section, the architecture of the service profiling system should consist of at least a few components. It should include the repository that will store the data gathered by the system and should have component(s) responsible for communication with the data sources. Moreover, it should provide interfaces that allow all interested parties to ask queries. Finally, it should have the profiling mechanism, responsible for analysing the data and deriving/computing the values of parameters, to be included in a service profile.

As an example of the architecture of service profiling system, the dynamic service profiling component of the Adaptive Services Grid project, may be presented. The main goal of the ASG project (Kuropka & Weske, 2006) was to develop a proof-of-concept prototype of a platform for adaptive services discovery, creation, composition, enactment, as well as negotiations and service profiling. In order to support the above-mentioned interactions, the ASG platform and mechanisms require the ability to differentiate and compare different services and service substitutes (services having the same functionality). There are some requirements that need to be met in order to make the service differentiation feasible. First, the non-functional parameters must be taken into account, as every customer perceives the service not only from the side of what functionality it gives, but is also interested in non-functional properties of the service. The next issue is to deliver a QoS model

Figure 4. Service delivery process in the ASG. ©Krause, 2005 (used with permission)

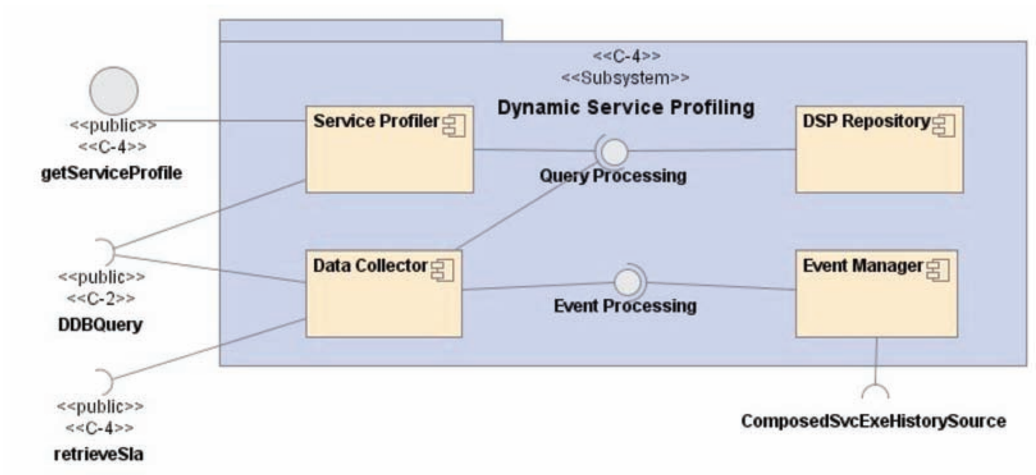


that everybody would accept. Such a standardized QoS model is the first step to the agreement on monitoring mechanisms, common SLAs, and other elements that should be a part of every mature marketplace. The last challenge is to create adequate description of a service that will give a user hints about the distinctive features of service substitutes. Thanks to the monitoring, it should be possible to analyse the information coming from service executions, SLA violations, and so forth. Based on the execution data and users' preferences, it is reasonable to create a service profile which reflects QoS values of a given service in a considered time horizon. Moreover, the user should be capable of ranking these profiles and choosing the most suitable Web service. Such a mechanism is implemented in the Adaptive Services Grid platform (Kuropka & Weske, 2006) using a dynamic service profiling (DSP) mechanism. The ASG service delivery process is presented in the figure below.

The architecture of a dynamic service profiling (see Figure 5) system, being a part of the entire ASG platform, consists of a few components (Abramowicz, Kaczmarek, Kowalkiewicz, & Zyskowski, 2006):

- Data collector, which is responsible for collecting data (by either a push or a pull method) from different sources, processing them, and saving properly aggregated to the DSP repository.
- Service profiler, which is responsible for deriving QoS attributes to answer requests. The Service profiler creates an up-to-date profile of a service (or a provider), whenever it receives a query. Two types of queries may be distinguished: a request for a profile of composed service, taking time horizon into consideration; and a request for profiles and a ranking of a set of atomic services, taking time horizon into consideration. When creating profiles, the service profiler uses the following data about services: data from the provider's declaration (service registry), and values of service attributes form the past execution (DSP repository). In order to create a profile, the appropriate values of characteristics, depending on the prognosis horizon, are computed. Then, based on the computed values a synthetic indicator for a service is created. As an interaction with a user is not

Figure 5. Architecture of DSP system



implemented, default user preferences are used. After computing the indicators for all of the services returned for the given query, services can be compared and the best of them can be identified.

- DSP repository, which is the internal persistent data storage fed by the data collector and responsible for storing all data relevant to service profiles. Only the data collector can change information in the DSP repository. Other subsystems have read-only access to the repository.
- Event Manager, which handles workflow events. The event manager is the subcomponent responsible for processing workflow events and receiving execution logs. If any crucial information is included in such an event, it is passed to the data collector for further analysis.

As verified in the prototype implementation within the ASG project, such an architecture fulfils goals and requirements of a service profiling system.

SUMMARY

This chapter familiarizes users with the idea of Web services profiling. As a background, the current initiatives in the field of Web services description, especially non-functional properties and methods to derive the values of these properties, were presented. Moreover, the readers were introduced to different approaches to the quality-of-service concept. The focus of the chapter was placed on Web service profiling successfully implemented within the ASG system. A service profile, in its final state, aggregates all measured values of quality parameters to give a user the holistic view on a service quality. Taking into account information from profiles, it is possible to select the most suitable service, with regard to the user-specific quality expectations.

REFERENCES

W3C. (2004). *Owl-s: Semantic markup for Web services*. Retrieved May 26, 2008, from http://www.w3.org/Submission/OWL-S/

W3C. (2007). *WSDL 2.0*. Retrieved May 26, 2008, from http://www.w3.org/TR/wsdl20

Aalst, W. D. (2003). Workflow mining: A survey of issues and approaches. *Data & Knowledge Engineering*, *47*(2), 237–267. doi:10.1016/S0169-023X(03)00066-1

Abramowicz, W., Filipowska, A., Kaczmarek, M., Kaczmarek, T., Kowalkiewicz, M., Rutkowski, W., et al. (2006). *Service interdependencies: Insights into use cases for service compositions*. Paper presented at the IFIP 2006.

Abramowicz, W., Haniewicz, K., Kaczmarek, M., & Zyskowski, D. (2006a). *Automatic Web services interactions - requirements, challenges and limits from the F-WebS system perspective*. Paper presented at the International Conference on Next Generation Web Services Practices, Seoul, Korea.

Abramowicz, W., Haniewicz, K., Kaczmarek, M., & Zyskowski, D. (2006b). *Filtering of Semantic Web services with F-WebS system*. Paper presented at the The Semantic Web: ASWC 2006 Workshop.

Abramowicz, W., Kaczmarek, M., Kowalkiewicz, M., & Zyskowski, D. (2005). *A survey of QoS computation for Web services profiling*. Paper presented at the 18th International Conference on Computer Applications in Industry and Engineering (ISCA), Honolulu.

Abramowicz, W., Kaczmarek, M., Kowalkiewicz, M., & Zyskowski, D. (2006). *Architecture for service profiling*. Paper presented at the Modelling, Design and Analysis for Service-Oriented Architecture Workshop in conjunction with the 2006 IEEE International Conferences on Services Computing (SCC 2006) and Web Services (ICWS 2006), Chicago.

Abramowicz, W., Kaczmarek, M., & Zyskowski, D. (2006). *Duality in Web services reliability*. Paper presented at the International Conference on Internet and Web Applications and Services (ICIW'06). Guadeloupe, French Caribbean.

Baida, Z., Gordijn, J., Omelayenko, B., & Akkermans, H. (2004). *A shared service terminology for online service provisioning*. Paper presented at the Sixth International Conference on Electronic Commerce (ICEC04), Delft, The Netherlands.

Casati, F., Castellanos, M., Dayal, U., & Shan, M. C. (2004, November 2004). *Probabilistic, context-sensitive and goal-oriented service selection*. Paper presented at the ICSOC'04, New York.

Eenoo, C. V., Hylooz, O., & Khan, K. M. (2005). *Addressing non-functional properties in software architecture using ADL*. Paper presented at the 6th Australian Workshop on Software and Systems Architectures - AWSA'05, Brisbane, Australia.

Farrel, J., & Lausen, H. (2006). *Semantic Annotations for WSDL*. DERI Innsbruck.

Kowalkiewicz, M., Ludwig, A., Kaczmarek, M., & Zyskowski, D. (2005). *Documented mechanisms for dynamic service prpofiling and agreement life-cycle management* (ASG internal deliverable, FP6-IST-004617).

Krause, H. (2005). *Next generation service delivery: Adaptive services grid, European project, 2007*.

Kuropka, D., & Weske, M. (2006, January). *Die adaptive services grid platform: Motivation, potential, funktionsweise und anwendungsszenarien*. Paper presented at the EMISA Forum.

Kuster, U., Koenig-Ries, B., Stern, M., & Klein, M. (2007, May 8-12). *DIANE: An integrated approach to automated service discovery, matchmaking and composition*. Paper presented at the WWW 2007, Banff, Alberta, Canada.

Lara, R., & Olmedilla, D. (2005, June). *Discovery and contracting of Semantic Web services*. Paper presented at the Position Paper for the Workshop on Frameworks for Semantics in Web Services, Innsbruck, Austria.

Liu, C., Peng, Y., & Chen, J. (2006). *Web services description ontology-based service discovery model.* Paper presented at the International Conference on Web Intelligence (WI 2006).

Liu, Y., Ngu, A. H. H., & Zeng, L. (2004, May). *QoS computation and policing in dynamic Web service selection.* Paper presented at the 13th International Conference on World Wide Web (WWW), New York.

Maximilien, E. M., & Singh, M. P. (2002a). Reputation and endorsement for Web services. *ACM SIGecom Exchanges*, *3*(1), 24–31. doi:10.1145/844331.844335

Maximilien, E. M., & Singh, M. P. (2002b). *Conceptual model of Web services reputation.* SIGMOD Record.

Maximilien, E. M., & Singh, M. P. (2004, November, 2004). *Towards autonomic Web services trust and selection.* Paper presented at the ICSOC'04, New York.

Menasce, D. A. (2002). QoS issues in Web services. *IEEE Internet Computing*, *6*(6), 72–75. doi:10.1109/MIC.2002.1067740

Noll, J. (2004). *ASG based scenarios in telecommunications, telematics and enhanced enterprise IT*. Retrieved May 26, 2008, from http://asg-platform.org

O'Sullivan, J., Edmond, D., & Hofstede, A. T. (2002). What's in a service? Towards an accurate description of non-functional properties of Web services. *Distributed and Parallel Databases*, (12): 117–133. doi:10.1023/A:1016547000822

Preist, C. (2004, November). *A conceptual architecture for Semantic Web services.* Paper presented at the International Semantic Web Conference 2004 (ISWC 2004).

Ran, S. (2003). A model for Web services discovery with QoS. *ACM SIGecom Exchanges*, *4*(1), 1–10. doi:10.1145/844357.844360

Roman, D., et al. (2006). *WWW: WSMO, WSML and WSMX in a nutshell.* Paper presented at the First Asian Semantic Web Conference (ASWC 2006).

Rosa, N. S., Cunha, P. R., Freire, L., & Justo, G. R. (2002, March 29, 2005). *Process NFL: A language for describing non-functional properties.* Paper presented at the 35th Annual Hawaii International Conference (HICSS), Hawaii.

Sheth, A., Cordoso, J., Miller, J., & Kochut, K. (2002, July). *QoS for service-oriented middleware.* Paper presented at the 6rh World Multiconference on Systemics Cybernetics and Informatics (SCI02).

Toma, I. (2006). *Non-functional properties in Web services*. DERI.

UDDI. (2004). *UDDI Version 3.0.2*. UDDI Spec Technical Committee.

Zeng, L., Benatallah, B., Dumas, M., Kalagnanam, J., & Sheng, Q. Z. (2003, May). *Quality driven Web services composition.* Paper presented at the 12th International Conference on World Wide Web (WWW), Budapest, Hungary.

This work was previously published in Managing Web Service Quality: Measuring Outcomes and Effectiveness, edited by K. Khan, pp. 96-113, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 7.14
On the Use of Web Services in Content Adaptation

Khalil El-Khatib
University of Ontario Institute of Technology, Canada

Gregor V. Bochmann
University of Ottawa, Canada

Abdulmotaleb El Saddik
University of Ottawa, Canada

ABSTRACT

The tremendous growth of the Internet has introduced a number of interoperability problems for distributed multimedia applications. These problems are related to the heterogeneity of client devices, network connectivity, content formats, and user's preferences. The challenge is even bigger for multimedia content providers who are faced with the dilemma of finding the combination of different variants of a content to create, store, and send to their subscribers that maximize their satisfaction and hence entice them to come back. In this chapter, the authors will present a framework for trans-coding multimedia streams using an orchestration of Web-services. The framework takes into consideration the profile of communicating devices, network connectivity, exchanged content formats, context description, users' preferences, and available adaptation services to find a chain of adaptation services that should be applied to the content to make it more satisfactory to clients. The framework was implemented as a core component for an architecture that supports personal and service mobility.

DOI: 10.4018/978-1-60566-330-2.ch008

INTRODUCTION

The tremendous growth of the Internet has introduced a number of interoperability problems for distributed multimedia applications. These problems are related to the heterogeneity of client devices, network connectivity, content formats, and user's preferences. The diversity of client devices, network connectivity, content formats, and user's preferences posed also some challenges in aligning and customizing the exchanged data between different users with different preferences. The challenge is even bigger for multimedia content providers who are faced with the dilemma of finding the combination of different variants of a content to create, store, and send to their subscribers that maximize their satisfaction and hence entice them to come back. Most content providers have taken the costly approach of creating different versions of content for different access devices and networks.

Content adaptation is an effective and attractive solution to the problem of mismatch in content format, device capability, network access and user's preferences. Using content adaptation, a number of adaptations is applied to the original content to make it satisfy the device constrains of the receiving device and the preferences of its user. Most currently available content adaptation modules are designed to make the Web easier to use. Examples of such adaptations modules include conversion of HTML pages to Wireless Markup Language (WML, 2001) pages, enlarging text size, reducing the size of an image, changing text and background colors for better contrast, removal of redundant information, audio to text conversion, video to key frame or video to text conversion, content extraction to list a few. These adaptation modules do not have though the same requirements and challenges of real-time multimedia content adaptations. Real-time multimedia applications involve large volumes of data making trans-coding a computationally very expensive task (Chandra & Ellis, 1999, Han et al.,1998). To address this challenge, some trans-coding services have been implemented in hardware and deployed on intermediate network nodes or proxies. The disadvantage of this approach is that there are always new types of clients that cannot be supported by the deployed hardware. A more suitable approach to address the computational challenge of multimedia trans-coding is based on the observation that the general trans-coding process can be defined as a combinatorial process (Mohan, Smith, & Li, 1999), and that multiple trans-coding services can be chained effectively together to perform a complex trans-coding task. So, instead of having all trans-coding done by one single trans-coding service, a number of trans-coding services can collaborate to achieve a composite adaptation task. For instance, trans-coding a 256-color depth *jpeg* image to a 2-color depth *gif* image can be carried out in two stages: the first stage covers converting 256-color to 2-color depth, and the second stage converts *jpeg* format to *gif* format. Using the software approach, transcoders can then be built more easily in software, and deployed and advertised more quickly to meet the needs of the users. Software-based trans-coding are also more reliable since its components can be simpler and they can also be replicated across the network. Moreover, transcoders can be modularized and re-used in different situations and contexts.

Given a composite adaptation task that can be carried out in a number of stages, and given that there could be a number of possible configurations to adapt the sender's content to make it presentable at the receiver's device, the challenge is to find the appropriate chain of available trans-coding services that best fits the capabilities of the device, and at the same time, maximizes the user's satisfaction with the final delivered content. In this chapter, we will discuss a Quality of Service (QoS) selection algorithm for providing personalized content through web-service composition. The function of the algorithm is to find the most appropriate chain of available trans-coding services between the sender and the receiver, and also to select the values for the configuration parameters for each trans-coding service. The proposed algorithm uses the user's satisfaction with the quality of the trans-coded content as the optimization metric for the path selection algorithm.

The rest of the chapter is organized as follows: In Section 2, we will introduce content adaptation and present the existing different models used in content adaptation. Section 3 lists all the required elements for providing customized content adaptation. In Section 4 we present our methodology for using the required element from Section 3 to construct a graph of trans-coding services; the algorithm for selecting the chain of trans-coding services is then presented. The selection criterion for the algorithm as well as its characteristics is also presented in Section 4, and finally, we end Section 4 with an example that shows step-by-step the results of the algorithm. Our conclusion is presented in Section 5.

CONTENT ADAPTATION

In today's Internet, there is a wide range of client devices in terms of both hardware and software capabilities. Device capabilities vary in different dimensions, including processing power, storage space, display resolution and color depth, media type handling, and much more. This variety on device capabilities makes it extremely difficult for the content providers to produce a content that is acceptable and appreciated by all the client devices (Fox, Gribble, & Chawathe, 1998), making application-level adaptation a necessity to cover the wide variety of clients.

There are three main approaches for handling this diversity in content formats: a static content adaptation, a dynamic content adaptation, and a hybrid of the static and dynamic approaches (Chang & Chen, 2002, Lum & Lau, 2002). The first two approaches differ in the time when the different content variants are created (Lei & Georganas, 2001) to match the requested format. In static adaptation, the content creator generates and stores different variants of the same content on a content server, with each variant formatted for a certain device or class of devices. Hafid and Bochmann (1996) presented an architecture for news-on-demand using this scheme. Static adaptation has three main advantages: (1) it is highly customized to specific classes of client devices, and (2) it does not require any runtime processing, so no delay is incurred, and (3) the content creator has the full control on how the content is formatted and delivered to the client. On the other hand, static adaptation has a number of disadvantages, mainly related to the management and maintenance of different variants of the same content (Lum & Lau, 2002): (1) different content formats need to be created for each sort of device or class of devices, and needs to be re-done when new devices are introduced, and (2) it requires large storage space to keep all variants of the same content.

With dynamic content adaptation, the content is trans-coded from one format to the other only when it is requested. Depending on the location where the trans-coding takes place, dynamic content adaptation technologies can be classified into three categories: server-based, client-based, and proxy-based. In the server-based approach (Mohan, Smith, & Li, 1999), the content server is responsible for performing the trans-coding; the content provider has all the control on how the content is trans-coded and presented to the user. Additionally, it allows the content to be trans-coded before it is encrypted, making it secure against malicious attacks. On the other hand, server-based adaptation does not scale properly for a large number of users and requires high-end content and delivery server to handle all requests.

As for the client-based approach (Björk et a., 1999, Fisher et al., 1997), the client does the trans-coding when it receives the content. The advantage of this approach is that the content can be adapted to match exactly to the characteristics of the client. But at the same time, client-based adaptation can be highly expensive in terms of bandwidth and computation power, especially for small devices with small computational power and slow network connectivity, with large volume of data might be wastefully delivered to the device to be dropped during trans-coding.

The third adaptation approach is the proxy-based approach (Chandra & Ellis, 1999, Chandra, Ellis, & Vahdat, 2000, Floyd & Housel, 1998, Fox, A., Gribble, Chawathe, Brewer, & Gauthier, 1997), where an intermediary computational entity can carry out content adaptation on the fly, on behalf of the server or client. Proxy adaptation has a number of benefits including leveraging the installed infrastructure and scaling properly with the number of clients. It also provides a clear separation between content creation and content adaptation. On the other hand, some content provider may argue that they prefer to have full control on how their content is presented to the user. Also, using proxies for adaptation does not allow the use of end-to-end security solutions.

CHARACTERIZATION AND REQUIREMENTS FOR CONTENT ADAPTATION

Advances in computing technology have led to a wide variety of computing devices, which made interoperability very difficult. Added to this problem is the diversity of user preferences when it comes to multimedia communications. This diversity in devices and user preferences has made content personalization an important requirement in order to achieve results that satisfy the user. The flexibility of any system to provide content personalization depends mainly on the amount of information available on a number of aspects involved in the delivery of the content to the user. The more information about these aspects is made available to the system, the more the content can be delivered in a format that is highly satisfactory to the user. These relevant aspects are: user preferences, media content profile, network profile, context profile, device profile, and the profile of intermediaries (or proxies) along the path of data delivery. We will briefly describe here each of these aspects; interested readers might refer to (El-Khatib & Bochmann, 2003) for more details.

User Profile: The user's profile captures the personal properties and preferences of the user, such as the preferred audio and video receiving/sending qualities (frame rate, resolution, audio quality…). Other preferences can also be related to the quality of each media types for communication with a particular person or group of persons. For instance, a customer service representative should be able to specify in his profile his/her preference to use high-resolution video and CD audio quality when talking to a client, and to use telephony quality audio and low-resolution video when communicating with a colleague at work. The user's profile may also hold the user's policies for application adaptations, such as the preference of the user to drop the audio quality of a sport-clip before degrading the video quality when resources are limited. The MPEG-21 standard (MPEG-21, 2001) is the most notable standards on user profiles.

Content Profile: Multimedia content might enclose different media types, such as audio, video, text, and each type can have different formats (Lei & Georganas, 2001). Each type has its format characteristics and parameters that can be used to describe the media. Such information about the content may include storage features, variants, author and production, usage, and many other metadata. The MPEG-7 standard (MPEG-7, 2000), formally named "Multimedia Content Description Interface", offers a comprehensive set of standardized description tools to describe multimedia content.

Context Profile: A context profile would include any dynamic information that is part of the context or current status of the user. Context information may include physical (e.g. location, weather, temperature), social (e.g. sitting for dinner), or organizational information (e.g. acting senior manager). The MPEG-21 standard includes tools for describing the natural environment characteristics of the user, including location and time, as well as the audio and illumination characteristics of the user's environment. Resource adaptation engines can use these elements to deliver the best experience to the user.

Device Profile: To ensure that a requested content can be properly rendered on the user's device, it is essential to include the capabilities and characteristics of the device into the content adaptation process. Information about the rendering device may include the hardware characteristics of the device, such as the device type, processor speed, processor load, screen resolution, color depth, available memory, number of speakers, the display size, and the input and output capabilities. The software characteristics such as the operating system (vendor and version), audio and video codecs supported by the device should also be included in the device profile. The User Agent Profile (UAProf) created by the Wireless Application Forum (WAP) and the MPEG-21 standard, both include description tools for describing device capabilities.

Network Profile: Streaming multimedia content over a network poses a number of technical challenges due to the strict QoS requirements of multimedia contents, such as low delay, low jitter, and high throughput (Ng, Tan, & Cheng, 2001). Failing to meet these requirements may lead to a bad experience of the user (Katchabaw, Lutfiyya, & Bauer,, 1998, Poellabauer, Abbasi, & Schwan, 2002). With a large variety of transport networks, it is necessary to include the network characteristics into content personalization and to dynamically adapt the multimedia content to the fluctuating network resources (Wu, Hou, Zhang, 2001). Achieving this requires collecting information about the available resources in the network, such as the maximum delay, error rate, and available throughput on every link over the content delivery path. A description tool for network capabilities, including utilization, delay and error characteristics are included in the MPEG 21 standard.

Profile of Intermediaries: When the content is delivered to the user across the network, it usually travels over a number of intermediaries. These intermediaries have been traditionally used to apply some added-value services, including on-the-fly content adaptations services (Chandra, Ellis, & Vahdat, 2000, Fox, Gribble, Chawathe, Brewer, & Gauthier, 1997). For the purpose of content adaptation, the profile of an intermediary would usually include a description of all the adaptation services that an intermediary can provide. These services can be described using any service description language such as the JINI network technology (JINI, 1998), the Service Location Protocol (Guttman, Perkins, Veizades, & Day, 1999), or the Web Service Description Language (WSDL, 2002). A description of an adaptation service would include, for instance, the possible input and output format to the service, the required processing and computation power of the service, and maybe the cost for using the service. The intermediary profile would also include information about the available resources at the intermediary (such as CPU cycles, memory) to carry out the services.

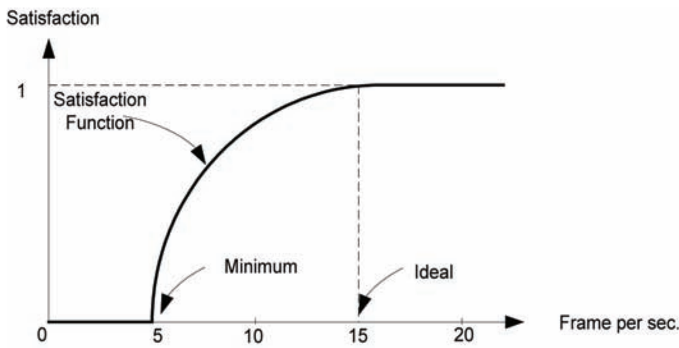
QOS SELECTION ALGORITHM

In this section, we will describe the overall QoS selection algorithm that finds the most appropriate chain of trans-coding services between the sender and the receiver, and also selects the configuration for each trans-coding service. We will first start by defining the user's satisfaction as the selection criterion for the algorithm, and then show how to construct the directed graph for adaptation, using the sender's content profile, receiver's device profile, and the list of available trans-coding services. After constructing the graph, we will show how to apply some optimization techniques on the graph to remove the extra edges in the graph, and finally present the actual QoS path and parameter selection algorithm.

User's Satisfaction as Selection Criteria

Most Internet users are indifferent about the underlying technologies such as protocols, codecs,

Figure 1. Possible satisfaction function for the frame rate.



or resource reservation mechanisms that enable their communication session. They are also indifferent about network level QoS characteristics, such as bandwidth, delay, or throughput. All what is important for these users in the end is making the communication session work in a satisfactory way: for instance, hearing without jitter and seeing without irregularity.

As we mentioned earlier, the user's preferences expressed in the user's profile can be classified as application layer QoS parameters. In order to compute the user's satisfaction with all values of the application layer configuration parameters, we have used the approach presented by Richards, Rogers, Witana, & Antoniades (1998), where each application level QoS parameter is represented by a variable x_i over the set of all possible values for that QoS parameter. The satisfaction or appreciation of a user with each quality value is expressed as a satisfaction function $S_i(x_i)$. All satisfaction functions have a range of [0..1], which corresponds to the minimum acceptable (M) and ideal (I) value of x_i. The satisfaction function $S_i(x_i)$ can take any shape, with the condition that it must increase monotonically over the domain. Figure 1 shows a possible satisfaction function for the frame rate variable.

In the case when there are more than one application parameter (frame rate, resolution, color depth, audio quality,…), Richards *et. al.* proposed using a combination function f_{comb} that computes the total satisfaction S_{tot} from the satisfactions s_i for the individual parameters (Equa. 1).

$$S_{tot} = f_{comb}(s_1, s_2, s_3 \ldots, s_n) = \frac{n}{\sum_{i=1}^{n} \frac{1}{s_i}} \tag{1}$$

Extending User's Satisfaction to Support Weighted Combination and Multi-User Conference Sessions

We think that the approach described above is a major step towards a simple user-friendly interface for user level QoS specification, however, further considerations could be taken into account as described below. A first improvement results from the observation that users in telecommunication session might find some media types more important than others. For instance, a user of a news-on-demand service might prefer to receive high quality audio with low quality video as compared to average quality audio and average quality video. In the case of a user watching a sport event the situation may be the opposite (if the user does not care about the audio of the commenter).

This preference to individual media can play a factor when it comes to the calculation of the total satisfaction S_{tot}. By assigning different weights w_i to the different parameters x_i, S_{tot} will reflect the user preference for different media types. The combination function for the total user satisfaction can be redefined as follows:

$$S_{tot}^{user} = f_{comb}(s_1, s_2, s_3 \ldots, s_n, w_1, w_2, w_3 \ldots, w_n) = \frac{n\bar{w}}{\sum_{i=1}^{n} \frac{w_i}{s_i}} \tag{2}$$

where w_i is the weight for the individual satisfaction s_i and $\bar{w} = \frac{\sum_{i=1}^{n} w_i}{n}$. Equa. 2 have similar properties as Equa. 1, which is to:

Prop. 1. One individual low satisfaction is enough to bring the total satisfaction to a low value.

Prop. 2. The total satisfaction of equal individual satisfactions s_i with equal weight is equal to the satisfactions s_i.

These constants weight factors (AudioWeightFactor, VideoWeightFactor,..) can be selected by the user, and stored in the user profile. The selection of these weights depends on the type of service the user is willing to receive when using a specific service or communicating with a given callee.

Additionally, we have so far considered only the QoS preferences of a single user. But all conversational multimedia applications involve several users. It is therefore important to determine how the possibly conflicting preferences of the different users are reconciled in order to come up with QoS parameters that are suitable for all participating users.

In certain circumstances, some given parameters may be determined simply based on the preferences of a single user. This may be the case in a two-way teleconference between two users A and B, where the parameters of the video visible by User A would be determined based on the preferences of User A alone, and the video in the opposite direction based on the preferences of User B. However, the situation may be more complex if the cost of the communication is paid by User A and the selection of the video received by User B has an impact on the communication cost.

In other circumstances, as for instance in the case of the joint viewing of a video clip by several participants in a teleconference, the selected quality parameters should be determined based on the preferences of all participating users. In such circumstances, we propose to use the same combination function for user satisfaction considered above and (optionally) introduce a weight for each of the participating users, called the *QoS selection weight*, which determines how much the preferences of the user influences overall QoS parameter selection. The total satisfaction (computed for all users) is then given by

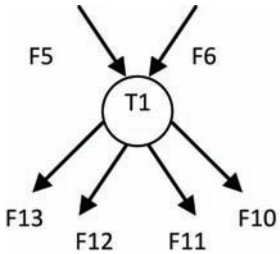
$$S_{tot} = f_{comb}(s_{tot}^{usr_1}, s_{tot}^{usr_2} \ldots, s_{tot}^{usr_m}, a_1, a_2 \ldots, a_m) = \frac{m\bar{a}}{\sum_{i=1}^{m} \frac{a_i}{s_{tot}^{usr_i}}} \tag{3}$$

where $s_{tot}^{usr_i}$ is the total satisfaction for user *i*, and a_i is the *QoS selection weight* for user *i*. In the case that the weight of a given user is zero, the preferences of this user are not taken into account for the selection of the QoS parameters.

Constructing a Directed Graph of Trans-Coding Services

Now that we have decided on the selection criteria, the first step of the QoS selection algorithm would be to construct a directed acyclic graph for adaptation, using the content profile, device profile, and the list of available trans-coding services. Using this graph, the route selection algorithm would then determine the best path through the graph, from the sender to the receiver, which maximizes the user's satisfaction with the final received adapted

Figure 2. Trans-coding service with multiple input and output links



content. The elements of the directed graph are the following:

1. Vertices in the graph represent trans-coding services. Each vertex of the graph has a number of properties, including the computation and memory requirements of the corresponding trans-coding service. Each vertex has a number of input and output links. The input links to the vertex represent the possible input formats to the trans-coding service. The output links are the output formats of the trans-coding service. Figure 2 shows a trans-coding service T1, with two input formats, F5 and F6, and four possible output formats, F10, F11, F12 and F13. The sender node is a special case vertex, with only output links, while the receiver node is another special vertex with only input links.

   To find the input and output links of each vertex, we rely on the information in different profiles. The output links of the sender are defined in the content profile, which includes as we mentioned earlier, meta-data information (including type and format) of all the possible variants of the content. Each output link of the sender vertex corresponds to one variant with a certain format. The input links of the receiver are exactly the possible decoders available at the receiver's device. This information is available through the description of the receiver's device in the device profile. The input and output links of intermediate vertices are described in the service description part of the intermediaries profile. Each intermediary profile includes the list of available trans-coding services, each with the list of possible input and output formats. Each possible input format is represented as an input link into the vertex, and the output format is represented as an output link.
2. Edges in the graph represent the network connecting two vertices, where the input link of one vertex matches the output link of another vertex.

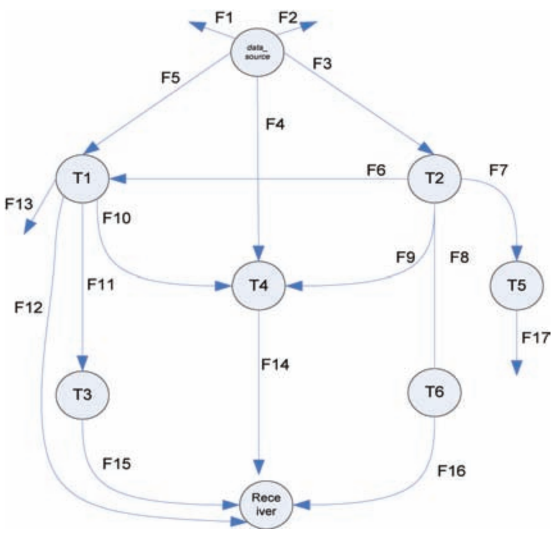
   To construct the adaptation graph, we start with the sender node, and then connect the outgoing edges of the sender with all the input edges of all other vertices that have the same format. The same process is repeated for all vertices. To make sure that the graph is acyclic, the algorithm continuously verifies that all the formats along any path are distinct.

   Figure 3 shows an example of an adaptation graph, constructed with one sender, one receiver, and seven intermediate vertices, each representing a trans-coding service. As we can see from the graph, the sender node is connected to the trans-coding service T1 along the edge labeled F5. This means that the sender S can deliver the content in format F5, and trans-coding service T1 can convert this format into format F10, F11, F12, or F13.

Adding Constraints to the Graph

As we have discussed earlier, the optimization criterion we have selected for the QoS selection algorithm is the user's satisfaction computed using the function f_{comb} presented in Section 4.2. The maximum satisfaction achieved by using a trans-coding service T_i depends actually on a number of factors.

Figure 3. Directed trans-coding graph



The first factor is the bandwidth available for the data generated by the trans-coding service T_i. The more bandwidth is available to the trans-coding service, the more likely the trans-coding service will be able to generate trans-coded content that is more appreciated by the receiver. The available bandwidth between two trans-coding services is restricted by the amount of bandwidth available between the intermediate servers where the trans-coding service T_i is running and the intermediate server where the next trans-coding service or receiver is running. We can assume that connected trans-coding services that run on the same intermediate server have an unlimited amount of bandwidth between them.

Other factors that can affect the user's satisfaction are the required amount of memory and computing power to carry out the trans-coding operation. Each of these two factors is a function of the amount of input data to the trans-coding service.

Graph Optimization

By looking at the graph in Figure 3, we can see that there are some edges like F1, F2 or F17 that are connected only to one trans-coder. These edges cannot be a part of any path from the sender to the receiver. The same principle also applies to trans-coders other than the sender and receiver that are not on any path from the sender to the receiver. T5 is an example of a trans-coder that cannot be used to send data through it on the way from the sender to the receiver. Removing these edges and vertices help reduce the computational time of the algorithm, since it helps pruning dead-ends from the graph. Applying optimization for the graph in Figure 3 would result in the graph shown in Figure 5. The pseudo-code for the graph optimization is shows in Figure 4.

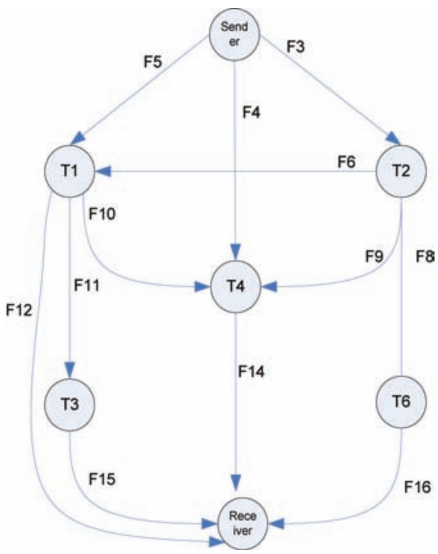
QoS Selection Algorithm

Once the directed acyclic adaptation graph has been constructed, the next step is to perform the

Figure 4. Pseudo-code for the graph optimization

```
graph_optimization(Transcoder t){
    if t ≠ receiver then {
            for all ne ≠ neighbor(t)
                    graph_optimization(ne);
            if is_empty(neighbor(ne)) then
                    delete(ne);
                    }
}
```

Figure 5. Optimized directed trans-coding graph



QoS selection algorithm to find a chain of trans-coding services, starting from the sender node and ending with the receiver node, which generates the maximum satisfaction of the receiver. Finding such as path can be similar to the problem of finding the shortest path in a directed weighted graph with similar complexity, except that the optimization criterion is the user's satisfaction, and not the available bandwidth or the number of hops.

Our proposed algorithm uses two variables representing two sets of trans-coding services, the set of already considered trans-coding services, called VT, and the set of candidate trans-coding services, called CS, which can be added next on the partially selected path. The candidate trans-coding services set contains the trans-coding services that have input edges coming from any trans-coding service in the set VT. At the beginning of the algorithm, the set VT contains only the *sender* node, and CS contains all the other trans-coding services in the graph that are connected to *sender*, and also the *receiver*. In each iteration, the algorithm selects the trans-coding service T_i that, when using it, generates the highest user satisfaction. The user satisfaction is computed as an optimization function of the audio and video parameters for the output format for T_i, subject to the constraint of available bandwidth between T_i and its ancestor trans-coding service, and also subject to the remaining user's budget. T_i is then added to VT. The CS set is then updated with all the neighbor trans-coding services of T_i. The algorithm stops when the CS set is empty, or when the *Receiver* node is selected to be added to VT. The complete description of the algorithm is given in Figure 6.

As indicated in Step 2 and Step 8, the algorithm selects from CS the transcoder T_i that can generate the highest satisfaction value for the receiver. To compute the satisfaction value for each transcoder T_i in CS, the algorithm selects the QoS parameter values x_i that optimize the satisfaction function in Equa. 2, subject only to the constraint remaining user's budget and the bandwidth availability that connects T_i to T_{prev} in VT. i.e.

$$bandwith\_requirement(x_1..x_n) \leq Bandwidth\_AvailableBetween(T_i, T_{prev}). \qquad (4)$$

Since each trans-coding service can only reduce the quality of the content, when the algorithm terminates, the algorithm would have computed the best path of trans-coding services from the *sender* to the *receiver*, and the user's satisfaction value computed on the last edge to the receiver node is the maximum value the user can achieve. To show this, assume that the selected path is the path $\{T_{11}, \ldots T_{1n}\}$ in Figure 7. If the path $\{T_{21}, \ldots T_{2m}\}$ is a better path, then T_{2m} should have converted the content into variant that is more appreciated by the user than the variant generated by T_{1n}. Since transcoders can only reduce the quality of content, all transcoders along the path $\{T_{21}, \ldots T_{2m}\}$, should have also produced a content with higher satisfaction function than the variant produce by T_{1n}, and hence all these transcoders should have been selected before T_{1n}, which contradicts with the assumption.

Figure 6. QoS selection algorithm

```
Step 1: // Let VT be the set of all considered trans-coding services.
        VT = {sender};
// Let CS be the set of all direct neighbor transcoders of all transcoders in VT
     CS = neighbor(sender);
        // Let user_budget be the amount of money the user is willing to pay
Step 2: // Each trans-coding service keeps a track of its parent trans-coding service. Let T_prev be the trans-coding services in CS connected to
        // T_i; Compute the perceived user's satisfaction for using all the trans-coding services in CS, subject to two constraints: the remaining
        // user budget and the available bandwidth between T_i and T_prev.
        For □ T_i □ CS
          Optimize( user_profile, input_format, output_format, Sat_T[i],
                              user_budget,cost,available_bandwith)
Step 3: // If there are no more transcoders to consider and the receivers can
        // not be reached from the sender through any transcoding path.
        if is_empty(CS) then
           TERMINATE(FAILURE)
Step 4: Select the trans-coding service T_i that has the highest satisfaction
        value Sat_T[i], for the user.
        CS = CS – { T_i };
Step 5: VT = VT + { T_i };
Step 6: Let T_i. previous_selected_transcoder = T_prev;
        T_i.accumulated_cost = T_i.previous.accumulated_cost + transcoding_and_transmission_cost [i];
Step 7: if T_i = receiver , then GOTO Step 10
Step 8: // compute the satisfaction for using all the neighboring transcoders of T_i and add them to CS
        For □ T_j □ neighbors( T_i );
          Optimize( user_profile, input_format, output_format, Sat_T[i], user_budget,cost[i],available_bandiwth)
           CS = CS □ {T_j };
Step 9: GOTO Step 3
Step 10:   Print the reverse path from the Receiver to the Sender by following the link "previous" of all transcoders, starting from the Receiver.
```

Figure 7. Graph selection

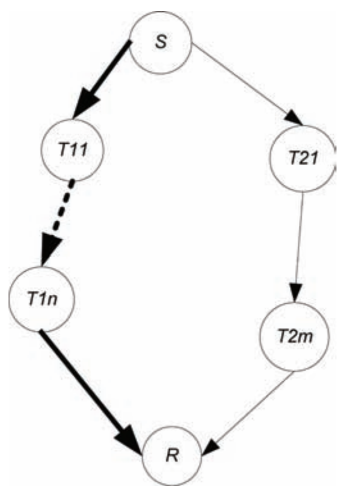
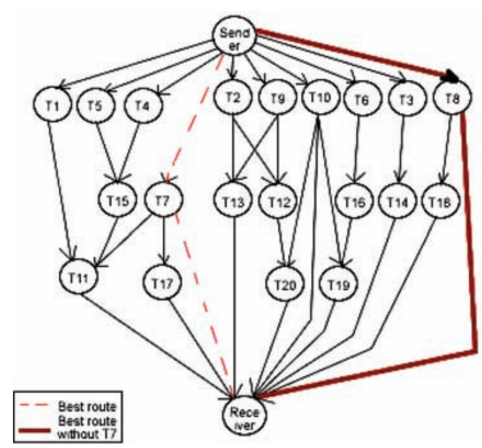


Figure 8. Example of trans-coding graph



Example

In this section, we will present an example to show how the QoS path selection algorithm works. We will assume that the graph construction algorithm has generated the graph shown in Figure 8. The graph also shows the selected path with and without trans-coding service T_7 as part of the graph. The selected trans-coding services, user satisfaction, as well as the best current path produced by the algorithm are also shown in Table 1. Each row in the table shows the results for one iteration of the algorithm.

Table 1. Results for each step of the path selection algorithm

Round	Considered Set (VT)	Candidate set (CS)	Selected transcoding service	Selected Path	Delivered Frame Rate	User satisfaction
1	{ *sender* }	{T1, T2, T3, T4, T5, T6, T7, T8, T9, T10}	T10	*sender*,T10	30	1.00
2	{ *sender*, T10}	{T1, T2, T3, T4, T5, T6, T7, T8, T9, T19, T20, *receiver*}	T20	*sender*, T10, T20	30	1.00
3	{ *sender*, T10, T20}	{T1, T2, T3, T4, T5, T6, T7, T8, T9, T19, *receiver*}	T5	*sender*,T5	27	0.90
4	{ sender, T10, T20, T5}	{T1, T2, T3, T4, T6, T7, T8, T9, T19, T15, receiver}	T4	sender,T4	27	0.90
5	{ sender, T10, T20, T5, T4}	{T1, T2, T3, T6, T7, T8, T9, T19, T15, receiver}	T3	sender,T3	23	0.76
6	{ sender, T10, T20, T5, T4, T3}	{T1, T2, T6, T7, T8, T9, T19, T15, T14, receiver}	T2	sender,T2	23	0.76
7	{ sender, T10, T20, T5, T4, T3, T2}	{T1, T6, T7, T8, T9, T19, T15, T14, T12, T13, receiver}	T1	sender,T1	23	0.76
8	{ sender, T10, T20, T5, T4, T3, T2, T1}	{T6, T7, T8, T9, T19, T15, T14, T12, T13, T11, receiver}	T11	sender,T1, T11	23	0.76
9	{ sender, T10, T20, T5, T4, T3, T2, T1, T11}	{T6, T7, T8, T9, T19, T15, T14, T12, T13, receiver}	T13	sender,T2, T13	23	0.76
10	{ sender, T10, T20, T5, T4, T3, T2, T1, T11, T13}	{T6, T7, T8, T9, T19, T15, T14, T12, receiver}	T12	sender,T2, T12	23	0.76
11	{ sender, T10, T20, T5, T4, T3, T2, T1, T11, T13, T12}	{T6, T7, T8, T9, T19, T15, T14, receiver}	T14	sender,T3, T14	23	0.76
12	{ sender, T10, T20, T5, T4, T3, T2, T1, T11, T13, T12, T14}	{T6, T7, T8, T9, T19, T15, receiver}	T8	sender, T8	20	0.66
13	{ sender, T10, T20, T5, T4, T3, T2, T1, T11, T13, T12, T14, T8}	{T6, T7, T9, T19, T15, receiver}	T7	sender, T7	20	0.66
14	{ sender, T10, T20, T5, T4, T3, T2, T1, T11, T13, T12, T14, T8, T7}	{T6, T9, T19, T15, receiver}	T6	sender, T6	20	0.66
15	{ sender, T10, T20, T5, T4, T3, T2, T1, T11, T13, T12, T14, T8, T7, T6}	{T9, T19, T15, receiver}	receiver	sender, T7, receiver	20	0.66

FUTURE RESEARCH DIRECTION

In this section, we will outline some potential directions for future research works.

In this chapter, we have not addressed the issues regarding autonomic service management of real-time multimedia services. One of the challenges is the efficient, autonomous management of these real-time content adaptation services in future generation networks. The autonomous service management is crucial for the self-management of real-time multimedia services. According to Ganek and Corbi of IBM (Ganek and Corbi, 2003), the autonomous or self-management aspects include self-optimizing, self-healing, self-configuring, and self-protecting. The existing approach and framework contributes towards a system that is not fully autonomic in all four management aspects.

Current, we are looking at nature-inspired automatic service management solution that is inspired from the Bee colony metaphor. The allegory comprises how bee agents mimic functional services related to multimedia applications, in order to autonomously monitor and configure multimedia services. The objective of this research is to ensure complete autonomic behaviour of the four main management activities (configuration, repair, optimization and protection) of an autonomous system. Such direction could enable customization of the service for the current and future generation network conditions.

CONCLUSION

Content adaptation is a natural solution to address the problem of heterogeneity of Internet clients and users. In this chapter, we have presented a solution to the problem of heterogeneity which takes into consideration the capabilities of the client devices, network connectivity, content format, and users' preferences. An important part of the framework is the QoS path selection algorithm that decides on the chain of adaptation services to add and the configuration parameters for each service. The decision is based on the profile of communicating devices, network connectivity, exchanged content formats, context description, and available adaptation services.

REFERENCES

Björk, S., Holmquist, L. E., Redström, J., Bretan, I., Danielsson, R., Karlgren, J., & Franzén, K. (1999). WEST: a Web browser for small terminals. *Proceedings of the 12th annual ACM symposium on User interface software and technology* (pp.187-196). Asheville, North Carolina, United States.

Chandra, S., Ellis, C., & Vahdat, A. (2000). Application-Level Differentiated Multimedia Web Services Using Quality Aware Transcoding . *IEEE Journal on Selected Areas in Communications*, *18*(12), 2265–2544. doi:10.1109/49.898736

Chandra, S., & Ellis, C. S. (1999). JPEG Compression Metric as a Quality Aware Image Transcoding. *Second Usenix Symposium on Internet Technologies and Systems* (USITS '99). (pp. 81–92) Boulder, CO.

Chang, C. Y., & Chen, M. S. (2002). Exploring Aggregate Effect with Weighted Transcoding Graphs for Efficient Cache Replacement in Transcoding Proxies. *Proceedings of the 18th IEEE International Conference on Data Engineering (ICDE-O)* (pp. 383—392). San Jose, CA, USA.

El-Khatib, K., & Bochmann, G. v. (2003, December). *Profiles in Content Adaptation.* Technical report. University of Ottawa, Canada.

El-Khatib, K., Zhang, Z. E., Hadibi, N. & Bochmann, G. v. (2004). Personal and Service Mobility in Ubiquitous Computing Environments. *Journal of Wireless communications and Mobile Computing*, *4*(6), 595-607.

Fisher, B., Agelidis, G., Dill, J., Tan, P., Collaud, G., & Jones, C. (1997). CZWeb: Fish-Eye Views for Visualizing the World-Wide Web. In *Proceeding of the 7th Int. Conf. on Human-Computer Interaction (HCI International '97)* (pp. 719-722). Elsevier, Amsterdam.

Floyd, R., & Housel, B. (1998). Mobile Web Access Using eNetwork Web Express. *IEEE Personal Communications*, *5*(5), 47–52. doi:10.1109/98.729724

Fox, A., Gribble, S. D., & Chawathe, Y. (1998). Adapting to Network and Client Variation Using Infrastructural Proxies: Lessons and Perspectives. (1998). [Springer Berlin/Heidelberg.]. *IEEE Personal Communications*, *5*(4), 10–19. doi:10.1109/98.709365

Fox, A., Gribble, S. D., Chawathe, Y., Brewer, E. A., & Gauthier, P. (1997). Cluster-Based Scalable Network Services. In *Proceeding of the 16th ACM Symp. On Operating Systems Principles* (pp. 78–91). Saint-Malo, France.

Ganek, A. G., & Corbi, T. A. (2003). The dawning of the autonomic computing era. *IBM Systems Journal*, *42*(1), 5–18.

Guttman, E., Perkins, C., Veizades, J., & Day, M. (1999). Service Location Protocol. Version 2. http://ietf.org/rfc/rfc2608.txt.

Hafid, A. & Bochmann, G.v., (1996). Quality of Service Negotiation in News-on-Demand Systems: an Implementation. In *Proceedings of the Third International Workshop on Protocols for Multimedia Systems* (pp. 299-315). Springer Berlin/Heidelberg.

Han, R., Bhagwat, P., LaMaire, R., Mummert, T., Perret, V., & Rubas, J. (1998). Dynamic adaptation in an image trans-coding proxy for mobile WWW browsing. *IEEE Personal Communication, 5*(6). JINI network technology (TM) (1998). Http://java.sun.com/product/JINI

Katchabaw, M., Lutfiyya, H., & Bauer, M. (1998). *Driving resource management with application-level quality of service specifications,* (pp. 83-91) ACM Press.

Lei, Z., & Georganas, N. D. (2001). Context-based Media Adaptation in Pervasive Computing. On *Proceeding Can.Conf. on Electr. and Comp. Engg.* (pp. 913-918). Toronto, Canada.

Lum, W. Y., & Lau, F. C. M. (2002). On Balancing Between Trans-coding Overhead and Spatial Consumption in Content Adaptation. [Atlanta, USA.]. *Mobicom*, *2002*, 239–250.

Mohan, R., Smith, J. R., & Li, C. S. (1999). Adapting Multimedia Internet Content for Universal Access. *IEEE Transactions on Multimedia*, *1*(1), 104–114. doi:10.1109/6046.748175

MPEG-21 (2001): International Standards Organisation. Information technology – multimedia framework (MPEG-21) – part 1: Vision, technologies and strategy. ISO/IEC 21000-1. Accessed on Jan 10, 2007.

MPEG-7. http://www.chiariglione.org/mpeg/standards/ mpeg-7/mpeg-7.htm. Accessed on Jan 10, 2007.

Ng, C.W., Tan, P.Y., & Cheng, H. (2001). *Quality of Service Extension to IRML*. IETF INTERNET-DRAFT, ‘draft-ng-opes-irmlqos-00.txt’.

Poellabauer, C., Abbasi, H., & Schwan, K. (2002). Cooperative run-time management of adaptive applications and distributed resources. In *Proceeding of the Tenth ACM Internationalconference on Multimedia,* (pp. 402-411) ACM Press.

Richards, A., Rogers, G., Witana, V., & Antoniades, M. (1998). Mapping user level QoS from a single parameter. In *Second IFIP/IEEE International Conference on Management of Multimedia Networks and Services*. (pp. 14-20). Versailles, France.

Smith, J. R., Mohan, R., & Li, C.-S. (1999). Scalable Multimedia Delivery for Pervasive Computing. *ACM Multimedia*, (pp. 131 – 140). Orlando, Florida, United States.

WML: Wireless Markup Language (2001). *Wireless Markup Language (WML) 2.0 Document Type Definition.*

WSDL. Web Service Description Language (2002). http://www.w3.org/TR/wsdl.

Wu, D., Hou, Y. T., & Zhang, Y. (2001). Scalable Video Coding and Transport over Broad-band Wireless Networks. *Proceedings of the IEEE*, *89*(1), 6–20. doi:10.1109/5.904503

ADDITIONAL READING

Ahmed, I., Wei, X., Sun, Y., & Zhang, Y. Q. (2005). Video transcoding: an overview of various techniques and research issues. *IEEE Transactions on Multimedia*, *7*(5), 793–804. doi:10.1109/TMM.2005.854472

Ardon, S.(n.d.). MARCH: a distributed content adaptation architecture. *International Journal of Communication Systems*, *16*, 97–115.

Dey, A. K. (2001). Understanding and Using Context. *Springer Personal and Ubiquitous Computing*, *5*(1), 4–7. doi:10.1007/s007790170019

El Saddik, & Hossain, M. S. (2007). Multimedia Streaming for wireless communication. In B. Furht (Ed.), *Encyclopedia of Wireless and Mobile Communications*. CRC Press, Taylor & Francis Group.

El Saddik, A., & Hossain, M. S. (2006). Multimedia content repurposing. In B. Furht, (Ed.), *Encyclopedia of Multimedia*. Berlin, Germany: Springer Verlag.

Han, J. R. et al.(1998, December). Dynamic adaptation in an image transcoding proxy for mobile WWW browsing. *IEEE Personal Commun.*, *5*(6).

Han, R., & Smith, J. R. (1999). Internet Transcoding for Universal Access. In J. Gibson (Ed.), *Multimedia Communications Handbook*.

Hossain, M. S., & El Saddik, A. (2008). A Biologically Inspired Multimedia Content Repurposing System in Heterogeneous Network Environments. *ACM/Springer Multimedia Systems J.*, *14*(3), 135-144.

Hossain, M. S., Alamri, A., & El Saddik, A. (2007). A framework for qos-aware multimedia service selection for wireless clients. In *Proc. the 3rd ACM Workshop on Wireless Multimedia Networking and Performance Modeling (WMuNeP 07)*, Chania, Crete Island, Greece, October 22 - 22.

Liang, Y., Chebil, F., & Islam, A. (2006). Compressed domain transcoding solutions for MPEG-4 visual simple profile and H.263 baseline videos in 3GPP services and applications. *IEEE Transactions on Consumer Electronics*, *52*(2), 507–515. doi:10.1109/TCE.2006.1649672

Lum, W. Y., & Lau, F. C. M. (2002). On Balancing between Transcoding Overhead and Spatial Consumption in Content Adaptation. In *Proc. MobiCom'02*, Atlanta, Georgia, USA, Sep. 23-26. (pp. 239-250).

Maheshwari, A., Sharma, A., Ramamritham, K., & Shenoy, P. (2002). TransSquid:Transcoding and caching proxy for heterogeneous e-commerce environments. In *Proc. 12th IEEE Int. Workshop Research Issues in Data Engg*, San Jose, California, USA, 26 Feb. - 1 March, 2002, (pp. 50-59).

Mao, M., So, H. W., Kang, B., & Katz, R. H. (2001). Network support for mobile multimedia. In *Proc. 11th Intl. Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV-2001)*, New York, USA.

Nahrstedt, K., & Balke, W. T. (2004). A taxonomy for multimedia service composition. In *Proc. 12th ACM Conf. Multimedia (ACM MM 04)*, New York, NY, USA, 10–16 October 2004, (pp. 88-95).

Nahrstedt, K., & Balke, W. T. (2005). Towards building large scale multimedia systems and applications: Challenges and status. In *Proc. the First ACM Intl.Workshop Multimedia Service Composition*, Hilton, Singapore, (pp. 3-10).

Nguyen, V. A., & Tan, Y. P. (2005). Efficient video transcoding between H.263 and H.264/AVC standards. In *Proc. IEEE Intl. Symposium on Circuits and Systems (ISCAS'05)*, Kobe, Japan, May 23-26.

Richards, A., Rogers, G., Witana, V., & Antoniades, M. (1998, November). Mapping user level QoS from a single parameter. IIn *2nd IFIP/IEEE Intl. Conf. Manage. Multimedia Networks and Services*, Versailles.

Shin, I., & Koh, K. (2004). Hybrid Transcoding for QoS Adaptive Video-on-Demand Services. *IEEE Transactions on Consumer Electronics*, *50*(2).

Smith, J. R., Mohan, R., & Li, C. S. (1999). Scalable Multimedia Delivery for Pervasive Computing. In *Proc. ACM Multimedia' 99*, Orlando, FL, USA, Oct.30 - Nov.5.

Vetro, A., Xin, J., & Sun, H. (2005). Error resillence video transcoding for wireless communications. *IEEE Wirel. Commun*, *12*(4), 14–21. doi:10.1109/MWC.2005.1497854

This work was previously published in Services and Business Computing Solutions with XML: Applications for Quality Management and Best Processes, edited by P. C. K. Hung, pp. 121-135, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 7.15

Reconceptualising Information Literacy for the Web 2.0 Environment?

Sharon Markless
King's College, London, UK

David Streatfield
Information Management Associates, UK

ABSTRACT

This chapter questions whether the shift from the Web as a vehicle for storing and transmitting information to the new Web as a series of social networking environments, requires significant changes in how students interact with information when they are studying within a formal learning environment. It explores the origins and growth of the idea of information skills development, the translation of this work into frameworks and sequential models and the adaptation of these models to take account of changes in information storage and transmission brought about by the Internet. The chapter then examines the changing contexts and changes in learning being brought about by the Web 2.0 environment and questions whether adjustment of existing information literacy models is a sufficient response to deal with these changes. We conclude that although Web 2.0 developments are not fundamentally undermining the nature of teaching and learning they do provide important possibilities for more effective information literacy development work. A non-sequential framework is offered as a contribution to supporting HE students when seeking to obtain, store and exploit information simultaneously in the informal social world of Web 2.0 and in their formal academic discipline.

DOI: 10.4018/978-1-60566-208-4.ch022

THE RISE OF INFORMATION SKILLS

In the early 1980s a spate of books appeared in the UK containing a new term in the title: 'information skills'. This term was the brainchild of a working party concerned about school pupils' competence in *"using libraries, exploring references and making notes"* (Marland, 1981, p7) and arose out of the Schools Council's desire to explore what a curriculum for a changing world might comprise. The working party report asserted that *"Individuals today have an increasing need to be able to find things out…never before have our lives depended so much on our ability to handle information successfully"* (Marland, 1981, p9). Narrow concerns about library skills and user education were replaced by a focus on students' problems in finding and using information to tackle assignments and conduct their research within a formal learning environment. This intervention was due to the interest in these skills by educationalists, who, working alongside librarians, ensured wider adoption for information skills and a clearer place for the concept within the learning process.

However, despite this development and the appearance of a number of books exploring the place of information skills in learning (see, for example, Markless and Lincoln, 1986, and Wray, 1985) the concept of information skills was far more widely accepted by librarians than by teachers. This resulted in heavy emphasis on competence in resource use and on finding information.

MODELS OF INFORMATION SKILLS

From the outset writers wanted to show the need for students to develop these 'new' information skills. The issue was presented as one of skills deficit and consequently led to a plethora of information skills frameworks and models, spelling out what students should be able to do. (Many of these models were later 'rounded up' and described by Loertscher and Woolls, 2002.) Model constructors conceived the requisite process as tying together distinct elements of information-related behaviour into a logical, sequential process which could then be taught (e.g. Marland, 1981; Brake, in Markless and Lincoln 1986).

An important retrospective review of these models and frameworks (Eisenberg and Brown, 1992) concluded that

while each author may explain this process with different terms … all seem to agree on the overall scope and the general breakdown of the process … it appears that the various process models are more alike than different and it may be possible and desirable to begin speaking about a common process approach to library and information skills instruction. (p. 7)

The approach to information skills as a 'common process' to be applied to library research and information handling unfortunately tended to result in a disregard for the context of learning. Skills were perceived as generic; the sequential process outlined in the models was to be adopted at all ages and across different subjects. The process formed a 'curriculum' to be taught to students and applied by them whenever necessary. This view was hardly challenged in the early world of information skills although research on information behaviour in context and on critical thinking skills was calling into question the whole notion of easy transfer, which is also a well-established assumption in mainstream education (Perkins and Salomon, 1992).

Perhaps the most influential of these generic information skills models was advanced as the Big6. This model was created by Eisenberg and Berkowitz (1990); it was widely disseminated in book form and continues to be heavily promoted in the USA and internationally through their website and through an extensive programme of workshops. We will use this Big6 framework as the basis of our critique for the remainder of this chapter because it is one of the frameworks most widely used in USA and UK schools to support information skills teaching and because its authors were amongst the first to integrate ICT into information skills in a distinct and transparent manner.

THE BIG SIX SKILLS™ APPROACH

The main elements of this model are outlined below:

1. Task Definition: (determine the purpose and need for information)
   - Define the problem
   - Define the information requirements of the problem
2. Information Seeking Strategies: (examining alternative approaches to acquiring the appropriate information to meet needs)
   - Determine the range of possible resources
   - Evaluate the different possible resources to determine priorities
3. Location and Access: (locating information sources and information within sources)
   - Locate sources (intellectually and physically)
   - Finding information within resources
4. Use of Information: (using a source to gain information)
   - Engage (e.g. read, hear, view) the information in a source
   - Extract information from a source
5. Synthesis: (integrating information drawn from a range of sources)
   - Organize information from multiple sources
   - Present information
6. Evaluation: (making judgements based on a set of criteria)
   - Judge the product (effectiveness)
   - Judge the information problem-solving process (efficiency)

   Eisenberg and Berkowitz (1990)

It is not surprising that when the concept of information skills was new, and people sought to understand its scope, frameworks such as the Big6 were widely adopted. They provided a foundation on which to build learning activities and assessment. Would such frameworks survive intact into the 'information age' of ICT?

THE SHIFT TO INFORMATION LITERACY: A BROADER VIEW?

With the advent of Worldwide Web and the extensive accompanying investment in ICT in educational institutions of all kinds, concerns about students' ability to find and use information grew exponentially and a new vocabulary began to emerge in formal education - that of information literacy. The notion of information literacy developed in the USA in the 1980s in response to a move towards more active learning in universities and the concomitant need to move away from terms implying passive instruction (Martin, 2006). Use of the term expanded considerably in the 1990s (Bawden, 2001) and has gained some worldwide influence, leading to a declaration by UNESCO (2003) stressing the global importance of information literacy within the information society. A parallel growth in the UK has seen the term widely adopted in academic libraries and national educational bodies (but with most school libraries until now still preferring to focus on information skills - Streatfield and Markless, 2007).

Did the new term signify any fundamental change in thinking or signal a new characterisation of the skills or processes previously called information skills? National Information Literacy Standards in Australia (CAUL, 2001) and the USA (ACRL, 2000) echoed much of what was in the earlier process models, as did the information literacy model proposed in the UK by the Society of College, National and University Libraries (1999). Despite the fact that 'literacy' is a problematic and contested concept (it has been variously described as encompassing notions of functional competence and skills, of sets of wider cognitive abilities, and as part of a contextualised approach to learning in its social and economic context - Bowden, 2001), information literacy was usually reduced in presentation to a series of skills, procedures and technicalities. This inhibited approach attracted some criticism for being too mechanistic and some writers moved towards a conceptualization that includes attitudes, underpinning knowledge and meta-cognitive abilities (Kuhlthau, 1988; Bruce, 1997). Although Kuhlthau recognised the importance of student attitudes and emotions in her information process model, these elements have not been integrated into other process models - although the commentaries accompanying these models usually refer in some way to motivation and attitudes.

INFORMATION SKILLS AND THE INTERNET

In this phase of its development, the Internet was viewed primarily as a new information storage and delivery system for which existing information skills frameworks could simply be expanded or adapted to take account of the growth in access to information via the Internet. Eisenberg and Johnson (1996) exemplified this view when they explicitly integrated ICT into the Big6 Skills model, saying that

Students need to be able to use computers flexibly, creatively and purposefully… (they) should be able to use the computer as part of the process of accomplishing their task. (p. 2)

During the 1990s, the creators of the Big6 confidently extended the model to include student use of ICT when solving learning problems. They claimed that various computer and information technology skills were integral parts of the Big6 Skills. This claim was sustained as their model continued to be implemented in schools across the USA and the UK (Eisenberg and Berkowitz, 2000). Adherents of this and other process models confidently asserted that the basic principles of information seeking and use, derived from years of watching and helping students to interact with print-based information, remained unchallenged.

We have chosen to exemplify current process models by citing the Big6 when looking at whether the concept of information literacy needs to be repackaged or reconceptualised because:

- the model crystallizes the general process approach favoured until now and serves as an adequate exemplar of the model-driven approach
- it serves our purpose because it was the only model advanced until recently that systematically encompasses the ICT dimension
- It is still currently being used and promoted in that form.

The Big6 framework is useful for this purpose because it is a systematic and widely adopted model. Our comments should not be construed as an attack on this particular framework.

EXAMINING THE PROCESS MODELS

What are the assumptions underpinning the Big6 and similar models and what are their main characteristics?

- A sequential view of the process of student research, conceived as a series of logical steps
- Use of prescriptive language to convey an 'ideal approach' to information-seeking and use (e.g. "After students determine their priorities for information-seeking they must locate information from a variety of sources"; "once the information problem has been formulated, the student must consider all possible information sources and develop a plan for searching").

This approach is commonplace in this period, despite the warning offered a decade earlier by Tabberer and Altman (1986) about the danger of idealising study behaviour and promoting 'the right way to …' They stressed that success came by diverse routes and as a result of different choices made in different situations. They warned that students did not always gain much by being confronted with 'the ideal' because there is a range of influences that prevent adoption of 'best behaviour'.

- The process models were designed to support information skills teaching (i.e. to provide a 'curriculum' for the teachers and a pathway to be followed by students when doing their research).
- A particular and limited conception of information-related behaviour is represented in these models, with much emphasis on information seeking, location and access. Use of information is reduced to determining relevance and extracting pertinent items of information (by taking notes or resorting to cut and paste). The words knowledge, understanding and making sense of, seldom occur in these models, nor does the idea of creating one's own viewpoint. The apparent assumptions are that this shortcoming will be addressed in the subject teaching or that the acts of extracting and organising relevant information will themselves stimulate the construction of meaning. What happens instead is frequently cut and paste activity leading to more or less unintentional plagiarism. In these models, synthesis is not about transforming information to encapsulate new knowledge
- Overall they present ways to support teaching ("innovative instructional methods") designed to provide a framework to guide teachers or librarians when preparing appropriate activities or tasks for their students.

These models reflected the main uses conceived for the Web in this period as a vehicle for storing and transmitting information.

INFORMATION LITERACY AND WEB 2.0: CHANGING THE CONTEXT, CHANGING THE LEARNING?

The 'orthodoxy' of information skills within formal learning environments, as enshrined in the Big6 Model, is being increasingly challenged. Recent research into information literacy is moving away from technological processes and skills-based models, recognising the complexities inherent in finding and using information. A more experiential perspective that recognises the contextual and affective elements of information literacy is emerging (Williams and Wavell, 2007). Two complementary developments have influenced this shift in focus: greater interest amongst information literacy researchers and practitioners in the processes of learning (especially theory about variation in learning and constructivist approaches); and an electronic environment that is increasingly being shaped by its users.

Have traditional views of information literacy really been rendered obsolete? Does learning through Web 2.0 require different skills and abilities? Are a new range of cognitive and metacognitive strategies needed to learn effectively within the Web 2.0 environment? Or, does the Web 2.0 environment provide tools that enable teachers to engage students more effectively in well-established learning processes than could be achieved hitherto?

In our view, learning is not fundamentally different within Web 2.0, nor does the 'new' social software change the basic processes of learning. Where Web 2.0 has made a difference is in making it easier to engage with some aspects of learning that were previously difficult to address (for example, real collaboration and groupwork, peer critique, hearing students' authentic voices and construction of new knowledge). None of these important aspects of effective learning are new: all can be found in the education literature of the 20th Century, from Dewey to Ausubel, and from Vygotsky to Marton. However, despite their importance, few of these elements have found their way into information literacy models or practice.

When the Worldwide Web was primarily a vehicle for storing and delivering information it was easy to portray information literacy as an ordered sequence of skills to be ***transmitted*** to students, whilst ignoring other approaches to learning. Web 2.0 effortlessly undermines this approach with its disregard for authority, hierarchy and order and its focus on the voice of the individual and on ever changing constructed groups. Any contemporary approach to information literacy must consider how to engage more effectively with learners, by understanding these multiple aspects of how they can learn.

Before we examine in a little more detail some of these key elements of learning and their relationship to information literacy and social software, we need to note two other factors that may influence this relationship: the reluctance of individuals and institutions to change; and the ways in which the 'Google generation' of 'digital natives' may interact with information and learn in new and different ways. What are the key elements of learning as they relate to information literacy and social software? Some at least of these key elements are:

1. Reluctance to change (institutions and teachers)

Faced with the unfamiliar challenge of a new world of social networking, some education institutions have tended to react in a predictably conservative way by blocking access to elements such as Face book and Second Life. As a result of such embargos, as well as a reluctance by teachers to engage with this new world, students are frequently operating in different electronic environments during formal learning from those in their out of hours experience (especially in schools). This makes teaching of information literacy more problematic.

To somewhat over-dramatize the dilemmas created: as a teacher, how can you fully engage with students in helping them to exploit information if you don't have easy access to what may constitute their major sources of information? Or, from a student perspective, why should you bother to engage with all this 'information literacy stuff' if your perception is that all you have to do to get the information and help that you need, is to resort to your social networks? When you are away from the institution, if you can effortlessly manipulate multi-media information to build your own web pages, why jump through what might be seen as sterile and irrelevant information literacy hoops when you are in formal learning mode? Again, as the world of Web 2.0 becomes increasingly sophisticated, the version of ICT encountered in formal learning is likely to appear ever more limited and pedestrian.

2. Digital natives and others

" Future students in higher education belong to a generation that has grown up with a PC mouse in their hands, a TV remote control, a mobile phone, an i-pod, a PDA and other electronic devices for communication and entertainment … computer games, the Internet, MSN, wikis and blogs being an integral part of their lives" (Veen, 2007, p.1). Prensky has labelled these young people 'digital natives' and has asserted that they now exhibit different characteristics from their forbears (the digital immigrants) due to the extent of their exposure to technology in all its forms.(Prensky, 2001). He claims that changes in activity during development may result in different neural wiring via processes of 'neuro-plasticity'; a view recently echoed by Martin Westwell of the Institute for the Future of the Mind (2007). Both advocates assert that current students have much better visual skills, do better at visual-spacial tests, are able to deal with lots of information at once, and can process this information and make decisions quickly. On the other hand, this generation of students may have shorter attention spans, be easily distracted, may not maintain focus well when interrupted and may have less ability to reflect on topics than the previous generation. Veen (2007) adds to this list of differences, talking about non-linear learning behaviour; clicking and zapping to deal with information overload; using exploratory approaches to new situations; and becoming experienced at problem solving at a young age. “We now have a new generation with a very different blend of cognitive skills than its predecessors – the digital natives.” (Prensky, 2001)

As a result of Web 2.0 developments, we can also anticipate that ‘digital natives’ may have different social skills.. This is because the Internet is increasingly used for socialisation rather than just information-seeking, with even those seeking information often doing so via peer groups. Westwell claims that more people use Second Life and Facebook than use Google. Whether or not we believe all these claims, Oblinger and Oblinger (2005) have forecast that the next generation of students entering higher education will be digitally literate, highly Internet-familiar, connected via networked media, used to immediate responses, and preferring experiential learning. This generation will be highly social: they will prefer to work in teams and will crave interactivity in image-rich environments as distinct from text-intensive environments.

Where does this leave traditional information literacy, with its focus on using libraries and finding primary sources, its reliance on laborious sequential steps and its scant reference to collaboration or to multi-media resources? If Westwood and others are correct, their picture of our ‘new’ students implies that not only have they gained from their early digital experiences but they have also lost in terms of opportunities for reflection and ‘slow-learning’. This picture of gains and losses calls into question the widespread claims that elements of Web 2.0 (wikis etc.) automatically help to develop meta-cognitive skills. However, it is also interesting to note that traditional information literacy frameworks do not emphasise reflection and its role throughout learning.

WEB 2.0, INFORMATION LITERACY AND FORMAL LEARNING

Where do all these changes leave information literacy? How might traditional models of information literacy need to be altered to accommodate the experience and expectations of students within formal education? Where does Web 2.0 fit in?

- The sequential view of skills deployment is now being questioned. Learning tasks make a range of different demands on students, which call into question the notion of applying the same series of steps to meet all these demands. Observations of pupils from 5-18 in schools and students in further education colleges show that they seldom follow the prescribed sequence (Streatfield and Markless, 1994; Moore, 1997; Markless and Streatfield, 2000). Formal studies of information-seeking behaviour in universities again challenge this premise (Foster, 2006). To be fair, most of the process models that are set out in steps are accompanied by some form of caveat recognising or even advising that it is not necessary to follow the prescribed sequence. However, there is usually little help offered on how to use the model in a non-sequential way, with the result that the framework tends to be taught as a sequence. The desire to inflict sequences on students is remarkably resilient in the world of information literacy. Even writers who are responding to the Web 2.0 environment tend to present a sequence of processes to be learned in order to become 'information fluent' (e.g. the five-stage process of Jukes (2007): asking questions; accessing data; analysing and authenticating information; applying it to real-life problems; assessing product and process). This approach takes no account of the influence of context on any sequence, the influence of learners' cognitive styles, or the need to make sense of any information and transform it into knowledge.

In addition, a core characteristic of Web 2.0 tools is that they transfer power, ownership and authority to the participants. This inevitably gives people license to design their own routes through learning tasks in any way that suits them. Finding information is less likely to involve systematic information seeking than, for example, interest groups, peer web pages or social bookmarking.

These observations lead to the key question - can the Big6 or any similar information literacy model be adapted to take account of how students actually find and use information, especially in the Web 2.0 environment?

- Although the importance of learning as construction is recognised within the rhetoric of information skills pedagogy and "Information literacy is often seen as the school library version of constructivism" (Moore, 2005 p.3), much of the observed planning and practice[1] suggests heavy reliance on transmission, learner practice, and feedback, all heavily structured into manageable segments and strongly 'teacher' controlled (that is, the classic behaviourist approach). Early voices such as Kuhlthau's (1993), which present information-seeking as a process of seeking meaning, were at first largely ignored in practice. In recent years there have been intensified efforts to ensure that people who are teaching information literacy adopt constructivist approaches (e.g. Todd, 2001). Limberg (2007) asserts that to learn is not to receive knowledge and information, but is about changing the relationship between a person and the world. She claims that information-seeking is too often focussed on teaching technical procedures and on fact-finding rather than on students formulating authentic questions and constructing their own positions. The concept of authenticity is central to Limberg's ideas on information literacy. Contrived questions and tasks, designed solely to meet externally imposed assessment and with no other consequences for the student, will not engage and motivate students. Without a real and personal interest, students will be satisfied with the superficial answer, the first 'hit', or 'good enough' information. There is no incentive to go beyond using technical skills to collect facts.

Again, the latest outputs from the USA-based Center for International Scholarship in School Libraries (Kuhlthau and others, 2007) focus on the concept of 'guided inquiry' as the basis for teaching and learning of information skills. The main characteristics of guided inquiry are:

- active engagement by students in the learning process
- students building on what they already know
- high levels of reflection
- a recognition of the importance of social interaction and of students' different ways of learning

  (Kuhlthau and Todd 2007)

All these are recognisable characteristics of learning as construction (see, for example, Papert and Harel, 1991). There is little doubt that constructivist approaches are particularly suited to

Web 2.0 tools. In this environment, students can construct artefacts such as video presentations, blog entries and wiki pages both individually and collaboratively. Teachers can join in with collaborative editing and can scaffold students' work. It seems likely that the constructivist approach to teaching and learning so well supported by Web 2.0 tools may finally lead to information literacy teaching becoming more attuned to how students learn.

If constructivist principles are used to inform and guide information literacy work, students will be required to develop a repertoire of strategies that are conspicuously absent from most information literacy models. This will involve:

- reflection: the ability to reflect constructively and to use that reflection in planning for their own development
- evaluation of the processes undertaken as well as of the products of their study
- making sense (deep understanding) of the information that they obtain, linked to the ability to transform the information to reflect their own emerging views

We do not think that these aspects of learning can simply be grafted onto existing frameworks or inserted after any particular element of a linear, sequential model. They are part of an iterative process of learning not well represented in existing information literacy frameworks.

THE IMPORTANCE OF CONTEXT

The importance of context in relation to information behaviour is well established (e.g. Streatfield and Wilson, 1980; Dervin, 1992; Ingwersen and Jarvelin, 2005). Context in information-related behaviour is recognised as multi-dimensional: with different facets reflecting features of the task; characteristics of the learner; and features of the system. Louise Limberg observed in a conference presentation that "Influential studies have abandoned the idea of information literacy as a set of generic skills applied anywhere. Information literacy is not generic but should be seen as social practice …" (Limberg, 2007). Looking at secondary schools, Williams and Wavell (2007) warned that if we are trying to to develop pupils' information literacy we cannot ignore content in favour of technicalities and procedures - if we do so, we will get trivial learning outcomes. Nevertheless, as we have already noted, information literacy advocates have persisted in offering generic skills development frameworks that take little no account of context.

How can the importance of context be reflected in an information literacy framework? We believe that a different type of framework is needed; one that moves away from offering a list of abilities to be taught or applied in an unvarying sequence, irrespective of context.

Alongside the challenge of producing an appropriate information literacy framework we face another problem: how can we teach information literacy in ways that respect the influence of context? Current views on skills development (e.g. Luke, 2006; Williams and Wavell, 2006) assert that if students are to develop their information-related skills through assignments there is a need for:

- Authentic tasks that are recognised as relevant by the students (tasks that have meaning to students on a personal or academic level; not contrived to allow them to practice particular skills)
- Immersion in authentic contexts (realistic environments, current information drawn from the real world, engagement with real world problems and concerns)
- High quality tasks related to current academic work (e.g. asking students to conduct critical evaluation of sources to construct a position for an essay, rather than

offering general guidance on evaluating information)
- Learning embedded in the relationships, values and discourse of the learning community (inherently social)
- Timely teacher interventions in order to move learners on at transition points in their work

Web 2.0 can once again be a powerful support for increasing authenticity and enabling the deployment of information literacy strategies in a variety of meaningful contexts. The possibility of a public platform for their work may help students to take more seriously the underlying information literacy processes involved in producing that work.

STUDENT REFLECTION

If we are to take context into account when deciding on information literacy strategies, this immediately introduces the concept of variation. Bowden and Marton (1998) argued that not only do students need to experience variation in order to learn, but they must also explore variation by comparing and analysing their experiences. To do this, students need to:

- actively engage in discussion and reflection about finding and using information in order to uncover variation in their conceptions
- confront variation in their own experience and in the experience of others.
  (Based on Bruce, 2007, pp. 51-52)

Since at least the 1970s, reflection has been seen as a mainstay of learning and this concept has found its way into many models of learning (e.g. Kolb, 1975; Schon, 1983).

Reflection is a particularly important element in developing the processes underpinning learning and is therefore potentially important in any systematic approach to information literacy. Reflection is taken for granted in most models of information literacy or placed at the very end of the process. This approach in not likely to enable the development of the meta-cognitive strategies necessary to perform problem-solving with information. It is likely to be difficult to integrate reflection into existing information literacy frameworks in any meaningful way (see the discussion about constructivism above). The possibilities for learning provided by Web 2.0 may provide a way forward. For example, peer critique and the collaborative production of artefacts may automatically stimulate reflection. If not, engagement in these processes should provide opportunities for a more formal emphasis on reflection as part of information literacy teaching.

COLLABORATIVE LEARNING

Collaborative learning has long been seen as a desirable process: for example, groupwork is a key element of training courses for teachers in all sectors. Web 2.0 tools have turned many students into sophisticated social networkers via YouTube, Facebook, blogs and discussion boards (Ipsos MORI, 2007). The same tools can also be used to facilitate collaboration in formal learning settings, whether the focus is on creating specific interest groups, building learning communities or enabling the collaborative production and editing of artefacts.

Collaborative learning requires many skills of communication and interaction, but does it make fundamentally different information literacy demands on learners than those made when individually finding and using information? There is little in recent research to indicate that this is the case (Williams and Wavell, 2007; Kuhlthau, 2007). The influence of context (subject, learner characteristics and teacher expectations) is not just about whether students are working individually or in groups to find and use information. At the same time, Web 2.0 can be seen as working counter to collaboration through increased personalisation of learning paths. Overall, this aspect of Web 2.0 raises important issues in the wider context of approaches to learning by providing increased scope for a variety of activities. It may offer valuable avenues for the teaching of information literacy but does not seem to fundamentally affect the information handling skills required.

LEARNERS' EXPECTATIONS OF INFORMATION

Web 2.0 inevitably raises questions of ownership and authority of information. It is an environment in the course of creation by its participants. These participants individually and collaboratively generate content in a form, format and structure that best suits their own needs and preferences. This process works well when the primary focus is on participation in social networks or developing personal interests. However, it can create major difficulties when the same processes are applied in formal learning. Keen (2007) claims that we are diving headlong into an age of mass mediocrity because of the absence of gatekeeper expertise and the increase in user-created content. This view is echoed by Gorman in his Britannica Blog (2007) which identifies an erosion of traditional respect for authenticity and expertise in a world in which everyone is an expert "ignorant of the knowledge they will never acquire and the rich world of learning that search engines cannot currently deliver to them."

Most students should be able to operate both in the social world of web 2.0 and in more formal learning environments (even before we take account of the growing presence of academic interests and institutions on Web 2.0). However, to operate effectively in formal learning environments, student autonomy may have to give way to recognised academic authority. Students' preferred use of Wikipedia and social bookmarking, alongside their facility in creating new 'knowledge' through remixing text, image and audio, or through the collaborative creation and editing of web pages may come into conflict with the necessity to conform to academic norms of using externally-validated information. Students will not be able to simply replicate their social/leisure on-line behaviour when engaging in formal academic tasks. Information literacy should help in this arena: traditional information literacy models do focus on evaluating sources of information, on considering authority and credibility. Such an emphasis should raise students' awareness of the problems associated with following their own preferences and concentrating on their own perspectives. A new balance may need to be drawn between encouraging students to use the range of pathways to information that are open to them in Web 2.0 and ensuring that they have the ability to choose the most appropriate for academic study.

However, do we also need to respond more positively to students' expectations of information? Should the information literacy field legitimise elements of students' preferred information-related behaviour? For example, should we ensure that information literacy frameworks encompass such concepts as 'good enough' information, trial and error, and peer 'expertise' rather than focusing primarily on a set of competencies that appear to be designed to turn all learners into systematic researchers, regardless of the task context?

DOES FINDING INFORMATION REALLY MATTER ANY MORE?

One question likely to worry traditional information literacy proponents is whether there will be a continuing need for skills in information seeking, given an information world in which search engines are become increasingly sophisticated and in which Web 2.0 offers a range of enticing alternatives to systematic searching. According to Carol Kuhlthau (2007) what is important in the 21st century is the ability to use information for problem-solving ***not*** "the technology of finding."

IS A NEW MODEL OF INFORMATION LITERACY NEEDED TO MEET THE CHALLENGE OF WEB 2.0?

We are not convinced that the Web 2.0 environment on its own necessitates the development of new sets of abilities for finding and using information. It does, however, move learning into new directions (e.g. increased collaboration, more authentic tasks, peer critique, non-linear aaproaches to information). In doing so, learning with Web 2.0 tools should put increasing pressure on proponents of information literacy to move in the direction of well recognised learning principles and practices. In particular, information literacy can be enhanced in a formal learning environment by exploiting some possibilities offered through Web 2.0 tools:

- Enhanced group work and shared tasks
- Cooperative creation of multi-media artefacts
- Collaborative editing and critiquing
- Searching for information (e.g. using social bookmarking and folksonomies)
- Organising information in new ways (e.g. using tagging)
- Increasing authenticity of work by presenting ideas to others in a more public space and using a wider range of media
- Providing 'just-in-time' scaffolding to support students
- Facilitating student reflection using records of individual and group processes and providing virtual contemplative spaces

None of these aspirations are new to formal education but some have been difficult to achieve hitherto without the benefits of advances in Web 2.0.

If the information literacy community is prepared to design materials, activities and support mechanisms based on the opportunities offered by Web 2.0, can they adapt existing information literacy frameworks to scaffold their work? Is a framework needed at all to enable information literacy development in formal education settings?

Any model or framework will be flawed because it cannot fully take account of the influence of context on information use or the problems inherent in producing any generic view of information literacy. However, whilst doing research and development work in many further and higher education institutions and schools, we have found that staff and students want to put some sort of framework in place. They want a public statement that clarifies what is encompassed by information literacy; a guide to support curriculum planning; and something that students can refer to when doing research and tackling academic tasks.

The following framework (Markless and Streatfield, 2007) was originally designed to address problems being encountered by the University of Hertfordshire. The University was trying to develop an institution-wide approach to supporting students when finding and using information in an electronic environment. At first it was thought that an existing framework could be used or adapted to meet the needs of staff and students. However, consideration of the issues explored in this chapter made the shortcomings of such an approach apparent. We concluded that many of the traditional information literacy models had been built on a series of assumptions about learning and information behaviour that were problematic and that the increasing use of Web 2.0 threw these assumptions into stark relief. We therefore needed to offer a different solution that is more in keeping with the changing learning environment.

The solution offered is essentially a framework to support student choice in learning rather than information literacy teaching. The framework is designed to enable students to get help where and from whom they need it rather than to usher them through a regimented programme of information skills development. Some of the individual elements of the framework hark back to those designed in the 1980s and 90s. The skills and strategies included are not all new, although we have moved away from a heavy emphasis on systematic searching. In addition this framework is designed to be approached and used differently from traditional frameworks such as the Big6. The drivers behind our approach are student choice and reflection to support effective learning rather than laying out a sequence of steps to be taught.

During three key stages (which do tend towards the sequential) students ***choose*** which strategy to adopt at different points in their research. Help and guidance is available for each of the key elements. Importantly, if one avenue fails students can go back to the big picture and choose another route; they are not trapped in a sequence that they are expected to follow. The framework is designed for students to construct their own problem-solving approaches to finding and using information, either individually or collaboratively. The impact of context on learning should lead students to make different choices about which strategies to employ and which skills to draw on depending upon the nature of the task they are addressing and the wider social context in which they are operating. The framework is designed to take advantage of technological developments that allow individuals to make choices, navigate between options and then save their search paths for future reflection.

The framework that we designed drew on two research-based published models, a non-linear model of information-seeking behaviour[2] devised by Allen Foster (2004; 2006) and a model of information and critical literacies offered by Ross Todd (2001). Foster worked with academics to show the fallacies inherent in the assumption that researchers looked for information using a fixed sequence of steps. Todd's overview of information literacy emphasised transformation and construction of knowledge because he wanted to encourage students to stop interpreting research tasks or assignments merely as processes of collecting information. Instead they are encouraged to think in terms of forming their own perspectives, creating new insights and presenting their own authentic voices.

Our new framework builds on these ideas as well as addressing some of the concerns discussed earlier in this chapter. It is presented below as a series of figures.

Figure 1 provides students with an overview of what is involved with finding and using information. It was important to avoid the trap of presenting information literacy as a series of steps (anything between 5 and 12 stages in many traditional frameworks. The Big6 actually contains 12 steps as there are two stages in each of the 6 main elements of the framework.) Students can choose to engage in any one of the 3 main elements depending on the nature of the academic task they are tackling.

Figures 2, 3 and 4 are what the students see when they click on the relevant box in figure one. There is no set path through any of these figures; if the student is at the beginning of a project they may look at figure two and decide that the best place to begin is with networking. If they click on networking they will get some ideas about who they might contact and how they might work with peers. Equally, a student might decide to begin with problem definition. Again, this element is populated with activities, ideas and advice about how to analyse a problem. Whatever path a student or group of students decides to take through this framework they are offered on-line support at the point of need. They are also encouraged to save their path so that they can see how they worked through a particular assignment - the choices they made. Students compare their chosen paths

Figure 1.

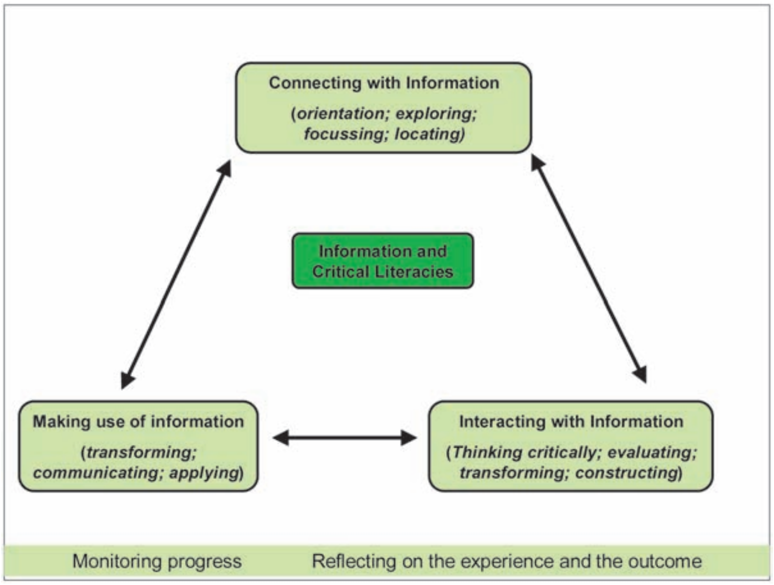
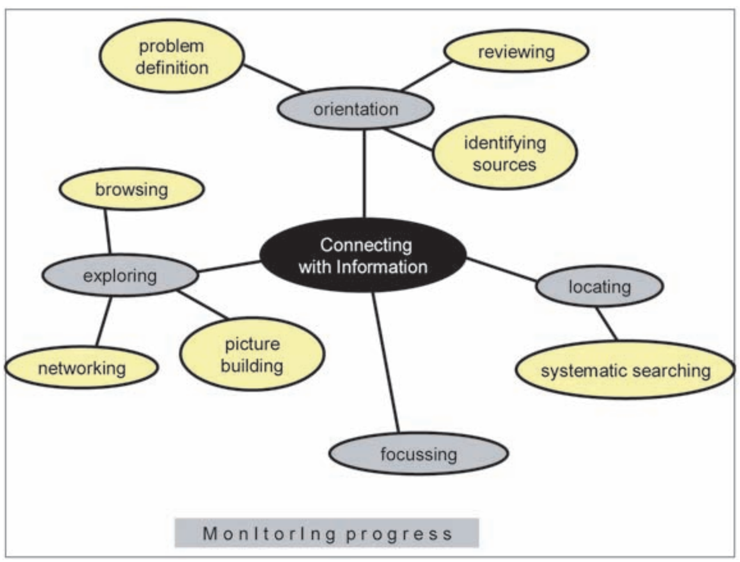


Figure 2.



with those of their peers and consider which have been most effective. They can also review their approaches across a number of different tasks to see whether and how they have adapted their approach to the context. Encouraging reflection on the process of learning was an important element in our design of a non-sequential framework.

Where our framework is being used, each element is being populated with material designed by both academic staff and librarians. The framework itself has stimulated some useful collaborative work between staff interested in specific elements most relevant to their discipline. Their aim is to provide support for students who want to use

Figure 3.

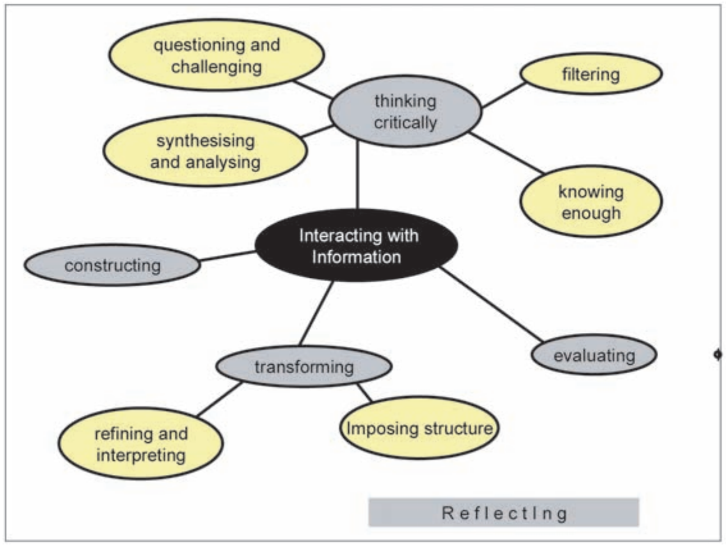
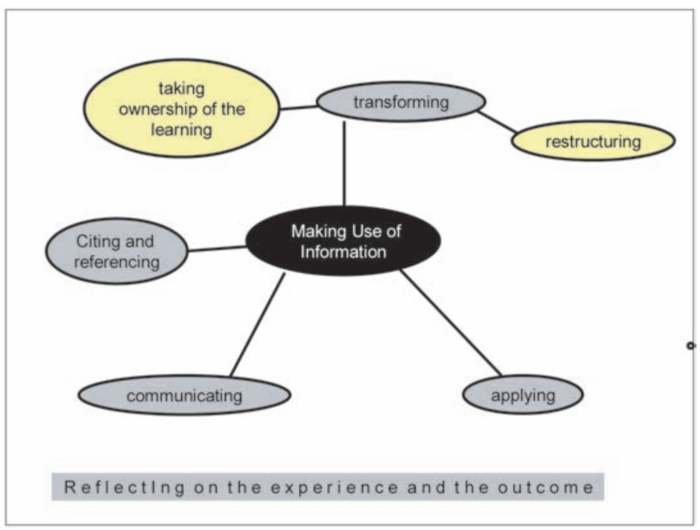


Figure 4.



a particular information skill or strategy in the course of their academic work without dictating a path through that work.

This framework is very much a work in progress. It is being tested at universities in the UK and Spain and is currently being translated into Arabic. The real test for this framework will be if students do not see it as a straightjacket, constraining their normal interactions with information, but find it useful no matter what the information environment in which they are working.

CONCLUSION

The traditional information skills models (such as the Big6) that grew out of early interest in enhancing the use of libraries and text-based resources tended to be based on a simplistic view of learning and information behaviour. These models served a purpose in introducing students to the formal world of academic information and, as such, continued to be of use when responding to the arrival of the Internet in its original conception as an information storage and transmission vehicle. However, the shift in focus towards ever greater information access through the Internet combined with greater attention to teaching and learning of information literacy based on constructivist education principles, has led to the traditional information skills approaches being increasingly questioned.

The changes being brought about by the advent of Web 2.0 have served both to provide a further challenge to traditional approaches to information literacy and potential solutions to some of the inherent problems in traditional approaches. The value of teacher-led, text-focussed, sequential models must now be in question because they are based on very un-web 2.0 propositions. Instead, the social networking possibilities offered by Web 2.0 provide fresh opportunities for supporting social learning, including peer information seeking, evaluation, critique of strategies and capturing of processes when helping students to engage with information literacy development. Accordingly, a new framework is tentatively offered here as an alternative to the Big6 and similar models, with the intention of allowing students to construct and revisit their own paths to information discovery, organisation, sense-making and exploitation in the evolving world of Web 2.0.

REFERENCES

Association of College and Research Libraries (ACRL) (2000). *Information literacy competency standards for higher education.*

Bawden, D. (2001). Information and digital literacies: a review of the concepts. *The Journal of Documentation*, *57*(2), 218–259. doi:10.1108/EUM0000000007083

Bowden, J., & Marton, F. (1998). *The university of learning: beyond quality and competence in higher education.* London: Kogan Page.

Bruce, C. S. (1997). *The seven faces of information literacy.* Adelaide: Auslib Press

Bruce, C. S., Edwards, S., & Lupton, M. (2007). Six frames for information literacy education: A conceptual framework for interpreting the relationships between theory and practice. In S. Andretta, (Ed.), *Change and challenge: information literacy for the 21st century.* Adelaide: Auslib Press.

Council of Australian University Libraries. (2001). *Information literacy standards.* Canberra: CAUL

Dervin, B. (1992). From the mind's eye of the user: the sense-making qualitative-quantitative methodology. In J. D. Glazier & R. R. Powell (Eds.), *Qualitative Research in Information Management.* Englewood, CO: Libraries Unlimited.

Eisenberg, M. B., & Berkowitz, R. E. (1990). *Information problem-solving: the Big Six skills approach to library and information skills instruction.* New Jersey: Ablex Publishing Corp.

Eisenberg, M. B., & Berkowitz, R. E. (2000). *Teaching information and technology skills: The Big6 in secondary school*s. New Jersey: Ablex Publishing Corp.

Eisenberg, M. B., & Brown, M. K. (1992). Current themes regarding library and information skills instruction: research supporting and research lacking. *SLMQ 20*(2) (Winter) http://archive.ala.org/aasl/SLMR/slmr_resources/select_eisenberg.html

Eisenberg, M. B., & Johnson, D. (1996). Computer skills for information problem-solving: learning and teaching technology.In *context ERIC Digest 1996, 4*

Foster, A. E. (2004). A non-linear model of information seeking behavior. *Journal of the American Society for Information Science and Technology*, *55*(3), 228–237. doi:10.1002/asi.10359

Foster, A. E. (2006). A non-linear perspective on information seeking. In A. Spink & C. Cole (Ed.), *New directions in human information behaviour.* New York: Springer.

Gorman, M. (2007). *The Siren Song of the Internet Part 2*.http://www.britannica.com/blogs/2007/06/the-siren-song-of-the-internet-part-ii/

Ingwersen, P., & Jarvelin, K. (2005). *The turn: integration of information seeking and retrieval in context.* New York: Springer

Ipsos, M. O. R. I. *(2007).* Student expectations study: key findings from on-line research and discussion evenings held in June 2007 for the Joint Information Systems Committee *JISC.*

Jukes, I. A. (2007). *Born to be wired: NetSavvy and communication literacy for an information age.* web.mac.com/iajukes/thecommittedsardine/Presentations.html [Accessed 15 December 2007]

Keen, A. (2007). *The cult of the amateur: how today's Internet is killing our culture and assaulting our economy.* London: Nicholas Brealey.

Kolb, D. A., & Fry, R. (1975). Towards an applied theory of experiential learning. In C. Cooper (Ed.), *Theories of group processes.* London: John Wiley.

Kuhlthau, C. C. (1988). Developing a model of the library search process: cognitive and affective aspects. *RQ* (Winter) (pp.232-242).

Kuhlthau, C. C. (1993). *Seeking meaning: a process approach to library and information services*. Norwood, NJ: Ablex; [Second edition published 2004].

Kuhlthau, C. C. (2007). From information to meaning: confronting the challenges of the 21st century. Keynote paper presented at *Information: interactions and impact conference*, Aberdeen June.

Kuhlthau, C. C., Caspari, A. K., & Maniotes, L. K. (2007). *Guided inquiry: learning in the 21st century.* New York: Libraries Unlimited Inc.

Kuhlthau, C. C., & Todd, R. J. (2007). Guided inquiry: a framework for learning through school libraries in 21st century schools. http://cissl.scils.rutgers.edu/guided_inquiry/characteristics.html [accessed 21 December 2007]

Limberg, L. (2007). What matters: shaping meaningful learning through teaching information literacy Presentation at *Information: interactions and impact conference*, Aberdeen June.

Loertscher, D. V., & Woolls, B. (2002). *Information literacy: a review of the research. A guide for practitioners and researchers* 2nd Edition Salt Lake City: Hi Willow Publishing.

Luke, A. (2006). On critical literacy: learning to question texts and discourses Keynote paper at *Bridging the Gap* Conference Yokohama November

Markless, S., & Lincoln, P. (Eds.). (1986). *Tools for learning* British Library R and D Report 5892 London: British Library Board.

Markless, S., & Streatfield, D. R. (2000). *The really effective college library.* Library and Information Commission Research Report 51 Twickenham, Middx. IMA for the LIC.

Markless, S., & Streatfield, D. R. (2007). Three decades of information literacy: Redefining the parameters. In S. Andretta (Ed.), *Change and challenge: information literacy for the 21st century.* Adelaide: Auslib Press.

Marland, M. (Ed.). (1981). *Information skills in the secondary curriculum: the recommendations of a Working Group sponsored by the British Library and the Schools Council*. London: Methuen Educational

Martin, A. (2006). Literacies for the Digital Age. In A. Martin & D. Madigan (Eds.), *Digital literacies for learning.* London: Facet Publishing.

Moore, P. (1997). Teaching information problem solving in primary schools: an information literacy survey. *J of Contemporary Educational Psychology*, *20*, 1–31. doi:10.1006/ceps.1995.1001

Moore, P. (2005). An analysis of information literacy education worldwide in School. *Libraries Worldwide*, *11*(2), 1–23.

Oblinger, D., & Oblinger, J. (Eds.). (2005). *Educating the net generation.* Educause. www.educause.edu/educatingthenetgen/

Papert, S., & Harel, I. (1991). *Constructionism.* New Jersey: Ablex Publishing Corp.

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In *International encyclopedia of education*, 2nd edition. Oxford, Pergamon Press.

Prensky, M. (2001). Digital natives, digital immigrants. [Accessed 21 December 2007]. *Horizon*, *9*(5). www.markprensky.com/writing/Prensky%20-%20Digital%20Natives.%20Digital%20Immigrants%20-%20Part1.pdf.

Schon, D. (1983). *The reflective practitioner.* New York: Basic Books.

Society of College, National and University Libraries. (1999). *Seven pillars of information literacy.*http://www.sconul.ac.uk/activities/inf_lit/sp/model.html [Published 1999; re-published 2004 - viewed December 21 2007]

Streatfield, D. R., & Markless, S. (1994). *Invisible learning? The contribution of school libraries to teaching and learning. Report on ... a research project* Library and Information Research Report 98 London: British Library.

Streatfield, D. R., & Markless, S. (2007). Information literacy. In J. H. Bowman (Ed.), *British librarianship and information work 2001-2005.* Aldershot, Hampshire: Ashgate 2007 (pp. 413-430)

Streatfield, D. R., & Wilson, T. D. (1980). *The vital link: information in social services departments.* London: Community Care and the Joint Unit for Social Services Research.

Tabberer, R., & Altman, J. (1986). *Study and information skills in schools.* London: British Library.

Todd, R. (2001). Transitions for preferred futures of school libraries. Keynote paper to International Association of School Libraries (IASL) Conference, Auckland, Symposium. http://www.iasl-slo.org/virtualpaper2001.html [Accessed 15 December 2007]. [Since developed by Professor Todd in various conference papers and presentations].

UNESCO. (2003). *Conference report of the information literacy Meeting of Experts.* Prague, September.

Veen, W. (2007). Homo Zappiens and the need for new education systems. Paper presented at the 2nd international convention *New Millennium Learners: Schools, ICT and learning* Florence. March

Westwell, M. (2007). Bending minds: how technology can change the way we think. Keynote paper presented at *Information: interactions and impact conference*, Aberdeen June.

Williams, D., & Wavell, C. (2006*). Untangling spaghetti? The complexity of developing information literacy in secondary schools*, Scottish Executive, Web publication of research report.

Williams, D., & Wavell, C. (2007). Making connections: the nature and impact of information mediation in the development of information literacy in schools. Paper presented at *Information: interactions and impact conference*, Aberdeen June.

Wray, D. (1985). *Teaching information skills through project work*. London: British Library. key terms

KEY TERMS AND DEFINITIONS

Constructivist Learning: Learning as an individual or social act of construction, leading to sense-making and the building of meaning.

Information Literacy: A set of abilities for seeking and using information in purposeful ways related to task, situation and context. (Limberg, 2007)

Information Skills: The sets of skills and competencies required to find and use information, usually in a formal education context.

ENDNOTES

1. This observation is based on our experience over twenty years of observing information skills/ literacy lessons and examining lesson plans.
2. Although Foster describes his model as non-linear, it may be more helpful to regard it as a non-sequential model

This work was previously published in the Handbook of Research on Social Software and Developing Community Ontologies, edited by S. Hatzipanagos and S. Warburton, pp. 316-334, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 7.16
Aspect-Oriented Framework for Web Services (AoF4WS): Introduction and Two Example Case Studies

Ghita Kouadri Mostefaoui
Oxford University Computing Laboratory, UK

Zakaria Maamar
Zayed University, UAE

Nanjangud C. Narendra
IBM India Research Lab, India

ABSTRACT

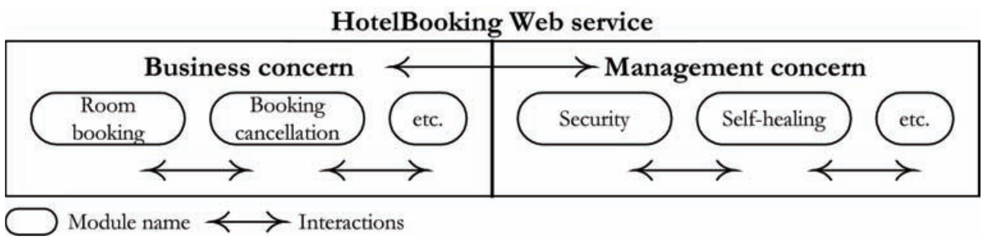
This chapter presents our research initiative known as aspect-oriented framework for Web services (AoF4WS). This initiative looks into the role of aspect-oriented programming in enhancing Web services with nonfunctional properties that are orthogonal to the primary functional properties of Web services, without the need for extensive reprogramming. This enhancement achieves a separation between the functional and nonfunctional aspects of Web services, thereby resulting in easier adaptability and maintainability. We have initially chosen to focus on security and self-healing nonfunctional requirements. The AoF4WS initiative is therefore demonstrated using two projects, SC-WS and SH-WS, which respectively stand for security concerns of Web services and self-healing Web services. Our contributions are relevant to the design phase in an aspect-oriented software development lifecycle.

DOI: 10.4018/978-1-60566-042-4.ch011

INTRODUCTION AND MOTIVATIONS

Web services are an attractive approach for implementing loosely-coupled business processes, which usually spread over companies' boundaries (Ma, 2005). Over the last few years several efforts have been put into the development of standards related to Web services definition, announcement/discovery, and composition, just to cite a few. The dynamic nature of the business world highlights the continuous pressure on businesses to reduce expenses, increase revenues, generate profits, and

Figure 1. Concern separation in a Web service



remain competitive. This calls for a quick reaction to the market trends, a quick handling of users' needs, a quick adaptation to unforeseen changes, and last but not least, a quick understanding of forthcoming challenges. To boost the acceptance level of Web services by the IT community as the technology of choice when developing flexible processes, Web services need to adapt to changing nonfunctional requirements with minimal reprogramming and minimal maintenance effort, so that they can be kept independent from the core Web services functionality. Security and self-healing are samples of nonfunctional requirements, and we will be highlighting them in this chapter.

Integrating security and self-healing capabilities into Web services calls for a clear separation between "business" and "management" concerns along which a Web service is defined (Figure 1). For this purpose, we adopt an aspect-oriented programming (AOP) approach to specify and implement this separation (Cottenier & Elrad, 2004; El-Manzalawy, 2005). This approach is part of our long-term research initiative known as aspect-oriented framework for Web services (**AoF4WS**). This initiative aims at examining the role of aspects in decoupling various concerns in Web services like security and self-healing. The separation between "business" and "management" sides emphasizes the noninvasive requirement that needs to be taken into consideration during the development cycle of a nonfunctional requirement. The mechanisms related, for instance, to security should be confined into one module and thus, should not scatter over the rest of modules of the Web service. Figure 1 illustrates the way concern separation occurs in a fictive Web service referred to as HotelBooking. The business side focuses on details directly related to hotel booking, like checking room availability, rate verification, and confirming client reservation. The management side of a Web service gathers all modules, such as security, self-healing, and monitoring that back the operations of this Web service. Constituents of the management side to be implemented as aspects need to be factored out of the core logic of the Web service.

In the following, we describe the two projects that we have chosen as part of the Ao4FWS initiative. The SC-WS project stands for security concerns of Web services and the SH-WS project stands for self-healing Web services. In Section 2 we present some basic definitions necessary for understanding the chapter. A motivating scenario is also presented in this section. SC-WS and SH-WS projects are described in Section 3 and Section 4, respectively. The chapter concludes in Section 5.

BACKGROUND

Some Definitions

Web Service

For the World Wide Web Consortium (W3C), a Web service is a software application identified by a URI, whose interfaces and binding are capable of being defined, described, and discovered by XML artifacts and supports direct interactions with other software applications using XML-based messages via Internet-based applications. Several standards are associated with Web services like ebXML registry services, Web service description language (WSDL), universal description, discovery, and integration (UDDI), simple object access protocol (SOAP), and WS-security (WSS).

Aspect-Oriented Programming

Ortiz, Hernández, and Clemente (2004a) define aspects as units of encapsulation that are built upon two elements: join points and advices. Join points determine the places or pointcuts where the behavior alteration of an application will happen, and advices identify the new code to be injected in response to this alteration. Aspect-oriented programming has emerged as a programming paradigm that allows gathering of a concern code into one single module. Security, performance, and logging are examples of such concerns that need to be separated from the code of the core application. This approach provides a clean separation between functional and nonfunctional concerns allowing fast update and maintenance of the application code.

Aspect-Oriented Programming for Web Services

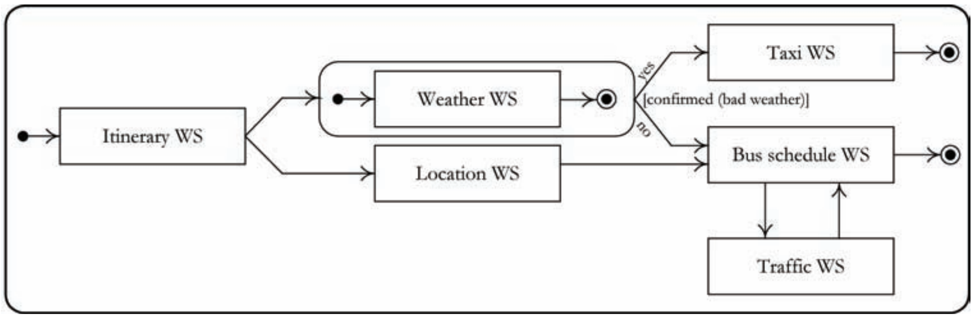
Cibrán and Verheecke (2003) promote modularizing Web services management with AOP. That was motivated because of the hard-wiring technique that is nowadays used for integrating Web services into applications. Hard-coding has several deficiencies when it comes to working out how to adapt to changes, what if a service or network fails, and how to deal with issues related to peers, such as checking for availability, switching to other services, and so forth. Charfi and Mezini (2004) apply AOP to workflow languages like business process execution language for Web services (BPEL4WS) in order to achieve modular and dynamic adaptability of Web services composition. This is done using aspect-oriented BPEL (AO4BPEL), which extends BPEL with additional features that permit, for instance, viewing business rules as aspects. In another work, Ortiz Hernández, and Clemente (2004b) adopt an aspect-oriented approach to develop solutions for Web services composition (of type orchestration) and interaction patterns. Their work was motivated by the lack of standards associated with composition. More particularly, Ortiz et al. raised multiple questions related, for instance, to the possibility of reusing interaction patterns previously implemented, and the efforts to put in for modularizing these patterns rather than scattering the code.

Motivating Scenario

In the following, we detail a motivating scenario, trip composite service (T-CS_1), that will be used for illustrating the two projects discussed in this chapter. The scenario is about Amin who is visiting Melissa back in her home city, Oslo. They agree to meet in a coffee shop, not far from Melissa's office since she finishes work late on that day. Amin has two options to reach the meeting place, that is, by taxi or by bus. Figure 2 illustrates the specification of Amin scenario using a combination of state chart diagrams and service chart diagrams (Maamar, Benatallah, & Mansoor, 2003).

At his hotel, Amin browses some Web sites about transportation in Oslo. A site has itinerary WS that proposes routes between two specific places, for example, between Amin's hotel and the coffee shop. The proposed routes are subject to weather forecasts: cold weather results in recommending taxis, otherwise public transportations like tramways and buses. Parallel to consulting with weather WS itinerary WS requests details about the origin and destination places using location WS. The use of location WS is highly appreciated by Amin since he is not familiar with the city.

Figure 2. Specification of Amin scenario



In case weather WS forecasts bad weather, a taxi booking is made by taxi WS upon Amin's approval. In case of pleasant day, Amin uses public transportation. The location of both Amin's hotel and coffee shop are submitted to bus schedule WS, which returns, for example, the bus numbers Amin has to ride. Potential traffic jams force bus schedule WS to regularly interact with traffic WS that monitors the status of the traffic network. This status is fed into bus schedule WS so adjustments to bus numbers and correspondences between buses can occur.

SC-WS PROJECT: DECOUPLING SECURITY CONCERNS IN WEB SERVICES

Related Work

The open and dynamic nature of the environment in which Web services operate poses various challenges and threats to their security. New Web services appear while others disappear without prior notice. Furthermore, messages among component Web services of a composite Web service have to be checked for integrity, confidentiality, and authentication purposes. The need to secure Web services is discussed by Moorthy and Gandhirajan (2005), as the use of Web services continues to increase. This increase is dependent on how much Web services are a serious development alternative to other rival middleware like CORBA and RMI. Indeed, some still consider Web services as distributed objects that react upon request only (Birman, 2004). Enhancing Web services with extra capabilities can happen along three perspectives as reported by Maamar, Benslimane, and Narendra (2006). The first perspective is about deploying Web services that assess the environment before they take part in any composition. The second perspective is about reducing the semantic heterogeneity gap between independent Web services that have all agreed to participate in a composition. Finally, the third perspective is about conciliating contextual information of Web services using ontologies.

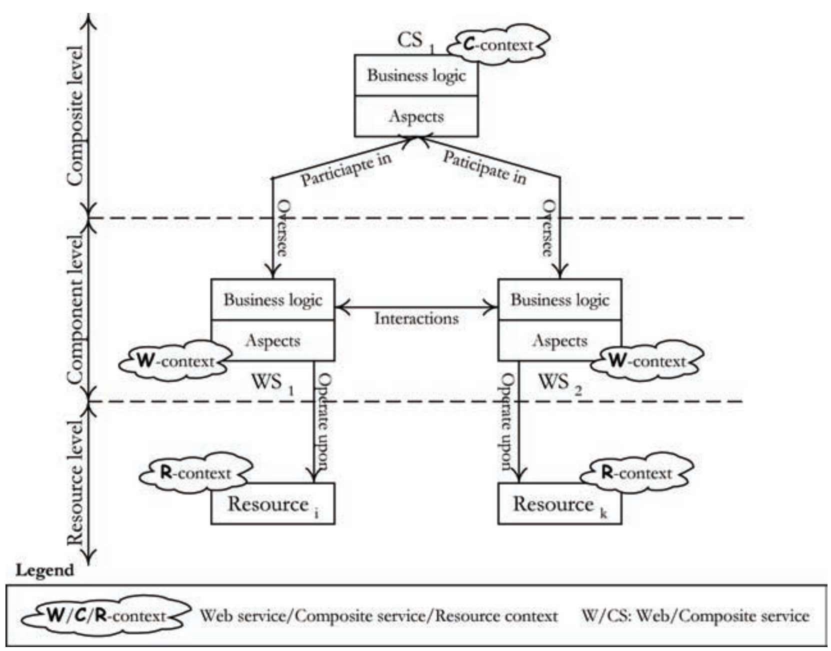
WS-security (2002) is a Microsoft and IBM specification dedicated to Web services security. It is an emerging standard for securing messages among Web services engaged in interactions. For this purpose, WS-security defines how security tokens are contained in SOAP messages. WS-security is extensible to other security models, such as secure sockets layer (SSL), kerberos, and public key infrastructure (PKI). Nowadays, the majority of secure communication measures rely on the transport layer security (TLS). TLS secures interactions by using encryption and makes servers and clients collaborate in order to decide on the authentication process to adopt during data transfer. Unfortunately, TLS does not scale well to complex transactions like those involving Web services (Nakamur, Hada, & Neyma, 2002). Traditional security techniques such as virtual private network (VPN) and SSL cannot secure the large number of requests that Web services expect to receive. The W3C's Web services architecture adopts PKI to secure communications over public networks (W3C, 2005). PKI, however, has a complex infrastructure that negatively affects its deployment cost, processing time, and so forth. Moreover, PKI has the reputation of being quite cumbersome. This could prove to overkill the Web services security to be engaged in intense interactions (Sandhu, 2003). The extensible access control markup language (XACML) is an OASIS standard, which describes both a policy language and an access control decision service interface. A policy is extensible and aims at describing general access control requirements. The request/response style for setting access controls allows forming a query to ask whether or not a given action should be allowed; examples of the queries are permit, deny, indeterminate, or not applicable.

Architecture

Taking into account the context in which Web services operate has been proven to be mandatory when developing Web services. Context-aware Web services result in considering the features of the environment in which the Web services are to be executed (Maamar et al., 2006). These features are multiple and can be related to users (e.g., stationary user, mobile user), their level of expertise (e.g., expert, novice), computing resources (e.g., fixed device, mobile device), time of day (e.g., in the afternoon, in the morning), and so forth. Context is relevant to both functional and nonfunctional properties of Web services.

Figure 3 presents the way aspects are handled in the SC-WS project (Kouadri Mostéfaoui, Maamar, Narendra, & Sattanathan, 2006). Three levels of abstraction exist in this figure: composite, component, and resource. The constituents of each level are related to a particular type of context denoted by C-context, W-context, and R-context, respectively. The rationale of each context

Figure 3. Overview of the SC-WS architecture



type is given by Maamar, Kouadri Mostefaoui, & Mahmoud (2005). The connection between composite, component, and resource levels is implemented with "participate in," "oversee," and "operate upon" relationships, respectively. Some key features of the SC-WS project are as follows: multilevel concern separation using aspects, and contextual tracking of the security requirements of Web services.

The composite level is about context-aware specifications of composite Web services. Each specification is split into two parts: business logic and aspects. The business-logic part reflects the overall objective that the composite Web service has to reach (e.g., hotel booking) using a set of component Web services. The aspect part reflects the cross-cutting concerns that are included in the operation of the composite Web service, and which are orthogonal to this overall objective. The business logic specifies the process by which user functional requirements are met, whereas aspects model user nonfunctional requirements such as security and reliability.

The component level is about context-aware Web services. Similar considerations apply to Web services, which are split into two parts: business logic and aspects. The business-logic part shows the actions that a component Web service has to individually or collectively carry out in order to enable reaching the composite Web service's overall objective. The aspect part shows the nonfunctional requirements that manifest themselves as cross-cutting concerns affecting the actions and interactions of the Web service.

The resource level is about context-aware resources. Resources represent the computing means on which Web services operate. The scheduling of execution requests of Web services is prioritized when enough resources are not available to satisfy them all at once. Moreover, resource allocation to Web services is subject to the context in which the Web services evolve. For instance, the computing requirements of a Web service need to be checked against the computing capabilities of the resources prior to performing resource allocation.

Configuration of Security Aspects

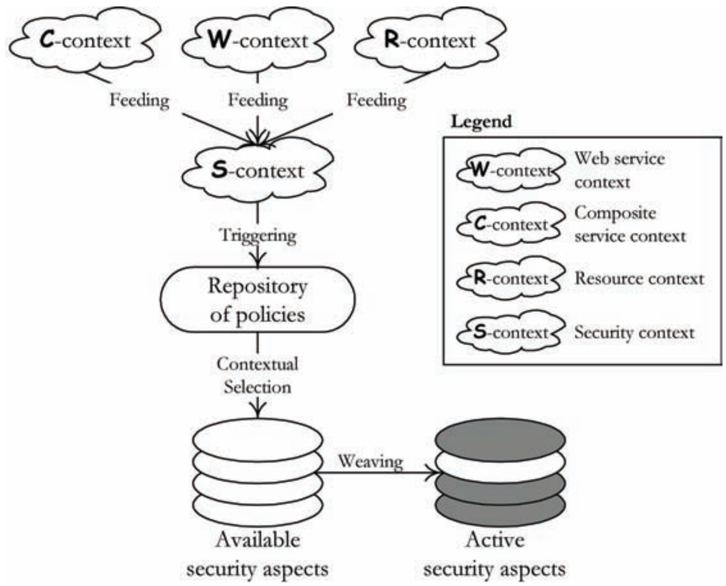
The development of the SC-WS project happened along two dimensions. The first dimension is the need for an adaptive approach that triggers security services (to be implemented as aspects) upon request. For instance, in a specific situation, only authentication aspect is activated, while an extra-logging aspect is activated in another situation. We refer to this dimension in the SC-WS development as composite configuration. It only targets the composite Web services. The second dimension shows the need for a fine tuning of each security aspect associated with composite configuration. For instance, the authentication aspect can be set to accept a timeout of 10 seconds when requesting clients' credentials. We refer to this dimension in the SC-WS development as component configuration. It only targets Web services.

The identification of a configuration that includes both composite and component levels calls for an additional technology to support aspect-oriented programming in modularizing cross-cutting concerns at each level. This technology corresponds to frames. Frames permit achieving this support and are defined as wrappers around code snippets (e.g., source code, HTML code). A frame contains variation points that permit adding, deleting, or adapting functionality in a specific application. This happens using various commands like overriding, extension, substitution, selection, and iteration.

Composite Configuration of Security Aspects

Figure 4 illustrates the operation of the SC-WS project in a configuration of type composite. This operation consists of selecting the security aspects that should be able to protect the whole Web services environment (these aspects are referred

Figure 4. Operations in SC-WS

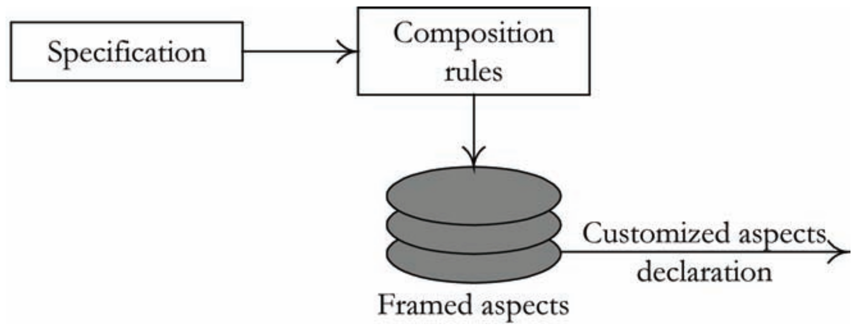


to as active in Figure 4). The selection process combines contextual information and policies. Contextual information offer details on the environment that surrounds each element (e.g., Web service, composite Web service, resource), and policies suggest the appropriate security aspects based on these details.

In addition to W/C/R-contexts of Web services, composite Web services, and resources in the SC-WS project, a new type of context that is just dedicated to security is added (Figure 4). S-context is fed with details obtained out of W/C/R-contexts and gets involved in triggering policies for weaving active security aspects. According to Kouadri Mostéfaoui, a security context is a state of the working environment that requires taking one or more security actions. A security context is formed by a set of information collected from the user's environment and the application environment and that is relevant to the security infrastructure of both the user and the application (Kouadri Mostéfaoui, 2004). In Figure 4, the feeding process is an event-trigger system that gathers contextual information from appropriate sources like contexts of Web services and contexts of resources. Additional sources could be used for feeding the security context, such as physical sensors in the environment or user inputs (Schmidt, Beigl, & Gellersen, 1999). Based on the content validity of each context, policies are triggered. In the following, and relying on a previous work by Sattanathan, Narendra, and Maamar (2005), we overview some arguments that populate each type of context and illustrate the specification of a policy in Ponder. We have selected Ponder due to its expressiveness and ease of use (Damianou, Dulay, Lupu, & Sloman, (2001).

Some arguments in W-context are: signature (establishes the identity of the Web service so that messages to peers are identified), security mechanism (sets the encryption/decryption mechanism needed for authenticating messages received from peers), security status (indicates the status of authenticating the received message in terms of success or failure), and violation (indicates the type of security violation that a message was subject to). Arguments in C-context are similar to arguments of W-context but are interpreted at the composition level. Some arguments in R-context are: signature (establishes the identity of the Web

Figure 5. Framed security aspects generation



service that operates on top of the resource), and violation (indicates the type of security violation that the Web service is involved in). Finally some arguments in S-context are as follows: signature per Web service/composite Web service/resource, security mechanism per Web service/composite Web service/resource, security status per Web service/composite Web service/resource, and security violation per Web service/composite Web service/resource. The main role of S-context is to report on which authentication mechanisms (i.e., username/password pairs, binary certificate, etc.), certificate algorithms, and so forth are supported by all components, whether Web service, composite Web service, or resource, and when they are active.

Policies are information which can be used to modify the behavior of a system (Lupu & Sloman, 1999). The use of policies in the SC-WS project permits managing Web services at a higher level where guidelines for conducting composition of Web services are separated from guidelines for securing Web services. The following is a policy in Ponder that authorizes activating a certain security aspect following the invocation request that a Web service (WS_1) receives from a peer (WS_2). This security aspect depends on the types of authentication and encryption mechanisms featuring WS_2. In this policy, details about these mechanisms are available in the S-context of WS_2 .

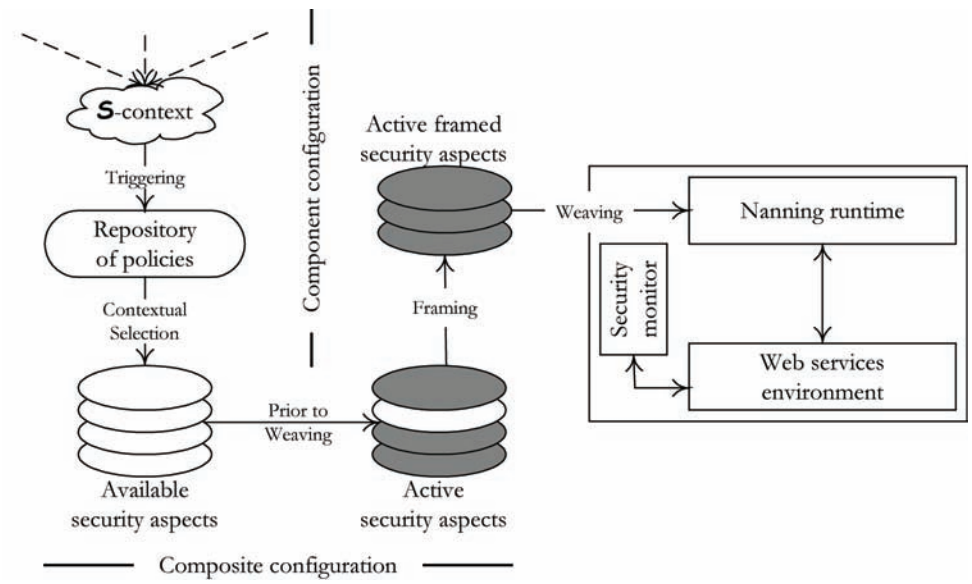
```
inst oblig AuthorizeWS2{
on ServiceRequest(s,t);
when S-context.
authentication(s,"Kerberos",1) and S-con-
text.encryption (s,"DES",1)
subject s = / WS2;
target t = /WS1;
action t.activate(aspecti);
```

Component Configuration of Security Aspects

Figure 5 illustrates the operation of the AoF4WS in a configuration of type component. This configuration aims at supporting the customization of each active-security aspect that was identified in composite configuration and according to the requirements that a user sets. This is achieved using frames. Some examples of user requirements are authentication to happen within 10 seconds and AES-128 is the encryption algorithm. In Figure 5, we also present the way a customized aspect is defined, which is adapted from Greenwood and Blair (2004):

- **Specification:** It is about the developer's security specification to be customized. This consists of setting the values of the metavariables and selecting the different options available in the framed aspect. For example, a specification sets the concrete

Figure 6. Composite and component configuration of security aspects



value of a timeout variable that is contained in a framed authentication aspect code.

- **Composition rules:** They control the way aspects are bound together. For example, an aspect's parameter (e.g., timeout) already defined in the specification can be constrained to a specific interval. The weaving process will then occur upon these constraints.
- **Framed aspect:** It is a parameterized version of the original security aspect that was established in the configuration of type composite. In addition to the generalized aspect code that a famed aspect contains, more elements are added, such as conditional compilation and parameterization.

The reader may wonder about the relationship between policies defined earlier and composition rules. Policies are responsible for selecting security aspects to be activated according to the current context of the whole Web services environment. A composition rule defines how the selected security aspects will be woven in order to secure specific Web services. A composition rule is seen as a policy at the micro level of the aspect. The composition rules apply to the set of aspects once these aspects are selected and customized following their respective specifications.

Putting it all Together

In previous contributions (e.g., Greenwood & Blair, 2004; Loughran & Rashid, 2003), weaving of aspects—for generic applications—is based on a simple schema, that is, on the adaptation of the aspects at the composite level (see above for more details on composite configuration of security aspects). The SC-WS—more specific to Web services—adds an extra step that consists of running an adaptation at the component level by integrating a set of relevant contextual information. Compared to Figure 5, Figure 6 illustrates the operation of the SC-WS after combining composite and component configuration. The new elements in the SC-WS are as follows:

- **Web services environment:** Includes the different component Web services of a composite Web service.

- **Security monitor:** Provides the right set of configured security aspects to the Web services environment.
- **Nanning runtime:** Is based on the Nanning runtime tool (nanning.codehaus.org/) for runtime weaving of security aspects. AspectWerkz, JAC, JAsCo, AOPAlliance, and Prose are examples of other weaving tools.

Figure 6 shows the overall picture of the SC-WS in operation. A transaction in the Web services environment (e.g., a request to use a Web service's functionality) requires from the security monitor to set the needed security aspects (i.e., a request is automatically forwarded to the security monitor before being fulfilled). Component and composite configurations of the AoF4WS engage in collaboration to fulfill this transaction. In the first step, that is, composite configuration, a list of security aspects (e.g., authentication, logging) that need to be included in the security framework is produced. The selected security aspects are then framed in the second step, that is, component configuration. The second step is about how each single aspect will be customized in order to properly respond to the context of use (e.g., type of protocol used by the privacy service). The final set of framed aspects is then concretely woven using Nanning runtime and applied in order to secure the transactions in the Web services environment.

Illustration using Amin Scenario

Refer to Figure 2 for a specification of the Amin scenario. The T-CS_1 puts together an itinerary for Amin. Two of the most significant component Web services of T-CS_1 are taxi booking Web service (TB-WS_1) and bus schedule Web service (BS-WS_2). In the C-context of T-CS_1, Blowfish algorithm is set as part of the security mechanism. In the W-contexts of TB-WS_1 and BS-WS_2, DES and AES algorithms are set, respectively. TB-WS_1 uses a resource, which is an online database through which up-to-date information on available taxis are provided. In the R-context of this database, authentication information is presented so TB-WS_1 gets to access this database.

Based on the W/C/R-contexts discussed earlier, the S-context arguments of T-CS_1 are instantiated. Some of them are: DES algorithm for BS-WS_2, Blowfish algorithm for T-CS_1, security status for TB-WS_1 accessing the database resource ("access granted"), and security violation (if any has occurred; in our case, so far, no). The S-context information is then used to populate the policy repository with the appropriate policies. A sample of policy is to authorize the invocation of TB-WS_1 upon request of T-CS_1, assuming that the security conditions are met.

```
inst oblig AuthorizeTaxiWebService{
on ServiceRequest(s,t);
when S-context.
authentication(s,"Blowfish",1) and S-con-
text.encryption(t,"DES",1)
subject s = /T_CS1;
target t= /TB_WS1;
action s.invoke(t);
```

In other words, when trip Web service authenticates itself to taxi service, this latter is supposed to accept the invocation request from trip service. A similar bus schedule request can also happen, as shown below.

```
inst oblig AuthorizeBusScheduleWebSer-
vice{
on ServiceRequest(s,t);
when S-context.
authentication(s,"Blowfish",1) and S-con-
text.encryption(t,"AES",1)
subject s = /T_CS1;
target t= /BS_WS2;
action s.invoke(t);
```

Based on the policies defined above, the list of appropriate security aspects is generated (i.e., DES aspect, Blowfish aspect, and AES aspect). Since the above policies are needed for Amin, the appropriate authentication code is weaved into the respective Web services (TB-WS_1 and BS-WS_2) in order to ensure that the necessary security checks are carried out during the composition of these Web services. The actual weaving itself is carried out via frame technology, as follows. Prior to the weaving process, each framed security aspect—identified by the list generated earlier—is customized according to the values set in the specification, as illustrated in Figure 6 (component configuration of security aspects). Afterwards the framed versions of these framed security aspects are woven using the Nanning runtime.

Summary

In the SC-WS project, we argue for an adaptive security strategy in Web services environments using framed aspects, which are the combination of frames and aspect-oriented programming. Frames enhance aspect-oriented programming by separating the specification of a security aspect from the aspect code itself. This approach allows for a fine-grained variation of security aspects according to the context of use.

SH-WS PROJECT: SELF-HEALING WEB SERVICES

The development of self-healing Web services means enhancing them with self-healing properties, such as how to deal with a timeout response from a critical request and how to resume operation after a major crash. One of the recommendations we put forward while designing and coding self-healing functionalities is to keep them separate from the design and code implementing the business logic of a Web service. Concern separation permits avoiding cross-cutting issues and emphasizes the noninvasive requirement that needs to be integrated into the development strategy of self-healing Web services. The inappropriate handling of this requirement leads into a code that is scattered all over the Web service and, thus, becomes difficult to localize and maintain. This maintenance exercise is extensive, expensive, and error-prone. In this chapter we suggest using aspect-oriented programming to design and develop self-healing Web services. The use of self-healing permits develops Web services that are more agile and robust, responsive to (unpredictable) changes in the environment, thereby resulting in reduced downtime, capable of self-diagnosis, and proactively seeking to avoid "unsafe" configurations. Multiple challenges face the development of self-healing Web services, including how to trigger the self-healing process, how to model and track Web services engaged in self-healing operations, how to adjust control and data flow among these Web services, and how to automatically modify this flow with little disruption.

Related Work

Baresi, Ghezzi, and Guinea (2004) select the selection stage of Web services in order to illustrate the importance of self-healing mechanisms. Shutting-down a system because of a Web service failure is no longer acceptable, whether in critical-systems or not. The execution environment should be able to identify new Web services and even to reorganize the process to find a solution that uses what is available, if a perfect match does not exist. Baresi et al.'s proposal revolves around special-purpose probes, that is, monitors, to allow the execution environment to detect anomalous conditions such as a nonresponding Web service. Their proposal is built-upon two techniques: defensive process design and service runtime monitoring. The approach they follow is mainly based on assertions of prepost and invariant conditions; these facilities are inspired from some programming languages such as Eiffel. The resulting proposal suggests the use of such facilities in order to implement recovery actions at the code level.

Ardissono, Console, Goy, Petrone, Picardi, Segnan, et al. (2005) propose a framework for adding diagnostic capabilities to Web services, using a model-based perspective. The objective is to develop self-healing Web services. The framework associates each Web service with a local diagnoser that relates hypotheses about incorrect outputs of this Web service to misbehavior of the Web service itself, or to incorrect inputs from other peers in composition scenarios. Besides the local diagnoser, a global diagnoser is deployed at the composite level. It coordinates the local diagnosers, exchanging messages with them and sometimes computes diagnoses without relying on the feedback of these local diagnosers. Pencole, Cordier, and Grastien (2005) model Web service workflows as discrete-event systems using a model-based reasoning approach. Such a model is the first step towards tracing the evolution of the workflow and diagnosing faults at run-time. The result is a tool for monitoring and diagnosing Web services.

Although the research outcomes in the above projects are promising, two major concerns still remain to be handled: How do we develop self-healing Web services without altering their underlying code? And how do we interleave monitoring, diagnosis, and adaptation as part of the self-healing process without interrupting the execution of the unaffected Web services? Modularizing both concerns using software engineering techniques like object-oriented is hard to achieve. This is where aspect-oriented programming comes into play, as will be described below.

Monitoring Model

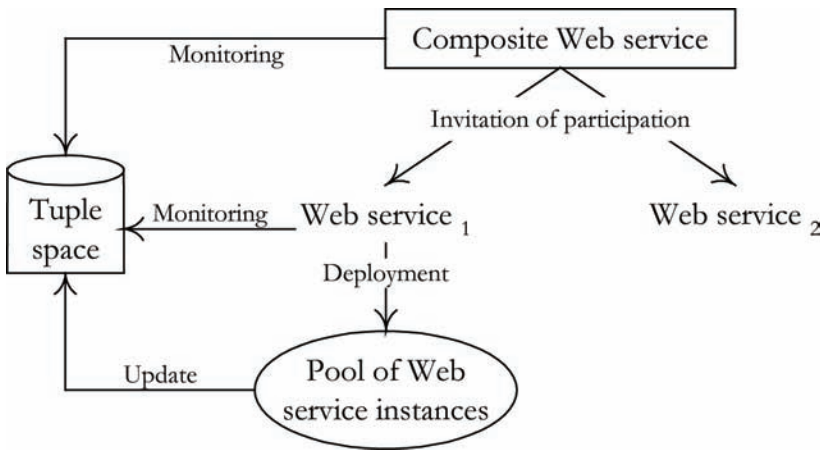
As part of the self-healing process, the ability of Web services to self-coordinate is important, that is, being able to monitor the progress of their execution without the overhead of a centralized coordinator. Of course, the composite Web service can always implement a centralized monitoring. However, this turns out to be a bottleneck to the architecture and imposes an unacceptable performance overhead. A decentralized monitoring model that could also permit centralized monitoring, if needed, is deemed appropriate.

Our monitoring model complies with the distributed tuplespace-based approach described by Maamar, Benslimane, Ghedira, & Mahmoud (2005) and depicted with Figure 7. The lifecycle of the monitoring model highlights the following elements: composite Web service issuing invitations of participation to component Web services, component Web services deploying Web service instances upon invitation acceptance, and monitoring being conducted at composite and component levels.

The operation of the monitoring model of Figure 7 assumes that a Web service instance is self-aware. The Web service instance has access to the description of the specification of the composite Web service. This permits to the Web service instance to know and assess the constraints under which it operates in terms of execution time, execution chronology, and acceptable data, just to cite a few. For monitoring purposes at the Web service level, a Web service instance is supposed to post all its activities on the tuplespace that is connected to its Web service. The composite Web service has all access to the tuplespace subject to some access rights verification. It is interesting to note that any failure or delay is automatically reported in the tuplespace, which permits immediate solutions to be deployed. To keep the chapter self-contained, the way a tuplespace operates is excluded. Readers are referred to Maamar, Benslimane, Ghedira, Mahmoud, and Yahyaoui's (2005) work.

Applied in the Amin scenario, the monitoring step is concretized as the following. When a Web service detects an error within its execution, it raises an exception by updating its respective tuplespace. Upon receipt of the exception, the tuplespace reports it to the rest of participating Web services that have expressed interests in this kind of exception. If bus schedule WS detects an error,

Figure 7. Monitoring model for self-healing Web services



this information will be recorded and propagated as an exception to traffic WS and weather WS.

Diagnosis Model

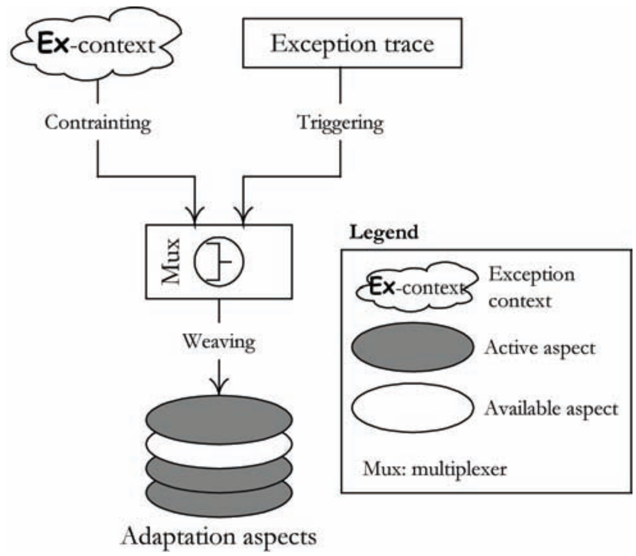
As part of the self-healing process, diagnosis continues monitoring. Diagnosis is the process of discovering the cause and nature of a fault (Ardissono et al., 2005). When a Web service instance fails, the cause can either lie within this Web service instance, or within a peer or resource that the Web service instance in question depends on. In the Amin scenario, if the bus schedule WS signals a failure, two options exist: either the bus schedule WS itself is not able to send the schedule due to an internal error, or the bus schedule WS has not received the right input from the location WS so it can send an appropriate bus schedule to Amin.

Two types of dependency exist between any pair of Web service instances WS-I_i and WS-I_j, where WS-I_j succeeds WS-I_i in the composite Web service specification: control flow and data flow. In a data flow dependency, WS-I_j requires data from WS-I_i in order to execute successfully. For example, the bus schedule WS requires the location data from the location WS so a bus schedule is provided. WS-I_i can create, modify, forward, or consume the data.

- **Data creation:** WS-I_i is the source of the data. WS-I_i is therefore to be diagnosed as the cause of failure since the data are being erroneous.
- **Data modification:** A different Web service instance is the source of the data. This requires checking whether the modification operation implemented by WS-I_i produces the expected result. If so, then the data need to be traced back to their source, in order to determine where the error occurred.
- **Data forwarding without modification:** Another Web service instance WS-I_k before WS-I_i in the composite Web service specification, is the possible failure cause. Therefore, the data should be traced back to their source until the failure-causing Web service instance is identified.
- **Data consumption:** WS-I_i is the final destination of the data that will become an input for the creation/modification of another data variable.

In a control flow dependency, the output of WS-I_i should satisfy certain preconditions that are needed for WS-I_j to execute successfully. In the Amin scenario, the positive confirmation of bad weather is the precondition needed for taxi WS to execute. The precondition value generated by

Figure 8. Aspects during adaptation



WS-I_i is based on the data created/modified by it. In case of any error in this data, for example, wrong confirmation of bad weather, the source would be WS-I_i itself.

The distributed tuplespace supports diagnosis by tracing the execution results in the reverse direction in which the execution occurred, while checking where the control and/or data flow dependencies have been violated. This will stop once all Web services where the data variables have been created have been covered by the tracing process. As described above, the execution results are stored/maintained in the tuplespace, and hence they can be retrieved from there for verification. In the Amin scenario, a partial list of data flow sequences could be: location WS → bus schedule WS, traffic WS → bus schedule WS, and weather WS → bus schedule WS. The consumption of location, traffic, and weather data results in creating the bus schedule for Amin. Any error in bus schedule WS means that the data flow sequences will be traversed in the reverse direction up to the sources of these data and verified to determine where the error occurred. First, the bus schedule calculation is checked; if that is correct, the location/weather/traffic calculations are checked to determine which of these data variables are erroneous. Since these three data variables are created afresh by their respective Web services, the checking will stop there.

Adaptation Model

Having detailed how the monitoring and diagnosis models operate, we discuss hereafter the adaptation as the last step of the self-healing process. The adaptation occurs using what we call exception trace, exception context, and multiplexing aspects (Figure 8). The exception trace relies on the data flow that is stored in the tuplespace and obtained by a reverse analysis of the content of the tuplespace. The exception context is the state of the Web services when the failure occurred. Multiplexing aspects trigger the adaptation actions based on the exception trace and the context.

Web services failures range from internal errors to some other external factors like network congestion. Handling exceptions is then highly dependant on the source and type of failure. To keep track of a failure source and type, an exception trace is generated out of the tuplespace. An effective adaptation strategy should be supported by an appropriate exception handling at the code level. This is dependant on the features of the programming language. In Java, java.lang.Exception and java.lang.RuntimeException packages are dedicated classes for exception handling.

The adaptation strategy we discuss in this chapter relies on the exception and the current context as well. Recovering from a timeout failure when the requested Web service is overloaded may be handled by a retry process. The same exception due to the Web service constant unavailability requires a dynamic binding to another Web service. As a result, Web services should be continually aware of the current context in order to adopt the right strategy for exception handling. The following suggests some examples of contextual elements constraining the adaptation strategy:

- **Exception time:** When the exception was observed.
- **Exception location:** At which point of the Web services composition the exception occurred.
- **Previous exceptions:** Number of times the exceptions were reported at the same point of the Web services composition.

In Figure 8, the multiplexer supports multiplexing aspects. Multiple aspects mean different strategies to adopt for handling different exceptions. Multiplexing is a well-known technique in the field of electrical engineering (Wikipedia, 2005). This technique supports mixing the inputs from multiple sources into one connection. In the self-healing process, multiplexing is used to simultaneously combine exception traces and current exception context so the right self-healing strategy is derived. In order to implement the concept of multiplexing aspects, we follow a policy-based approach. The primary use of policies is to perform actions according to the occurring events and detected changes. Our main motivation in adopting a policy-based approach is to support the adaptation of self-healing actions by taking advantage of the up-to-date information that context caters over occurred exceptions. We use the well-known Ponder language for representing the policies, due to its expressiveness and ease of use (Damianou et al., 2001). We use obligation policies as a class of Ponder policies, which are event-driven condition-action rules for actions that must be performed on specified objects of the target domain. For weather WS, the policy for activating the aspect relating to weather confirmation could be as follows:

```
inst oblig SelfHealing{
on Exception(t);
when context-elt1 = “4 PM” and context-
elt2 = “Weather Confirmation” and con-
text-el3 = “3”
subject t = /WeatherService;
action t.activate(aspecti);}
```

The activation of aspect$_i$ would result in changing weather WS code, so the correct weather information for Amin is recomputed. The contextual elements refer to those listed in the previous table: exception occurred at 4 PM, in weather confirmation module of weather WS, and this is the third time this error occurs. In case the error was in a different module, say, weather information gathered from a satellite imagery, a different aspect other than aspect$_i$ would need to be activated for self-healing purposes.

Summary

As a nonfunctional and must-have property, fault tolerance of Web services has been extensively investigated in the last few years. The focus has been more on error detection and recovery in composite Web services execution flow. In the current chapter, however, we have proposed a reusable framework that achieves self-healing of Web services and allows for consideration of recovery actions as structured units. Implemented as software aspects, the self-healing actions are triggered based on the type of the fault in the composite Web services flow and on the context of use. Our approach aims at separating the fault-tolerance mechanisms from the core code of the Web services in order to allow for a structured and easy maintenance of Web services. We have also positioned our self-healing technique within our previously introduced AoF4WS framework that uses aspects for decoupling various concerns in Web services. Future work will involve developing an implementation to demonstrate our technique.

CONCLUSION

The AoF4WS research initiative is devoted to highlight the added-value of aspect-oriented programming in implementing nonfunctional requirements of Web services. It is also intended to identify the core technologies that realize these requirements. In this chapter, we introduced the AoF4WS initiative that so far encompasses two research projects, namely SC-WS and SH-WS. Both projects use the well-known aspect-oriented programming technique to decouple cross-cutting concerns in Web services. Such a decoupling promotes a clear separation between the operational and nonfunctional aspects of Web service executions, thereby resulting in easier adaptability and maintainability. As a future work, we are looking at extending our framework by tackling other nonfunctional requirements, such as performance. In addition, we will be investigating the issues involved in implementing several aspects simultaneously, which could result in semantic interference between the aspects (Durr, Staijen, Bergmans, & Aksit, 2005).

REFERENCES

W3C. (2005). *Working group.* Retrieved June 1, 2008, from http://www.w3.org/

Ardissono, L., Console, L., Goy, A., Petrone, G., Picardi, C., Segnan, M., et al. (2005). Towards self-diagnosing Web services. In *Proceedings of the IFIP/IEEE International Workshop on Self-Managed Systems & Services (SelfMan'2005) held in conjunction with The 9th IFIP/IEEE International Symposium on Integrated Network Management (IM'2005),* Nice, France.

Baresi, L., Ghezzi, C., & Guinea, S. (2004). Towards self-healing service compositions. In *Proceedings of the First Conference on the Principles of Software Engineering (PRISE'2004),* Buenos Aires, Argentina.

Birman, K. (2004, December). Like it or not, Web services are distributed objects. *Communications of the ACM*, *47*(12). doi:10.1145/1035134.1035157

Charfi, A., & Mezini, M. (2004). Hybrid Web service composition: Business processes meets business rules. In *Proceedings of the 2nd International Conference on Service Oriented Computing (ICSOC'2004)*, New York.

Cibrán, M., & Verheecke, B. (2003). Modularizing Web services management with AOP. In *Proceedings of the 1st European Workshop on Object Orientation and Web Services (ECOOP'2003) held in conjunction with The 17th European Conference on Object-Oriented Programming (ECOOP'2003)*, Darmstadt, Germany.

Cottenier, T., & Elrad, T. (2004). Validation of aspect-oriented adaptations to components. In *Proceedings of the 9th International Workshop on Component-Oriented Programming (WCOP'2004) held in conjunction with The 18th European Conference on Object-Oriented Programming (ECOOP'2004)*, Oslo, Norway.

Damianou, N., Dulay, N., Lupu, E., & Sloman, M. (2001). The ponder specification language. In *Proceedings of the Workshop on Policies for Distributed Systems and Networks (Policy'2001)*, Bristol, UK.

Durr, P., Staijen, T., Bergmans, L., & Aksit, M. (2005). Reasoning about semantic conflicts between aspects. In *Proceedings of the 2nd European Interactive Workshop on Aspects in Software (EIWAS'2005)*, Brussels, Belgium.

EL-Manzalawy. Y. (2005). *Aspect oriented programming*. Retrieved August 2005, fromhttp://www.developer.com/design/article.php/3308941

Greenwood, P., & Blair, L. (2004). *Dynamic framed aspects for policy-driven auto-adaptive systems*. Retrieved August 2004, from http://www.comp.lancs.ac.uk/computing/aose/papers/dynFr_daw04.pdf

Kouadri Mostéfaoui, G. (2004). *Towards a conceptual and software framework for integrating context-based security in pervasive environments*. Unpublished doctoral thesis (No. 1463), University of Fribourg and Paris 6 University.

Kouadri Mostéfaoui, G., Maamar, Z., Narendra, N. C., & Sattanathan, S. (2006). Decoupling security concerns in Web services using aspects. In *Proceedings of the 3rd International Conference on Information Technology: New Generations (ITNG'2006)*, Las Vegas.

Loughran, N., & Rashid, A. (2003). Supporting evolution in software using frame technology and aspect orientation. In *Proceedings of the Workshop on Software Variability Management*, Groningen, The Netherlands.

Lupu, E., & Sloman, M. (1999, November/December). Conflicts in policy-based distributed systems management. *IEEE Transactions on Software Engineering*, *25*(6). doi:10.1109/32.824414

Ma, K. (2005, March/April). Web services: What's real and what's not? *IT Professional*, *7*(2). doi:10.1109/MITP.2005.47

Maamar, Z., Benatallah, B., & Mansoor, W. (2003). Service chart diagrams: Description & application. In *Proceedings of the Alternate Tracks of The 12th International World Wide Web Conference (WWW'2003)*, Budapest, Hungary.

Maamar, Z., Benslimane, D., Ghedira, C., Mahmoud, Q. H., & Yahyaoui, H. (2005). Tuple spaces for self-coordination of Web services. In *Proceedings of the 20th ACM Symposium on Applied Computing (SAC'2005)*, Santa Fe, NM.

Maamar, Z., Benslimane, D., & Narendra, N. C. (2006). What can context do for Web services? *Communications of the ACM.*

Maamar, Z., Kouadri Mostéfaoui, S., & Mahmoud, Q. (2005, July-September). On personalizing Web services using context. *International Journal of E-Business Research, Special Issue on E-Services*, *1*(3). IGI Global, Inc.

Moorthy, K. R., & Gandhirajan, A. (2005). *The foundations of Web services security.* Retrieved August, from http://www.developer.com/services/article.php/3496326

Nakamur, Y., Hada, S., & Neyma, R. (2002). Towards the integration of Web services security on enterprise environments. In *Proceedings of the Workshop on Web Services Engineering 2002 (WebSE'2002) held in conjunction with The IEEE/IPSJ Symposium on Applications and the Internet (SAINT'2002)*, Nara, Japan

Ortiz, G., Hernández, J., & Clemente, P. J. (2004a). Decoupling non-functional properties in Web services: An aspect-oriented approach. In *Proceedings of the 2nd European Workshop on Web Services and Object Orientation (EOOWS'2004) held in conjunction with the 18th European Conference on Object-Oriented Programming (ECOOP'2004)*, Norway.

Ortiz, G., Hernández, J., & Clemente, P. J. (2004b). Web services orchestration and interaction patterns: An aspect-oriented approach. In *Proceedings of the 2nd International Conference on Service Oriented Computing (ICSOC'2004)*, New York.

Sandhu, R. (2003, January/February). Good-enough security: Toward a pragmatic business-driven discipline. *IEEE Internet Computing, 7*(1). doi:10.1109/MIC.2003.1167341

Sattanathan, S., Narendra, N. C., & Maamar, Z. (2005). Towards context-based tracking of Web services security. In *Proceedings of The 7th International Conference on Information Integration and Web Based Applications & Services (iiWAS'2005)*, Kuala Lumpur, Malaysia.

Schmidt, A., Beigl, M., & Gellersen, H. W. (1999, December). There is more to context than location. *Computers & Graphics Journal*, *23*(6).

Web services security. (2002). *Version 1.0*. Retrieved August 2005, from http://www.verisign.com/wss/wss.pdf

Wikipedia. (2005). *The free encyclopedia*. Retrieved June 1, 2008, from http://en.wikipedia.org/wiki/Multiplexing/

Yan, Y., Pencole, Y., Cordier, M. O., & Grastien, A. (2005). Monitoring Web service networks in a model-based approach. In *Proceedings of the 3rd European Conference on Web Services (ECOWS'05),* Vaxjo, Sweden.

This work was previously published in Managing Web Service Quality: Measuring Outcomes and Effectiveness, edited by K. Khan, pp. 245-263, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 7.17

A Static Web Immune System and its Robustness Analysis

Tao Gong
Donghua University, China & Central South University, China

ABSTRACT

Static Web immune system is an important applicatiion of artificial immune system, and it is also a good platform to develop new immune computing techniques. On the Static Web system, a normal model is proposed with the space property and the time property of each component, in order to identify the normal state of the system that the artificial immune system protects. Based on the normal model, the Static Web immune sytsem is modelled with three tiers, that is the innate immune tier, the adaptive immune tier and the parallel immune tier. All the three tiers are inspired from the natural immune system. On the tri-tier immune model, the self detection mechanism is proposed and programmed based on the normal model, and the non-self detection is based on the self detection. Besides, the recognition of known non-selfs and unknown non-selfs are designed and analyzed. It is showed that the Static Web immune system is effective and useful for both theory and applications.

DOI: 10.4018/978-1-60566-310-4.ch011

A.1 INTRODUCTION

Human immune system is very important for human health, because it is able to detect, recognize, memorize and eliminate foreign viruses and inner faults, which are sometimes unknown and even quite complex (Perelson, Hightower & Forrest, 1996; Fauci, 2003; Chao, Davenport & Forrest, et al, 2004). Inspired from nature, artificial immune system is very important for computer world, because it is used to detect, recognize, learn, memorize and eliminate special objects, which are possibly unknown and even quite complex, such as computer viruses, faults and so on (De Castro & Timmis, 2002; Jerne, 1974; De Castro & Von Zuben, 2002). However, due to incomplete theories of immunology, one of bottlenecks for detecting the unknown non-selfs prevents the artificial immune system from developing. First, traditional detection approaches against viruses and faults are based on matching the features of the viruses and faults, and the features of unknown viruses and unknown faults are possibly unknown, thus 100% detection is impossible in theory (Balachandran, 2005; Gonzalez & Dasgupta, 2003). Second, the faulty mechanism for detecting viruses and faults causes lower possibility for recognizing the viruses and faults, and affects ability and efficiency for repairing the damaged computer system. To overcome the bottleneck of research on the artificial immune system and improve research on the basis of the anti-worm application and software fault diagnosis, a normal model of the static web immune system is proposed and built with the space-time properties of the components, and the normal model is used to represent the selfs.

A.2 BACKGROUND

Web system is popular on the Internet now and useful for many web users, and web security has become a serious problem due to viruses, worms and faults (Balthrop, Forrest & Newman, et al., 2004; Orman, 2003). To solve the security problem, some detecting techniques are used to recognize the non-selfs such as viruses and faults by matching the features of the non-selfs, but the traditional techniques have a difficult bottleneck in detecting unknown non-selfs especially such as brand-new viruses. To overcome the bottleneck, a new strategy for detecting the unknown non-selfs has been proposed with the normal model of the system that the artificial immune system protects. Current work has been done on the static web system and in fact many static web systems are useful and popular on the Internet, such as the webpage system for many companies and universities.

A.2.1 Space Property of Component

Suppose a static web system S is comprised of m web directories and n files in the cyberspace, and the system can be represented with the set

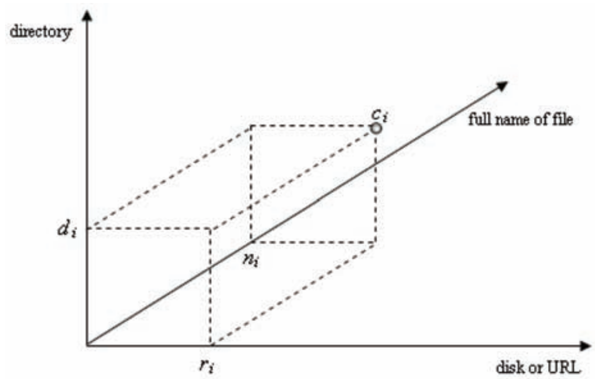
$$\{p_{ij}, d_k | \sum_{i=1}^{m} n_i = n; p_{ij} \in S; j = 1, \cdots, n; k = 1, \cdots, m\}$$

Here, p_{ij} denotes the jth file in the ith directory of the system *S*, dk denotes the kth directory in the system *S*, and ni denotes the sum of all files in the ith directory of the system *S*.

The components of the static web system are software parts, and the software is used to simulate the physical world in the cyberspace. In the physical world, every object has unique 3-dimension space coordinates and 1-dimension time coordinate, so that the state of the object is uniquely identified by its space-time coordinates (Einstein, 1920). Alike in the cyberspace, every software part has unique location for storing the space property because the storage of the software is based on the hardware in the physical world. The absolute pathname pi is used to represent the location information for storing the file and/or the directory, and the pathname consists of the name ri of the disk or the URL, the name di of the directory and the full name ni of the file ci, shown in Figure 1. The full name of the file includes the file-name of the file and the suffix name of the file, and the suffix name of the file is one of features that are useful for classifying the files.

According to the basic rules of the operating systems for managing the files, the absolute pathname of the file c_i uniquely identifies the location

Figure 1. 3-dimension information of the absolute pathname for files. ©2008 Tao Gong. Used with permission.



of the file in the computer. One absolute pathname belongs to only one file at a certain time, and at that time the file has only one absolute pathname.

A.2.2 Time Property of Component

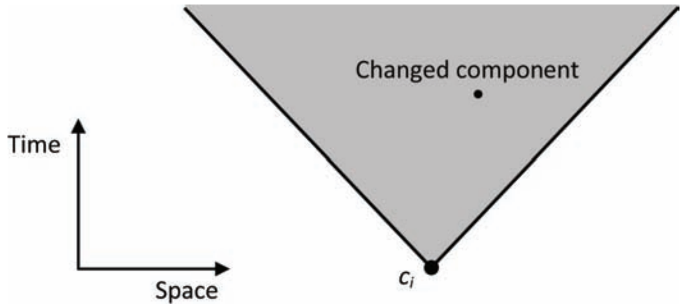
Time property of every object in both the physical world and the software world only has one dimension, and each time variable has four parts: date part, hour part, minute part and second part. However, the time property of natural object in the physical world is not changeable, but the time properties of files in the current operating system are changeable. The operating system is a kind of complex dynamic system in which the normal model is difficult to build. For simplification, the normal model is built for the static web system and the artificial immune system is designed to protect the static web system, because the normal model can be built on the condition that the time property in the static system is unchangeable to keep the normal state.

In general, according to the representing method of time, the data of time includes the data of the date, the data of hour, the data of minute and the data of second. Moreover, if the accuracy for representing the data of time is higher, the data of time may include the data of microsecond and so on.

When a component is changed, the time point of the cyberspace is certainly changed and the change on the component can be external or internal. The external change of the component can be shown with the space property of the component, but the internal change of the component must be detected by the operating system and represented with the time property. In some current operating systems, the change of the component such as file or directory is detected with the function for representing the time that the component has been changed for the last time. The approach may be imperfect, but it is feasible in the static web system on the condition that the operating system is robust. The external change of the component may occur on the disk name of the component, the directory name of the component or the file name of the component, and the future change of the component c_i is shown in Figure 2 (Hawking & Penrose, 1996).

On the other hand, the internal change of the component is affected by many factors such as the structure of the component, the length of the component, the special unit of the component and so on. For example, from the time point t_1 to t_2, the component c_i is changed with its structure into a new component c_j, shown with bigger scale in Figure 3, and the functions of the component will be also changed.

Figure 2. External change of the component c_i. ©2008 Tao Gong. Used with permission.



A.2.3 State of Static Web System

The state of the static web system is either normal or abnormal, and the normal state of the static web system is static. When the static web system is set up for usage, the initial state of the system is normal and the normal state of the static web system can not be transformed into another different normal state. In this chapter, suppose the operating system is always normal, the artificial immune system is designed to protect the static web system so that the static web system can keep normal against worms and software faults.

To keep the normal state of the system, the state of the static web system must be defined as normal or abnormal. The normal state of the static web system is the state that the current system has same structure and components as the initial system. The abnormal state of the static web system is the state that the current system has different structure and components from the initial system.

To describe the state of the static web system, a normal function $N(x)$ is defined as such.

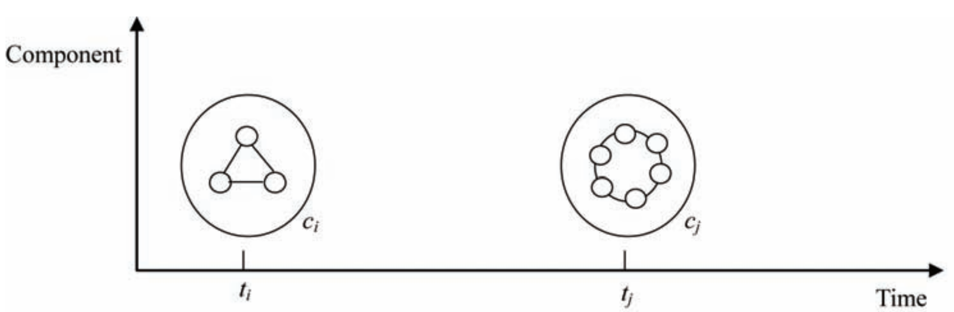
$$N(x) = \begin{cases} 1 & , x \text{ is normal}, \\ 0 & , x \text{ is abnormal}. \end{cases} \tag{1}$$

Here, x is an object in the static web system or the system itself.

Considering the composition of the static web system, the state of the static web system is uniquely identified by the states of the components for the system, and the components of the static system include files and directories. Suppose the state of the component c_i is represented as $N(c_i)$, the state of the system S can also be represented as such.

$$N(S) = \prod_{i=1}^{m+n} N(c_i), \quad c_i \in S. \tag{2}$$

Figure 3. Internal change of the component c_i. ©2008 Tao Gong. Used with permission.



A.2.4 Normal Model of Normal Static Web System

When and only when all the components of the system are normal, the system is normal. If one component among them is abnormal, the system is abnormal. Therefore, if and only if every component of the static normal web system is identified uniquely, the normal state of the static web system can be identified uniquely. The following theorem proves that the normal state of the static web system can be detected with the space-time properties of the components in the system.

Theorem A.1 Suppose a static web system *S* is comprised of *m*+*n* components, each of which has its unique absolute pathname (space property) and unique last revision time (time property). Let the absolute pathname of a component c_i be p_i and the last revision time of the component c_i be t_i. On the condition that the space-time properties of the static web system are correct, the absolute pathnames and last revision time of all the normal components in the normal static web system uniquely identify the normal state of the system.

[Proof] In the static web system there are n+m components of web pages and directories, which uniquely identify the system. Therefore, the normal state of every component in the static web system should be identified with its space property and time property before the whole web system is identified. For a certain absolute pathname p_i, the corresponding component c_i is unique. Moreover, if the last revision time t_i of the component c_i is certain, the current state of the component c_i is also unique. The revision time t_i is recorded when the static web system is normal, so that the component c_i is also normal at that time. Thus, the absolute pathname p_i and the last revision time t_i identify the normal state of the unique component c_i in both the space and time dimensions. For a certain component c_i of the web system *S*, the absolute pathname p_i of the component is unique, because this is kept by the rules for managing files and directories with the operating system. Besides, the revision time of the component c_i is also unique, because the last revision time of every thing should be unique. All in all, the absolute pathname and last revision time of every normal component in the normal static web system identify the normal state of the component. Based on the unique identification between the static web system and its components, the normal state of the static web system is uniquely identified with the absolute pathnames and last revision time of its all normal components. Theorem 1 is correct. ■

The last revision time is a representing parameter for the operating system and the parameter is different in the physical world. When a copy of a component is made in the cyberspace, the copy has same revision time (time property) as the component, so that the copy is called same to the component. In fact, the physical storage of the copy is different from that of the component. In the physical world, no two objects are completely same and the concept of sameness is local and partial. Therefore, the time property of the component shows the partial computing feature of the operating system in the cyberspace.

According to Theorem 1, the normal state of the static web system has three examples. The first example is the initial static web system, and the conclusion is self-evident. The second is the unchanged system after a period, and the space properties and the time properties of all the components in the system are unchanged. Therefore, the state of the current system is same as the initial system, so that the current system is also normal. The third example is the repaired system that was damaged and ever abnormal but now has been repaired by the artificial immune system. When the damaged components are being repaired, the backup components are copied to replace the damaged ones and the backup components have same last revision time as the components that were damaged. Thus, all the components of the repaired system have same space properties and time properties as those of the initial system, and the repaired system has the same state as the initial system.

The normal model of the static web system is a new attempt to build the normal model of the selfs and increase the probability for detecting the non-selfs by detecting the selfs. This work can establish a basis for investigating the approaches to build the normal model of the dynamic system and design more intelligent and effective immune algorithms for real complex systems. To update the normal model for the dynamic systems, the normal models must be adjusted according to some rules and protected by some security techniques.

A.3 TRI-TIER IMMUNE MODEL OF STATIC WEB SYSTEM

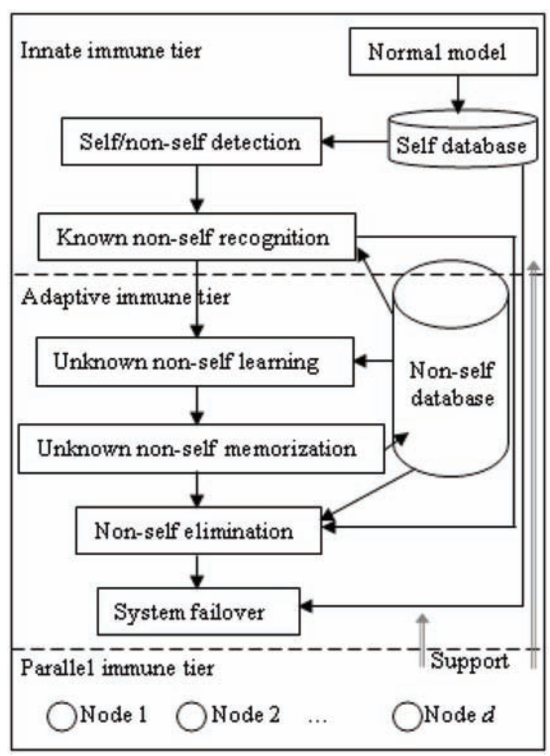
In biological immune systems, the innate immunity and the adaptive immunity are two main types of immunity, which are provided by many immune cells such B-cells and T-cells. The innate immunity is responsible for discriminating the selfs from the non-selfs and recognizing the known non-selfs. The adaptive immunity is responsible for learning unknown non-selfs and memorizing the new known non-selfs that were ever unknown but now known after learning. Based on the mapping model of natural computation, the two types of immunity are mapped to the first tier and the second tier of the immune computing model for the static web immune system. Considering another important trait of processing information in a parallel and distributed way in the biological immune systems, the parallel immune cells is an infrastructure for the immune system. Alike, the parallel computer can provide a powerful computing infrastructure for some complex artificial immune systems, and the parallel computer is used as the third tier of the immune computing tier for the static web immune system.

A.3.1 Tri-Tier Immune Structure

The tri-tier immune computing model of the static web immune system is based on the normal model of the static web system, and the normal model is built with both the space property and the time property of each component in the system. The tri-tier immune computing model is inspired from some natural immune theories, and includes inherent immune tier, adaptive immune tier and parallel immune tier, shown in Figure 4. The parallel immune tier is built on the parallel computing theorems, and used to increase efficiency (Gong & Cai, 2003).

In Figure 4, when the foreign object enters the static web system, the random detectors begin to detect whether the foreign object is self or non-self, by matching the features of the foreign object with

Figure 4. Tri-tier immune computing model of static web immune system. ©2008 Tao Gong. Used with permission.



the self information in the self database. The self is the part of the static web system, such as the webpage files of the static web system. And the non-self is not any part of the static web system or compatible with the static web system. The self information is used to define the features of the self and represent the normal state of the static web system. And all self information is stored in the self database. During the immune process of the artificial immune system, the percent of the selfs is maximized and the amount of the non-selfs in the static web system is minimized. When a foreign object is determined as a non-self, pattern recognition of the non-self is started in two ways. One is the way of matching features, and the other is the way of matching rules. The former is done through querying records in the database and matching the feature information of the detected non-self with the record information in the non-self database, where the entire known non-selfs store. The latter is done in a random way through searching some antibodies in the rule base. If the non-self is known by the artificial immune system, then the destroyer is called to eliminate the non-self. For computer viruses and software faults, the deletion command in the operating system is a kind of destroyer. Otherwise, the rule matching is used to recognize the non-self by the antibody and the rule-base on the adaptive immune tier. The rule includes two parts: the first one is the condition of the non-self feature, and the second one is the conclusion of the rule-based reasoning, which shows the type of the non-self and the elimination approach of the non-self. The rule matching is similar to the combination of DNA genes. And the immune algorithm is built on the random search of the rules. If the random search is done through evolutionary algorithm, then the immune algorithm is built on the evolutionary algorithm (Jiao & Wang, 2000; Deng & Korobka, 2001; Bell, 1970). Cooperative co-evolutionary adaptive genetic algorithm (CCAGA) is suitable for parallel computation, which is convenient to solve complicated problems (Cai & Peng, 2002).

In the immune computing model of the static web immune system, the immune computation has the threshold that is alike in the immune response of many biological immune systems. And the threshold is caused by the long-time cost of the random search. Moreover, the antibody search and rule matching are large-scaled in the chaos state.

A.3.2 Innate Immune Tier

The innate immune tier of the artificial immune system is inspired from the innate immunity of the biological immune system, and this tier is responsible for detecting all the selfs & non-selfs and recognizing the known non-selfs whose features have been stored in the non-self database. The normal model provides a certain representation of selfs and is useful for modeling the static web immune system. The normal model is the information basis of the model for detecting selfs and non-selfs with the space-time properties of the selfs, and the probability for detecting the selfs and the non-selfs can attain 100% in theory.

In the innate immune tier, recognition of foreign objects is relatively simple and quick, because this operation is only based on matching the feature information of the foreign objects and the records of the known non-selfs in the non-self database, and the operation for matching them can be done by the operator for querying. According to the results for querying, the corresponding methods for eliminating the non-selfs may be called from the records, for example some methods is used to delete the files that have been infected by some viruses. In the innate immune tier, there are also the operators for capturing the features of the non-selfs and querying, the non-self database, and so on. If a record of the feature information in the non-self database is matched with the features of the non-self that is being recognized, then the non-self is classified as a known non-self, otherwise it is a unknown non-self.

A.3.3 Adaptive Immune Tier

The adaptive immune tier of the artificial immune system is inspired from the adaptive immunity, and the adaptive immune tier is used to recognize the unknown non-selfs that can not be matched with any record on the feature information in the non-self database. The feature information of all the known non-selfs in the non-self database is represented on the dimension of features in the multi-dimension space, as called the feature space of the non-selfs, and the feature space is used to learn unknown features and types of the unknown non-selfs. By random searching and reasoning in the feature space, the most similar known non-selfs in features are found for the unknown non-selfs. These conclusions are memorized into the non-self database with the operator for memorizing the unknown non-selfs, so that the unknown non-selfs are transformed into new known non-selfs, which shows an immune memory of the artificial immune system. The advantage of such intelligence is to quickly recognize the same type of non-selfs in the innate immune tier as the memorized non-self for the next time and need no recognition in the adaptive immune tier, as the second immune response is much quicker than the first immune response.

The algorithm that is used to recognize the unknown non-selfs is called as artificial antibody, and many parallel antibodies are searching in the feature space.

A.3.4 Parallel Immune Tier

The parallel immune tier is inspired from the parallel immune response of many immune cells in the biological immune system, and parallel immune computation is used to break through the bottleneck of efficiency in the artificial immune system. The parallel immune tier provides a high-performance computing infrastructure for the innate immune tier and the adaptive immune tier, and can be used for solving the problems of limited computing and load balance. The computing capability and resource for one host may be limited for complex immune computation, and computing with overload is unsafe, easy to wither and of high risk. When information that the static web system processes is overload for single processor, the parallel immune tier is called and data are backed up and repaired in multi-threads.

The artificial immune system for the static web system eliminates the non-selfs such as some viruses and abnormity with immune agents and destroyers, but the capability for an immune agent is limited. When a large amount of unknown evil non-selfs attack the static web system, the speed for detecting recognizing and eliminating the non-selfs may be smaller than the speed for the non-selfs to clone transport and attack the system. Therefore, some parts of the artificial immune system may be infected by the non-selfs, and the load for immune computing will be increased. In this way, the speed for immune computing will be smaller and smaller, and then much smaller than the speed for the non-selfs to clone and attack the system. On the other hand, more parts of the artificial immune system will be infected and damaged by the non-selfs and in the end the static web system will be destroyed completely. Moreover, when the artificial immune system is quite abnormal, the immune computation and some destroyers may quicken damage to the static web system.

A.4 SELF/NON-SELF DETECTION OF STATIC WEB IMMUNE SYSTEM

The normal model is a new technique for the artificial immune system, and the model has powerful functions. Therefore, the static web immune system based on the normal model has some new traits, and one of them is the self/non-self detection based on the normal model. Many traditional approaches for detecting abnormity attain low probability for detecting some unknown abnormity, so that many researchers do not believe that there is probability for detecting all the non-selfs in a system. But the normal artificial immune system based on the normal model can detect all the selfs and all the non- selfs in the static web system, in theory.

A.4.1 Self Detection Based on Normal Model

The model for detecting selfs based on the normal model is comprised of the operator for visiting components, the operator for reading the space-time properties, the operator for querying, the static web system and the set of detecting results, shown in Figure 5. First, the operator for visiting components selects a component of the static web system as the object that is being detected in one turn. For the object that has been selected, both the space property and the time property of the object is measured with the operator for reading the space-time properties, and the space-time properties are encapsulated as an immune object. Next, the operator for querying is used to query whether there is any record that matches the immune object. If yes, then the object is determined as a self. Such is repeated until all the components in the static web system are all detected. At last, the results for detecting the selfs are stored into the set of detecting results.

A.4.2 Non-Self Detection Based on Self Detection

The algorithm for detecting the non-selfs is based on the model for detecting the selfs and the recursive logic. For a component that is being detected, the space property and the time property of the component are used to determine whether the component is a self, at first. For a sub-system, the sub-system is checked recursively to detect whether any component of the sub-system is not a self. Suppose the current static web system has l components that are being detected, the algorithm for detecting the non-selfs is show in Figure 6.

A.4.3 Probability for Detecting Selfs and Non-Selfs

According to the unique relationship between the state of the components and the space-time properties of the components, the function can be established between the state set $\{s(c_i)\}$ of the components and the bi-dimension set $\{(p_i, t_i)\}$ of the space-time properties as such.

$$s(c_i) = f((p_i, t_i)) = g(p_i, t_i). \tag{3}$$

Figure 5. Model for detecting selfs based on the normal model. ©2008 Tao Gong. Used with permission.

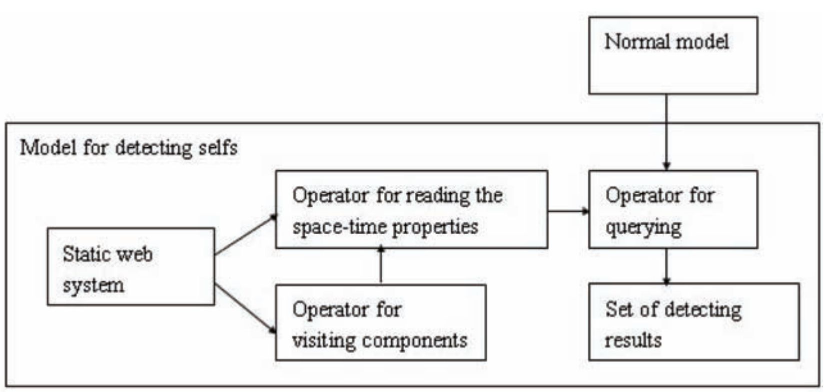
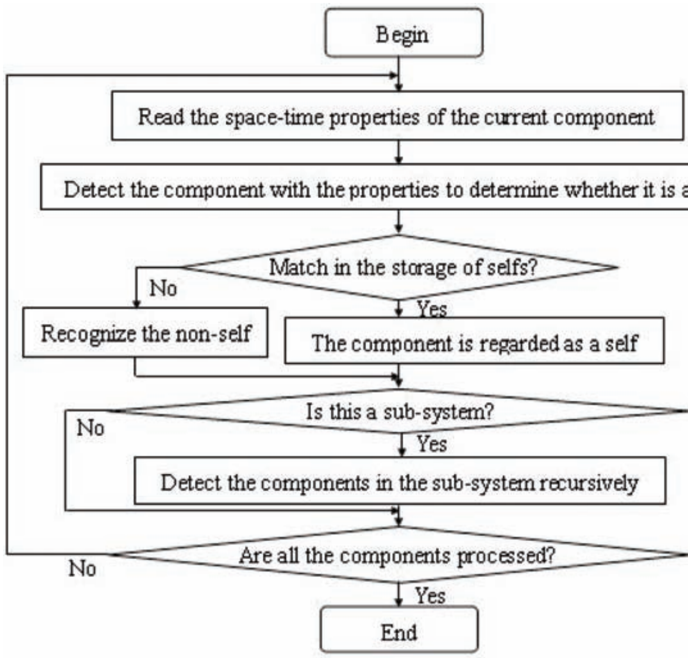


Figure 6. Algorithm for detecting the non-selfs by detecting the selfs. ©2008 Tao Gong. Used with permission.



Here, $f(.)$ represents the mapping from the vector (p_i, t_i) to the state $s(c_i)$ and $g(.)$ represents the mapping from the space property p_i and the time property t_i to the state $s(c_i)$.

To investigate how to improve the approach for detecting the non-selfs, some definitions should be given on the probability for detecting the non-selfs.

Definition A.1 Self/non-self detection of the artificial immune system is a random event, and the event is denoted with D. The measurement on the probability of the random event is called as the probability for detecting the selfs and the non-selfs, and the probability is denoted with $P(D)$. Suppose the number of the selfs that have been detected during immunizing is n_s, the number of the non-selfs that have been detected during immunizing is n_n, the sum of selfs before immunizing is s_s, the sum of non-selfs before immunizing is s_n, then the probability $P(D_s)$ for detecting the selfs can be represented as such.

$$P(D_s) = \frac{n_s}{s_s}. \tag{4}$$

The probability $P(D_n)$ for detecting the non-selfs can be represented as such.

$$P(D_n) = \frac{n_n}{s_n}. \tag{5}$$

Theorem A.2 Suppose an artificial immune system detects the non-selfs in the static web system, which the artificial immune system protects, by matching the features of the non-selfs, the feature set of the known non-selfs is denoted with $U=\{u_i, i=1, 2, \ldots, C\}$, and some non-selfs are unknown, then the probability for detecting the non-selfs is smaller than 1, i.e. $P(D_n)$<1, and the probability can not be 100% in theory and real applications.

[Proof] Suppose the feature set of the unknown non-selfs is denoted with I, the probability for detecting the unknown non-selfs by the artificial immune system is $P(D_n)=|U|/|I|=C/|I|$. Because the non-selfs are unknown, the set I can be an unlimited set, and the element number of the set I can be ∞. Moreover, for any limited set U, there is at least one feature r of one unknown non-self that is not included in the set U, so that the artificial immune system can not detect the unknown non-self. Therefore, $P(D_n)$<1. ■

Theorem A.3 Suppose the data of the space-time property set in the storage of selfs are correct, and the operator for detecting the selfs and the non-selfs by the artificial immune system is normal, the probability for detecting the selfs by the artificial immune system based on the normal model of the static web system can be 100% in theory, i.e. $P(D_s)=1$; the probability for detecting the non-selfs can also be 100%, i.e. $P(D_n)=1$.

[Proof] The normal states of all the components in the static web system, which the artificial immune system protects, are uniquely identified with the bi-dimension data of the space properties and the time properties. Thus, when the normal components are being detected with the normal model of the static web system, all the normal components can be matched in the storage of selfs, so that all the normal components can be detected. Therefore, no normal component has been detected as a non-self, and no abnormal object has been detected as a normal component. Hence, all the non-selfs can be detected.

$$P(D_s) = \frac{n_s}{s_s} = \frac{\sum_{i=1}^{n+m} N(s(C_i))}{s_s} = \frac{s_s}{s_s} = 1$$

$$P(D_n) = \frac{n_n}{s_n} = \frac{l - s_s}{s_n} = \frac{s_n}{s_n} = 1$$

In conclusion, with the artificial immune system and the normal model of the static web system, the probability for detecting the selfs can be 100%, and the probability for detecting the non-selfs can also be 100%. ■

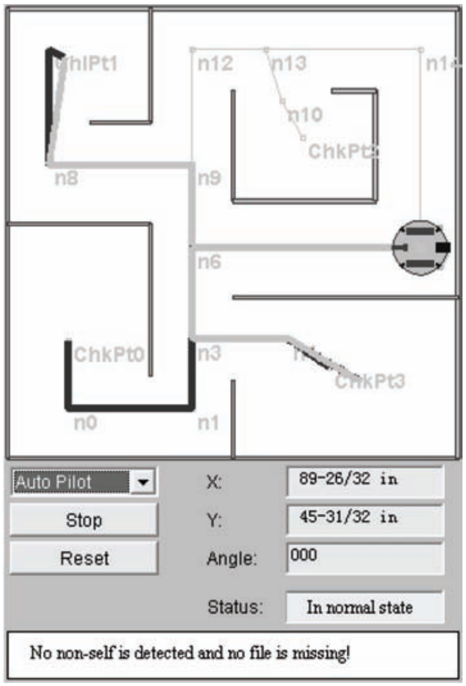
The complete detection of the non-selfs can detect more non-selfs from unknown non-selfs, so that the next step for recognizing the non-selfs will have more non-selfs to learn.

A.4.4 Experiments for Detecting Selfs and Non-Selfs

To test the approach for detecting non-selfs by detecting the selfs with the normal model, a web demo system was infected by some unknown worms and then its immune mechanism was activated. The unknown worms mean that the worms can not be matched in the storage of worms for the artificial immune system, and there are three variants of known worms and two complete-unknown worms among the unknown worms. The three variants are respectively modified from the loveletter worm, the happy-time worm and the Jessica worm (Levy, 2005; Arce & Levy, 2003; Zou, Gong & Towsley, 2002). The worms copy themselves into the file system of the static web system, overwrite the system files and the registration data of the operating system, and spread themselves via the e-mail system. If the static web system is attacked and affected by the worms, the users will feel that the system becomes much slower and more unstable, and the other computers or movable memorizers that connect the static web system may also be affected by the worms.

The experiment was setup on the web demo system of robots, as shown in Figure 7, and the machine was with double CPUs and 2GB memory. First, the selfs of the web demo system were represented with the normal model of the

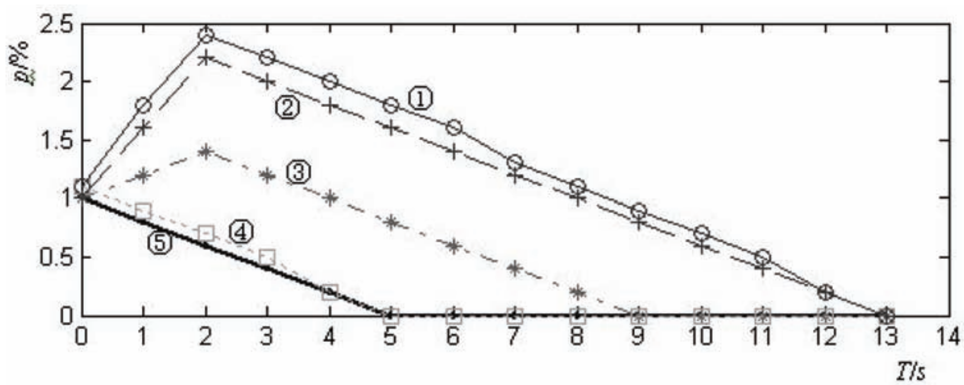
Figure 7. Initial web demo system when the system is normal. ©2008 Tao Gong. Used with permission.



system, and the normal model was uniquely identified by the space-time properties of all the normal components when the initial system was normal. The system was a part of web-based e-learning system, which was developed by some web programming languages such as HTML JSP, Java and JDBC. The system was used to show how a fire-fighting robot searches a path to find the fire and put out the fire, and the initial system was normal. Therefore, at the initial normal state the normal model of the system was built and used to represent the selfs.

To detect and eliminate the worms and the variants in the web demo system, the immune programs were tested in 50 experiments on the machine. The experiments of immune computation show that the non-self percent *p* decreases from the initial value to zero, and the non-self percent may increase first and then decreases due to the spread of the worms, as shown in Figure 8. The time is denoted as *T* and the unit of the time is second. The minimization of the non-self percent means that the elimination of the worms is accomplished and the danger from the worms disappears. The non-self percent means the ratio of the sum of current worms to the sum of components in the current system, and different curves represent different immunization against different worms. The first condition for the first curve is that some files are lost in the Web system; the second condition is that the initial ratio of the sum of worms to the sum of files is 1.1%; and the third condition is that when the artificial immune system detects and eliminates the worms, the worms are activated to copy themselves and produce new worms, until the process of the worm is closed, and then the worms are eliminated. The first condition for the second curve is that no file is lost in the Web system; the second condition is that the initial ratio of the sum of worms to the sum of files is 1%; and the third condition is that when the artificial immune system detects and eliminates the worms, the worms are activated to produce new worms, until the process of the worm is closed, and then the worms are eliminated. The first condition for the third curve is that no file is lost in the Web system; the second condition is that the initial ratio of the sum of worms to the sum of files is 1%; and the third condition is that when the artificial immune system detects and eliminates the worms, the worms are activated to infect some self files, until the process of the worm is closed, and then the worms are eliminated. The first condition for the fourth curve is that some files are lost in the Web system; the second condition is that the initial ratio of the sum of worms to the sum of files is 1.1%; and the third condition is that when the artificial immune system detects and eliminates the worms, the worms are not activated. The first condition for the fifth curve is that no file is lost in the Web system; the second condition is the

Figure 8. Non-self percent of immune computation. ©2008 Tao Gong. Used with permission.



initial ratio of the sum of worms to the sum of files is 1%; and the third condition is that when the artificial immune system detects and eliminates the worms, the worms are not activated.

A.5 RECOGNITION AND LEARNING OF NON-SELFS IN STATIC WEB IMMUNE SYSTEM

For the artificial immune system, the recognition of the non-selfs is classified into two types according to the difference among the non-selfs, and the two approaches for recognizing the non-selfs are inspired from the innate immunity and the adaptive immunity in the biological immune system. The first approach is used to recognize the known non-selfs by matching the features of the known non-selfs in the storage of non-selfs. The second approach is used to recognize some unknown non-selfs by the random search for finding the most similar known non-selfs to the unknown non-selfs in the feature space of the non-selfs and learning the unknown non-selfs.

A.5.1 Recognition of Known Non-Selfs

Before designing the algorithm for recognizing the known non-selfs on the model for recognizing the known non-selfs, the storage of non-selfs must be designed. The storage of non-selfs is used to represent and store the feature information of all the known non-selfs, and the feature information includes the space information of the non-selfs, the feature string of the non-selfs, the copying behavior of the non-selfs, the modifying behavior of the non-selfs to the registration data, the calling behavior of the non-selfs to the e-mails and so on.

The multi- dimension feature information of the non-selfs is encapsulated into the feature objects of the non-selfs, and the feature objects of the non-selfs consist of the set of feature information for the non-selfs. The set of feature information is called as the feature space of the non-selfs, shown in Table 1.

In Table 1, the feature information of the known non-selfs should be input into the storage of non-selfs first, and the software of database management can be used to import the data. For example, for the loveletter worms, if the worms are known for the artificial immune system, then the data of the loveletter worms can be input into the storage of non-selfs. The data include 1) Non-self No.: w00001, 2) Space information of the non-selfs (File-extension): vbs, 3) Feature string of the non-selfs: loveletter, 4) Copying behavior of the non-selfs: .Copy, 5) Modifying behavior of the non-selfs to the registration data: .RegWrite, 6) Calling behavior of the non-selfs

Table 1. Dimension of the feature space for the non-selfs

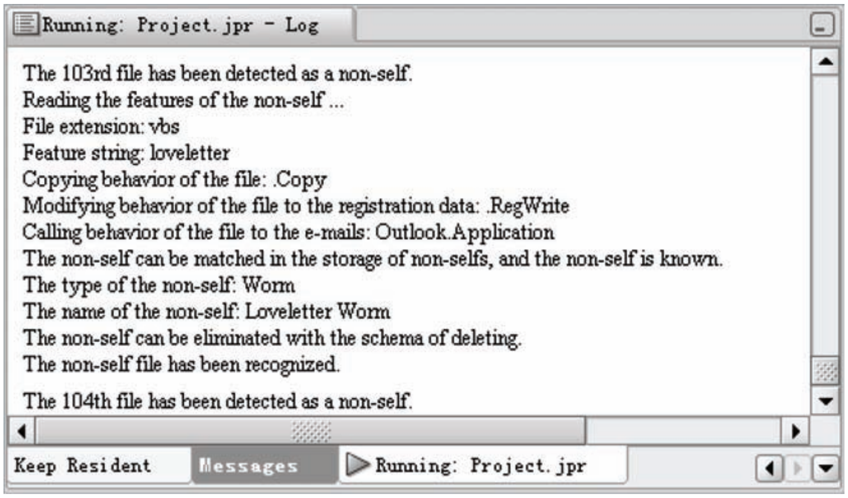
Dimension No.	Dimension of non-selfs' features	Description of non-selfs' features
1	u_1	Non-self No.
2	u_2	Space information of the non-selfs
3	u_3	Feature string of the non-selfs
4	u_4	Copying behavior of the non-selfs
5	u_5	Modifying behavior of the non- selfs to the registration data
6	u_6	Calling behavior of non-selfs to the e-mails
7	u_7	Non-self name
8	u_8	Type of the non-self
9	u_9	Eliminating schema of the non-self
…	…	…

to the e-mails: Outlook.Application, 7) Non-self name: loveletter worm, 8) Type of the non-self: worm, 9) Eliminating schema of the non-self: delete and so on.

At first, the algorithm for recognizing the known non-selfs queries the space information of the non-self in the feature space, which is represented in the storage of non-selfs. If a record is matched with the space feature of the non-self that is being recognized, then other features of the non-self are recognized for the next steps, otherwise the non-self is regarded as an unknown non-self. Afterwards, the feature string, the information on the copying behavior, the information on the modifying behavior to the registration data and the information on the calling behavior to the e-mails are extracted from the non-selfs. If the features of the non-selfs can be matched with any samples in the feature space that has been built with all the known non-selfs, then the non-self is a new sample of known non-selfs.

In the static web demo system, the algorithm for recognizing the known non-selfs reads the file extension of the non-self file at first. If at least one record is matched on the file extension, then the algorithm for recognizing the known non-selfs continues to recognize the other features of the non-self and classifies the non-self into an existing type of the known non-selfs; otherwise, the file is regarded as a unknown non-self, and the algorithm for learning the unknown non-selfs is used to recognize the unknown non-self. After the matching record is found, the information of the record is read according to the feature string, the copying behavior, the modifying behavior to the registration data, and the calling behavior to the e-mails and so on. For example, the loveletter worm has been recorded as a known non-self

Figure 9. Recognition of the known non-selfs by the algorithm. ©2008 Tao Gong. Used with permission.



in the storage of non-selfs, and the No. of the worm is w00001, so the result for recognizing the loveletter worm is shown in Figure 9.

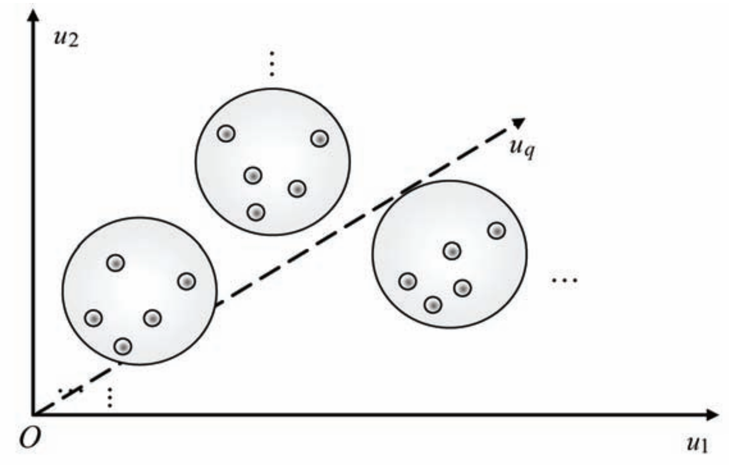
In Figure 9, the file that has been infected by the loveletter worm has been detected as a non-self. Afterwards, the algorithm for recognizing the known non-selfs is used to recognize the file's features such as the file extension, the feature string, the copying behavior of the file, the modifying behavior of the file to the registration data, the calling behavior of the file to the e-mails and so on. By matching the features of the non-self with the records in the storage of non-selfs, the non-self is recognized as a known non-self, and the type of the non-self is worm. The name of the non-self is Loveletter Worm, and the schema of deleting can be used to eliminate the non-self.

A.5.2 Feature Space that is Built with all the Known Non-Selfs

Suppose the feature dimension of all the known non-selfs is q, the feature vector of the non-self c_j among the known non-selfs is denoted with $(u_{j1}, u_{j2}, \ldots, u_{jq})$, the feature space of the non-selfs is represented with $\{(u_{j1}, u_{j2}, \ldots, u_{jq})\}$, shown in Figure 10, j=1, 2, …, M, and M denotes the sum of the known non-selfs. The known non-selfs are classified into many classes, and the number of the classes is limited and numerable at any time. After the process for learning the unknown non-selfs, the unknown non-selfs are classified into the type of the most similar known non-selfs to the unknown ones or given new classes, according to the feature vector of the unknown non-selfs. Sometimes, the unknown non-selfs can not be classified into any type of known non-selfs, and new classes must be created for the unknown ones at that time. With creation of new class repeated, the classes of unknown non-selfs may be unlimited but numerable.

In Figure 10, the dimension coordinate of the feature space for the non-selfs is represented with u_i, i=1, 2, …, q, small balls are used to denote the non-selfs, and the big circles represent the classes of the non-selfs. For the problem for learning unknown non-selfs with unlimited class, current approaches of machine learning are not quite suitable, so that the class of unknown non-selfs is regarded as a limited variable in some applications.

Figure 10. Feature space of non-selfs with unlimited class expending. ©2008 Tao Gong. Used with permission.



A.5.3 Learning of Unknown Non-Selfs

Suppose the current static web system that the artificial immune system protects has l components, among which the algorithm for recognizing known non-selfs has regarded l_2 components as unknown non-selfs. There is some feature information for K known non-selfs in the storage of non-selfs, and each non-self has q features that are coded in some tables. The problem for recognizing the unknown non-selfs can be solved by finding the most similar known non-self to the unknown non-self and/or creating a new class for the unknown non-self, and the problem for finding the most similar known non-self to the unknown non-self is a constrained optimization problem. Suppose the difference between the unknown non-self c_u and the known non-self c_i is represented with the function $f(c_u, c_i)$, the constrained optimization problem is described as such (Cai & Wang, 2006).

minimize $f(c_u, c_i)$ $c_u = (u_{u1}, x_{u2}, \cdots, x_{uq}) \in \Re^q$

$$c_i = (u_{i1}, x_{i2}, \cdots, x_{iq}) \in \Re^q$$

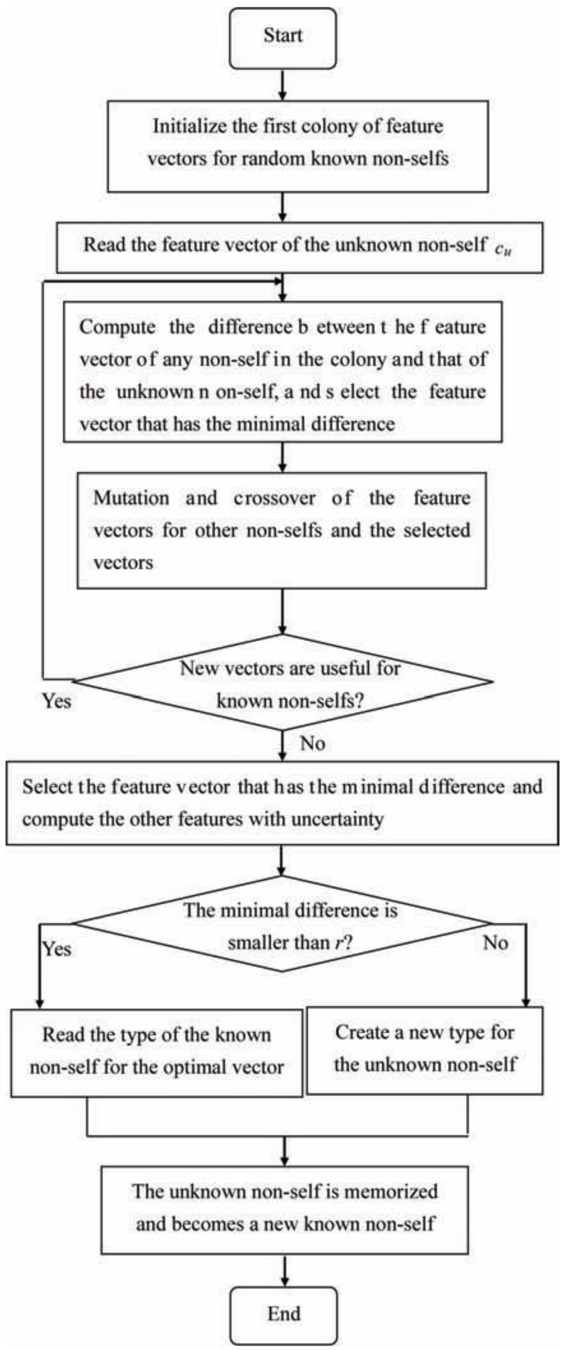
subject to: $f(c_u, c_i) < r$ (6)

Here, r represents the threshold that is used to determine if the unknown worm belongs to any class of known non-selfs.

The constrained optimization problem for finding the most similar known non-self to the unknown non-self can be solved with some evolutionary algorithms, and the algorithm for recognizing the unknown non-selfs is designed with immune memory, shown in Figure 11. The immune memory means that the unknown non-self can be transformed into a known one after learning, and the immune learning is enhanced.

- **Step 1.** Select some known non-selfs with random probability from the storage of known non-selfs, and build the first colony for evolutionary computation with the feature vectors of the selected non-selfs.
- **Step 2.** Read the measurable features of unknown non-self c_u, and build its feature vector.
- **Step 3.** Compute the distance between the feature vectors of the unknown non-self

Figure 11. Algorithm for recognizing the unknown non-selfs with immune memory. ©2008 Tao Gong. Used with permission.



c_u and each known non-self in the colony, and sort the distances that represent the difference between the unknown non-self and the known non-selfs. Select the minimal distance and keep the relevant feature vector and known non-self.

- **Step 4.** Mutate and cross in the colony of feature vectors with the feature vectors of other known non-selfs and the feature vector of the unknown non-self.
- **Step 5.** Match the new vector with the feature vectors of all the known non-selfs. If the new vector belongs to any other known non-self, call step 3 to continue the random search; otherwise end the loop, and select the feature vector that has the minimal distance. Then compute unknown features of the unknown non-self with the feature vector that has the minimal distance.
- **Step 6.** Determine whether the minimal distance is smaller than *r*. If yes, then the unknown non-self belongs to the type of the known non-self whose feature vector has the minimal distance; otherwise, the unknown non-self is regarded as a new class of non-selfs and a new class is created for the unknown non-self.
- **Step 7.** Memorize the feature vector of the unknown non-self, and then the unknown non-self becomes a new known non-self.

A.5.4 Probability for Learning Non-Selfs

Before the algorithm for recognizing the unknown worms is used, the unknown worms must be detected. The algorithm for detecting selfs and non-selfs on the normal model is used to detect whether the object is a self or non-self, and the algorithm for recognizing the known non-selfs is used to determine whether the non-self is a known non-self or unknown non-self. For recognizing more unknown non-selfs the stage for detecting selfs and non-selfs is very crucial, and the stage for recognizing known non-selfs is relatively simple.

With the space properties and the time properties of the components, the normal model uniquely identifies the normal state of each component and the normal state of the whole static web system that the artificial immune system protects. Therefore, based on Theorem A.3, the following theorem shows the advantages of the normal model to the algorithm for recognizing the unknown non-selfs.

Theorem A.4 Suppose the event that the artificial immune system detects the non-selfs by matching the features of the non-selfs is denoted with D_T, the event that the artificial immune system detects the non-selfs is denoted with D_N, the probability for recognizing the unknown non-selfs based on the normal model is represented with $P(R|D_N)$, the probability for recognizing the unknown non-selfs only based on matching the features of the non-selfs is represented with $P(R|D_T)$, then $P(R|D_N)>P(R|D_T)$.

[Proof] The artificial immune system recognizes the unknown non-selfs after the system detects some non-selfs as the unknown non-selfs. Thus, the probability for recognizing the unknown non-selfs is a conditional probability $P(R|D)$, which is the probability of the event that unknown non-selfs are recognized on the condition that the non-selfs are detected.

$$\because P(D_N) = 1,\ P(D_T) < 1,$$
$$\therefore P(D_N) > P(D_T)$$
$$\therefore P(R \mid D_N) = P(R) > P(R \mid D_T)$$
■

A.5.5 Experiments for Learning Non-Selfs

After the experiment for detecting selfs and non-selfs were done, the experiment for learning non-selfs was activated. The algorithm for detecting selfs and non-selfs on the normal model was used to detect whether the object is a self or non-self, and the algorithm for recognizing the known non-selfs was used to determine whether the non-self is a known non-self or unknown non-self.

In the web demo system shown in Figure 7, 50 independent trials for learning the non-selfs are executed with MATLAB offline. In the algorithm for learning unknown worms with evolutionary computation, the common recombination operator involve simulated binary crossover (SBX), simplex crossover (SPX) (Tsutsui, Yamamure & Higuchi, 1999) etc. In this test, the algorithm adopts SPX, which generates offspring based on uniform probability distribution and does not need any fitness information, as the recombination operator. In $\Re^n$, μ mutually independent parent vectors $(\vec{x_i}, i = 1, \cdots, \mu)$ form a simplex, and Figure 12 illustrates the density of the offspring produced with three-parent SPX.

The production of an offspring consists in: 1) employing a certain ratio to expand the original simplex in each direction $\vec{x_i} - \vec{o}$ ($\vec{o}$ is the center of μ vectors, $\vec{o} = (1) / (n+1) \sum_{i=1}^{\mu} \vec{x_i}$) and forming a new simplex; and 2) choosing one point form the new simplex as an offspring.

The evolutionary search was tested with some benchmark functions such as g1, g2, g3, g10,

Figure 12. Density of the offspring produced with three-parent SPX. ©2008 Tao Gong. Used with permission.

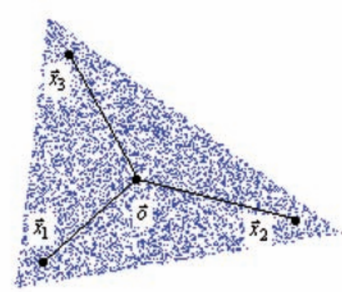


Table 2. Results of evolutionary searching for learning the non-selfs

fcn	optimal	best	median	mean	st. dev	worst	average percentage
g01	-15.000	-15.000000	-15.000000	-15.000000	2.522E-08	-15.000000	95
g02	-0.803619	-0.803241	-0.802556	-0.801258	3.832E-03	-0.792363	86
g03	-1.000	-1.000000	-1.000000	-1.000000	1.304E-12	-1.000000	91
g10	7049.248	7049.248020	7049.248020	7049.248020	1.502E-12	7049.248020	92
g11	0.750	0.750000	0.750000	0.750000	1.546E-12	0.750000	56
g12	-1.000000	-1.000000	-1.000000	-1.000000	0.000E+00	-1.000000	100

g11, g12 (Cai & Wang, 2006), and the evolutionary algorithm in the static web immune system showed good optimum and performance, shown in Table 2. For example, when the benchmark function g10 was tested with 100 experiments, the convergence curves of the two experiments among them are shown in Figure 13. At first, the evolutionary search with some constraints jumped up and down beside the optimum and then found much closer solutions to the optimum than before until the convergence of the algorithm was accomplished.

The immune memory is a kind of rote learning, and the memory part can be regarded as a function $m(\cdot)$ in mathematics. The input vector of the memory function is $(u_{j_{i_1}}, u_{j_{i_2}}, \cdots, u_{j_{i_o}})$, and the output vector of the memory function is the combination of unknown features and unknown type of the unknown worm, as denoted with $(u_{j_{l_1}}, u_{j_{l_2}}, \cdots, u_{j_{l_{q-o}}}, T)$. The immune memory can be searched directly and easily, and no repeated immune computation is needed for learning unknown features and type, when the memory function $m(u_{j_{i_1}}, u_{j_{i_2}}, \cdots, u_{j_{i_o}})$ is called (Cai & Xu, 2004).

$$(u_{j_{i_1}},\cdots,u_{j_{i_o}}) \xrightarrow{m} (u_{j_{l_1}},\cdots,u_{j_{l_{q-o}}},T) \xrightarrow{\text{storing}} ((u_{j_{i_1}},\cdots,u_{j_{i_o}}),(u_{j_{l_1}},\cdots,u_{j_{l_{q-o}}},T)) \quad (7)$$

In this example, the three variants v_1, v_2, v_3 are recognized to belong to three classes of the loveletter worms, the happy-time worms, and the Jessica worms respectively, and the three

Figure 13. Convergence of the two evolutionary searching curves. ©2008 Tao Gong. Used with permission.

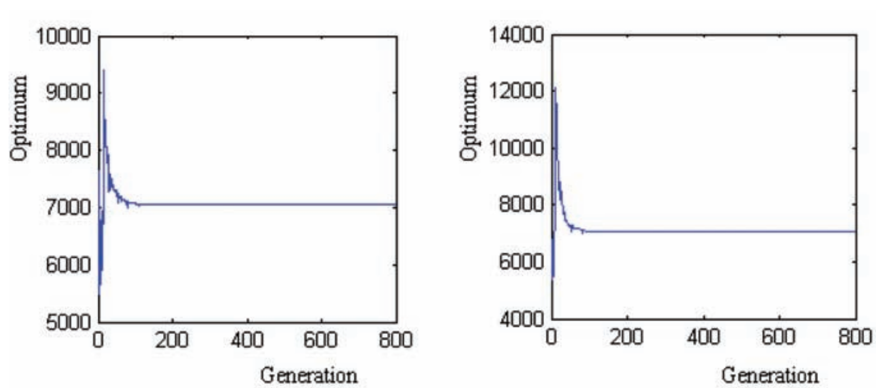
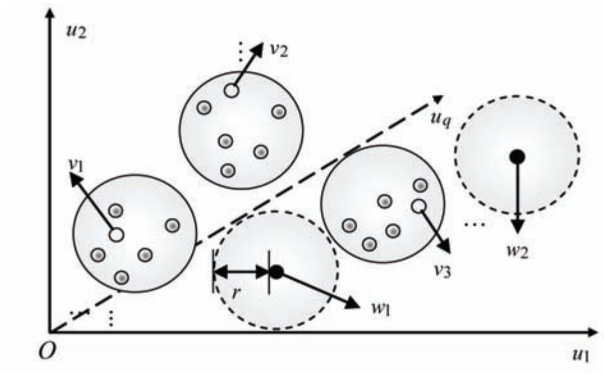


Figure 14. Learning results of the unknown worms in the feature space. ©2008 Tao Gong. Used with permission.



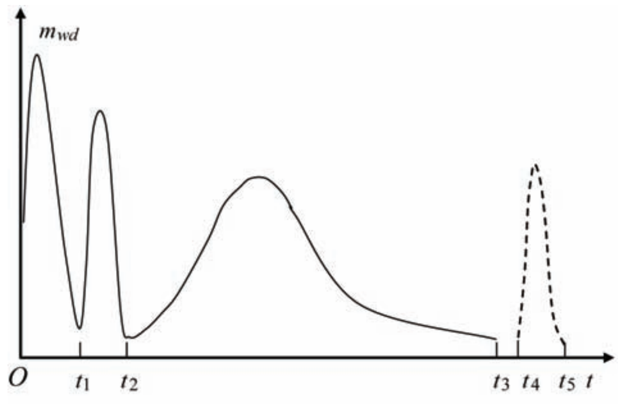
classes are known, shown as the real-line circle in Figure 14. The other unknown worms w_1, w_2 are recognized as two brand-new unknown worms, and two new classes are created for them, shown as the dashed circle.

The web demo system is immunized by the artificial immune system and is now immune from some worms. When many known worms and the unknown worms attack the demo system that the artificial immune system protects, the innate immune tier is activated to detect the worms and recognized the known worms. Detection of all the worms and recognition of the unknown worms are both quick because of the normal model and the storage of known worms, shown as the curve from the time point 0 to t_2 in Figure 15. After the innate immune tier confirms that the unknown worms are not any known worms, the adaptive immune tier begins to learn the unknown worms with random evolutionary search, shown as the curve from the time point t_2 to t_3 in Figure 15, and the most similar known worm is found to decide whether the unknown worms belong to any type of known worms or are really new type of worms. The learning results are memorized so that the unknown worms are transformed into new known worms in the end. If another variant of the loveletter worm attacks the demo system, the artificial immune system recognizes the variant as a known worm now and the immune response is quick from the time point t_4 to t_5 in Figure 15. Here, m_{wd} represents the sum of the worms that have been processed, and t represents the time coordinate.

In Figure 15, the primary immune response includes the self/non-self detection and recognition of the known worms, and the detection is accomplished from the time point 0 to t_1. The secondary immune response is from the time point t_2 to t_3, which is much longer than the primary immune response. In fact, the hypothetic immune response from the time point t_4 to t_5 is a part of the primary immune response after the secondary immune response.

The experiments are made on the web-based course system, and the web demo system is a part. Over a hundred of worms attack the web system, and many files are infected. The artificial immune system detects all the worms successfully with the normal model and the approach for detecting selfs and non-selfs. But only with some intelligent techniques such as the artificial neural network, the probability for detecting the non-selfs is smaller, and such result affects the process of recognizing

Figure 15. Immune response to the known worms and the unknown worms. ©2008 Tao Gong. Used with permission.



the worms. With the immune learning algorithm after detecting the worms with the normal model, the artificial immune system recognizes the unknown worms with the higher probability than the result with the artificial neural network with the BP algorithm.

A.6 CONCLUSION

The tri-tier immune computing model is a useful and effective model for the static Web immune system. The first tier is the innate immune computing tier, the second tier is the adaptive immune computing tier, and the last tier is the parallel immune computing tier.

The theorems prove that, on the condition that the time property is correct, the normal model and the tri-tier immune computing model are useful for increasing the probability of detecting selfs and non-selfs in the static web system. Many anti-worm experiments of immunization validate effectiveness and advantages of the static web immune system and the immune algorithms. The static web immune system can eliminate worms and repair itself with higher probability for detecting non-selfs than traditional approaches.

A.7 FUTURE DIRECTIONS

This work emphasized on static web immune system, and this is new try of new techniques for immune computing. But this is just a beginning, the next work is emphasizing and will emphasize on dynamic web immune system. The related issues abut the dynamic web immune system can be shown in the following:

1) Normal model of dynamic web immune system. How to build the normal model for the dynamic web immune system and how to transform the normal model from a normal state to another normal state are the difficult bottlenecks for representing the selfs.
2) Adaptive learning of unknown non-selfs. If the viruses are designed by the most intelligent programmers, how to design the adaptive learning mechanism of complete-unknown non-selfs is really a difficult problem for other most intelligent programmers.
3) Design of artificial immune system for the operating systems. It is very difficult to build a normal model for a Windows operating system, because the designers for the Windows operating system do not really know when the system is normal or abnormal.

ACKNOWLEDGMENT

This work was supported in part by the Shanghai Natural Science Foundation under Grant 08ZR1400400, the Shanghai Educational Development Foundation under Grant 2007CG42 and Donghua Univ. Foundation under Grant #104-10-0044017.

REFERENCES

Arce, I., & Levy, E. (2003). An analysis of the Slapper worm. *IEEE Security & Privacy*, *1*(1), 82–87. doi:10.1109/MSECP.2003.1177002

Balachandran, S. (2005). *Multi-shaped detector generation using real-valued representation for anomaly detection*. Masters Thesis, University of Memphis.

Balthrop, J., Forrest, S., & Newman, M. E. J. (2004). Technological Networks and the Spread of Computer Viruses. *Science*, *304*(5670), 527–552. doi:10.1126/science.1095845

Bell, G. I. (1970). Mathematical model of clonal selection and antibody production. *Journal of Theoretical Biology*, *29*(2), 191–232. doi:10.1016/0022-5193(70)90019-6

Cai, Z. X., & Peng, Z. H. (2002). Cooperative coevolutionary adaptive genetic algorithm in path planning of cooperative multi-mobile robot systems. *Journal of Intelligent & Robotic Systems*, *33*(1), 61–71. doi:10.1023/A:1014463014150

Cai, Z. X., & Wang, Y. (2006). Multiobjective optimization based evolutionary algorithm for constrained optimization. *IEEE Transactions on Evolutionary Computation*, *10*(6), 658–675. doi:10.1109/TEVC.2006.872344

Cai, Z. X., & Xu, G. Y. (2004). Artificial Intelligence: Principles and Applications (Third Edition, Graduate Book), Beijing: Tsinghua University Press.

Chao, D. L., Davenport, M. P., & Forrest, S. (2004). Modelling the impact of antigen kinetics on T-cell activation and response. *Immunology and Cell Biology*, *82*(1), 55–61. doi:10.1111/j.1440-1711.2004.01207.x

De Castro, L. N., & Timmis, J. (2002). *Artificial Immune Systems: A New Computational Intelligence Approach*. London: Springer-Verlag.

De Castro, L. N., & Von Zuben, F. J. (2002). Learning and optimization using the clonal selection principle. *IEEE Transactions on Evolutionary Computation*, *6*(3), 306–313. doi:10.1109/TEVC.2002.1011539

Deng, Y., & Korobka, A. (2001). Performance of a supercomputer built with commodity components. *Parallel Processing*, *27*(12), 91–108.

Einstein, A. (1920). *Relativity: the Special and General Theory*. Three Rivers Press, New York.

Fauci, S. A. (2003). HIV and AIDS: 20 years of science. *Nature Medicine*, *9*(7), 839–843. doi:10.1038/nm0703-839

Gong, T., & Cai, Z. X. (2003). Parallel evolutionary computing and 3-tier load balance of remote mining robot. *Trans Nonferrous Met Soc China*, *13*(4), 948–952.

Gonzalez, F., & Dasgupta, D. (2003). Anomaly detection using real-valued negative selection. *Journal of Genetic Programming and Evolvable Machines*, *4*(4), 383–403. doi:10.1023/A:1026195112518

Hawking, S., & Penrose, R. (1996). *The nature of space and time*, Princeton University Press.

Jerne, N. K. (1974). Towards a network theory of the immune system. *Annals of Immunology*, *125C*, 373–389.

Jiao, L. C., & Wang, L. (2000). Novel genetic algorithm based on immunity. *IEEE Trans on Systems, Man and Cybernetics — Part A. Systems and Humans*, *30*(5), 552–561.

Levy, E. (2005). Worm Propagation and Generic Attacks. *IEEE Security and Privacy*, *3*(2), 63–65. doi:10.1109/MSP.2005.57

Orman, H. (2003). The Morris Worm: A Fifteen-Year Perspective. *IEEE Security & Privacy*, *1*(5), 35–43. doi:10.1109/MSECP.2003.1236233

Perelson, A., Hightower, R., & Forrest, S. (1996). Evolution (and learning) of v-region genes. *Research in Immunology*, *147*, 202–208. doi:10.1016/0923-2494(96)87221-X

Tsutsui, S., Yamamure, M., & Higuchi, T. (1999). Multi-parent recombination with simplex cross-over in real coded genetic algorithms. in *Proc. Genetic and Evol. Comput. Conf.*, 657–664.

Zou, C. C., Gong, W., & Towsley, D. (2002). Code Red Worm Propagation Modeling and Analysis. In: Atluri V. eds. *Proc of the 9th ACM Conf on Computer and Communications Security*, ACM Press, New York, 138-147.

KEY TERMS AND DEFINITIONS

Adaptive Immune Tier: The immune computing tier, which learn and recognize the unknown non-selfs, is called as the adaptive immune tier of the artificial immune system.

Immune Memorization: The process for remembering the unknown non-selfs to transform the non-self into the known ones is called as the immune memorization.

Innate Immune Tier: The immune computing tier, which detects the selfs & non-selfs and recognize all the known non-selfs, is called as the innate immune tier of the artificial immue system.

Non-Self Database: The database that stores the feature information of the known non-selfs is called as the non-self database.

Normal Model of Normal Static Web System: The set of space-time properties for all the normal components of the normal static web system is called as the normal model of the normal static web system.

Parallel Immune Tier: The immune computing tier, which uses parallel computing to increase efficiency and load balance of immune computation, is called as the parallel immune tier of the artificial immune system.

Probability for Detecting Non-Selfs: The measurement on the probability of the random event that the artificial immune system detects the non-selfs is called as the probability for detecting the non-selfs.

Probability for Learning Unknown Non-Selfs: The measurement on the probability of the random event that the artificial immune system learns the unknown non-selfs is called as the probability for learning the unknown non-selfs.

Self Database: The database that stores the space-time information of the selfs is called as the self database.

Self/Non-Self Detection: The process for detecting the object to decide whether the object is a self or non-self is called as the self/non-self detection.

This work was previously published in the Handbook of Research on Artificial Immune Systems and Natural Computing: Applying Complex Adaptive Technologies, edited by Hongwei Mo, pp. 219-239, copyright 2009 by Medical Information Science Reference (an imprint of IGI Global).

Chapter 7.18
Mapping Policies to Web Rules: A Case of the KAoS Policy Language

Nima Kaviani
University of British Columbia, Canada

Dragan Gašević
Athabasca University, Canada

Marek Hatala
Simon Fraser University, Canada

ABSTRACT

Web rule languages have recently emerged to enable different parties with different business rules and policy languages to exchange their rules and policies. Describing the concepts of a domain through using vocabularies is another feature supported by Web rule languages. Combination of these two properties makes web rule languages appropriate mediums to make a hybrid representation of both context and rules of a policy-aware system. On the other hand, policies in the domain of autonomous computing are enablers to dynamically regulate the behaviour of a system without any need to interfere with the internal code of the system. Knowing that policies are also defined through rules and facts, Web rules and policy languages come to a point of agreement, where policies can be defined through using web rules. This chapter focuses on analyzing some of the most known policy languages (especially, KAoS policy language) and describes the mappings from the concepts for KAoS policy language to those of REWERSE Rule Markup Language (R2ML), one of the two proposals to Web rule languages.

DOI: 10.4018/978-1-60566-402-6.ch024

INTRODUCTION AND MOTIVATION

Rules are among the most frequently used knowledge representation techniques. They can generally be categorized as *reaction rules* (event-condition-actions), *integrity rules* (rules of consistency checking), *derivation rules* (implicitly, derived rules), and *production rules* (Boley, Tabet, & Wagner, 2001).

Facts can also be regarded as derivation rules with no premises.

Recently, rule markup languages have started to be considered as the vehicle for using rules on the Web and in other distributed systems, forming a new category of rule languages, referred to as Web rule languages (Wagner, Giurca, & Lukichev, 2006). The main strength of markup languages is their machine readability combined with their inherent potentials to easily be linked to distributed knowledge sources also represented in the form of markup languages (e.g., Data Type Definition (DTD), XML Schema (Fallside & Walmsley, 2004), Resource Description Framework (RDF) (Lassila & Swick, 1999), and Web Ontology Language (OWL) (Smith, Welty, & McGuinness, 2004)). Rule markup languages allow for the reuse, interchange, and publication of rules as well as their communication and execution in a distributed environment. More specifically, rule markup languages allow for specifying business rules as modular, standalone, units in a declarative way, and publishing and interchanging them between different systems and tools (Wagner, Giurca, & Lukichev, 2006). Rule markup languages play an important role in facilitating business-to-customer (B2C) and business-to-business (B2B) interactions over the Internet by enabling information exchange across various stakeholders. For example, they may be used to express derivation rules for enriching Web ontologies by adding definitions of derived concepts, or for defining data access permissions; to describe and publish the reactive behaviour of a system in the form of reaction rules; and to provide a complete XML-based specification of a software agent (Wagner, Giurca, & Lukichev, 2005).

RuleML and *REWERSE Rule Markup Language (R2ML)* are two newly and rapidly emerging Web rule languages that help with interchanging various types of rules from one rule domain to another. RuleML represents an initiative for creating a general rule markup language that will support different types of rules and different semantics (Boley, Tabet, & Wagner, 2001). R2ML more or less follows a similar idea to RuleML; however, its design follows Model Driven Architecture (MDA) defined by the Object Management Group (OMG) (Miller & Mukerji, 2003). Moreover, R2ML is a trial to cover a more comprehensive set of atoms and elements required to develop and define rules, thus bringing more flexibility to rule definition. These rule languages are designed to be conformant and compatible with the guidelines and use cases defined by the Rule Interchange Format (RIF) working group (Ginsberg, Hirtle, McCabe, & Patranjan, 2006).

Policies in the domain of autonomous computing are guiding plans that restrict the behaviour of autonomous agents in accessing the resources (Toninelli, Bradshaw, Kagal, & Montanari, 2005). The main advantage in using policies is the possibility to dynamically change the behaviour of the system by adjusting the policies without interfering with the internal code of the system. They can authorize/oblige the users to, or prohibit/dispense them from, accessing the resources or taking particular actions in the system. Policy languages can be considered as instructional sets that enable phrasing and putting systematic guidelines in place for a target agent. KAoS (Uszok, et al., 2003) is one of the most known policy languages that goes beyond the traditional policy systems by giving special care to the context to which the policies are applied. This is done by enabling these policy languages to use domain knowledge (a.k.a, vocabularies) readily available on the Internet and represented in knowledge representation markup languages such as XML-Schemas, RDF, and OWL.

Knowing that policies are also defined through the use of rules and facts that can share online knowledge bases, Web rules and policy languages come to a point of agreement. Web rules and policies get even closer bringing it into the consideration that most of the newly emerging policy languages also have chosen markup syntax for their specifications (e.g., KAoS). Following the hierarchy of rules represented earlier, a policy rule can be considered a reaction rule which generates an action (possibly permission or denial) upon the occurrence of an event and satisfaction of a series of conditions. On the other hand, a policy rule can be considered a derivation rule that leads to generating additional facts once a set of conditions are satisfied (e.g., a policy may result in concluding that a resource is no longer in use, thus adding an extra fact to the knowledge base). Further to this, a policy rule might be defined in the form of an integrity rule to preserve the integrity of information in a system. All this brings Web rules and policies to a joint end where Web rules can be used to describe policies. The inherent flexibility of Web rule languages in using various vocabularies overcomes the intricate problem of reusing policies across various domains. What is more interesting is that the recent policy languages (including KAoS) can be defined over domain knowledge represented in markup languages such as OWL (Smith, Welty, & McGuinness, 2004) and XML Schema (Fallside & Walmsley, 2004), making it a lot easier to integrate them with Web rules. Furthermore, Web rules have been designed to be compatible with different subsets of first order logic (FOL), ranging from description logic to rule-based logic programs. Consequently, the essence and the semantics of the policy rules specified in different policy languages can be captured agnostic to the logical domain in which the language is grounded. This in turn brings more flexibility and compatibility to Web services to be combined with policies defined in the form of Web rules. There are various benefits in providing bidirectional transformations between policies and Web rule languages. First and foremost, it enables sharing of policies between various business processes which in turn helps with policy reuse from a business system to another. Sharing policies across various business processes is another step towards facilitating B2B and B2C interactions, as discussed above. Additionally, the possibility of converting policies from one language to another, while preserving the semantics of the policies, facilitates the deployments of policies in a business system regardless of the underlying technologies used by its business processes.

In order to be able to decide how a markup policy language can best be described through using a Web rule language, there needs to be a deep analysis of the constructs and elements of one domain to be matched with the corresponding elements in the other domain. This chapter focuses on discussing the building blocks of defining policies (from their logics to their concepts) using markup policy languages, and to find the appropriate matches in a corresponding Web rule language. For the policy languages we will mainly focus on KAoS while for the Web rule we will narrow our focus down to R2ML. This will later on lead us to discuss the possibilities for using a unified modeling language to define and interchanging policies between different policy domains, or even more, to rely on a modeling language to facilitate the definition, design, and integration of policies with other parts of a system, including but not limited to business processes.

This chapter is organized as follows. In Section 2 we will provide some background information on different types of underlying logic for defining policies and rules, some of the most known policy languages, the idea behind rule interchange format, and also a background on Web rule languages. Section 3 discusses our policy interchange framework. In Section 4 we present the metamodels for R2ML and KAoS to discuss high level exchange of concepts between the elements of these languages along with the graphical QVT transformations from each language to the other one and back. Section 5 shows a real example of applying the QVT transformations to the concrete representation of a KAoS policy rule which is then followed by a discussion and conclusion to our findings in Section 6.

BACKGROUND

Before going into the details of how policy languages can be converted to Web rule languages and transformed from one business domain to another, in this section we provide brief background information on the existing policy and Web rule languages. We also discuss how following the principle of modeling can facilitate the process of defining and transforming policy rules from one policy language to another.

Rule Interchange Format

Most of the proposals on Web rule languages are trying to address the use cases and requirements defined by Rule Interchange Format Working Group (Ginsberg, Hirtle, McCabe, & Patranjan, 2006). *Rule Interchange Format (RIF)* (Ginsberg, Hirtle, McCabe, & Patranjan, 2006) is an initiative to address the problem of interoperability between existing rule-based technologies. RIF is desired to play as an intermediary language between various rule languages and not as a semantic foundation for the purpose of reasoning on the Web. It aims to be a widely adopted W3C consistent standard for exchanging rules among different stakeholders. RIF Working Group has defined 10 use cases which have to be covered by a language compliant to the RIF's properties, three of which are dealing with policies and business rules, namely: *Collaborative Policy Development for Dynamic Spectrum Access*, *Access to Business Rules of Supply Chain Partners*, and *Managing Inter-Organizational Business Policies and Practices*. *SWRL* (Horrocks, Patel-Schneider, Boley, Tabet, Grosof, & Dean, 2004) and *RuleML* (Boley, Tabet, & Wagner, 2001) are two of the ongoing efforts in this area trying to serve as rule languages to publish and share rule bases on the Web.

Web Rule Languages

Web rule languages are being developed following the RIF Request for Proposal (RFP) for designing a Web rule language that in the first place provides an intermediary language for exchanging rules between various stakeholders and business processes. The current Web rule languages are not trying to facilitate reasoning on the Web and as a result there is no reasoning engine or infrastructure developed for them. On the other hand, the idea is to have sound transformations from a source rule language to a target rule language, such that the reasoner at the destination (which is of course conformant to the rule language at the destination) can successfully reason over the transformed rules and derive the same conclusions as of the ones at the source. RuleML (Boley, Tabet, & Wagner, 2001) and R2ML (Wagner, Giurca, & Lukichev, 2005) are two of the most known Web rule languages. In this chapter, our focus is more on R2ML, but we briefly skim over RuleML before describing the architecture and of R2ML.

RuleML represents an initiative for creating a general rule markup language that will support different types of rules and different semantics (Boley, Tabet, & Wagner, 2001). It is conceptualized to capture the hierarchy of rule types as we discussed earlier. However, the current version of RuleML covers only some limited forms of rules (Wagner, Damasio, & Antoniou, 2005).

RuleML is built on top of logic programming paradigm of first order logic (i.e. predicate logic). In the tradition of logic programming which is also followed by RuleML, research is focused on computable interpretations of predicate logic, by exploring a great number of semantic extensions and variations. OWL (as well as SWRL (Horrocks, Patel-Schneider, Boley, Tabet, Grosof, & Dean, 2004)) stems from logic-based tradition

Figure 1. An example of a RuleML rule showing the "hasUncle" relationship

```
<Implies>
  <head>
    <Atom>
      <Rel>hasMother</Rel>
      <Var>x1</Var>
      <Var>x2</Var>
    </Atom>
    <Atom>
      <Rel>hasBrother</Rel>
      <Var>x2</Var>
      <Var>x3</Var>
    </Atom>
  </head>
  <body>
    <Atom>
      <Rel>hasUncle</Rel>
      <Var>x1</Var>
      <Var>x2</Var>
    </Atom>
  </body>
</Implies>
```

of artificial intelligence where research is based on classical predicate logic (two-valued) as the one and the only logic. An example of a RuleML rule that uses certain person's attributes to define a rule of "hasMother" and "hasBrother" implies "hasUncle" is shown in Figure 1.

In the next subsection, we skim over the second key web rule language proposal for RIF, known as REWERSE Rule Markup Language (R2ML).

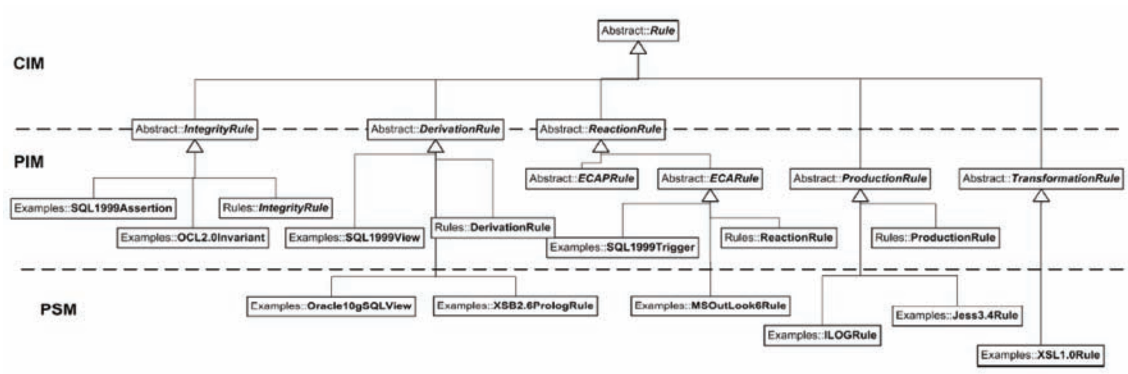
REWERSE Rule Markup Language (R2ML)

R2ML is part of the EU-funded REWERSE project, which follows the goal of providing a general rule markup language to make the deployment, execution, publishing, and communication of rules on the Web possible. The approach chosen to develop the R2ML is based on the known Model Driven Architecture (MDA) (Miller & Mukerji, 2003) and Meta Object Facility (MOF) (ODM, 2001) defined by the Object Management Group (OMG). This means that the whole language definition of R2ML can be represented by using UML diagrams as MOF uses UML's graphical notation. The language also has an XML concrete syntax defined by an XML schema.

As shown in Figure 2, R2ML considers three levels of abstraction for rules. At the *computation-independent business domain* level (called CIM in OMG's MDA), rules are statements that express (certain parts of) a business/domain policy (e.g., defining terms of the domain language or defining/constraining domain operations) in a declarative manner, typically using a natural or visual language. At the *platform-independent operational design* level (called PIM in OMG's MDA), rules are formal statements, expressed in some formalism or computational paradigm, which can be directly mapped to executable statements of a software system. At the platform-specific implementation level (called PSM in the OMG's MDA), rules are statements in a language of a specific execution environment such as XSB Prolog (XSB, 2007).

R2ML provides a vocabulary that enables users to define their own world in the form of objects and elements available in the domain of discourse.

Figure 2. Different abstraction levels considered for rules in R2ML



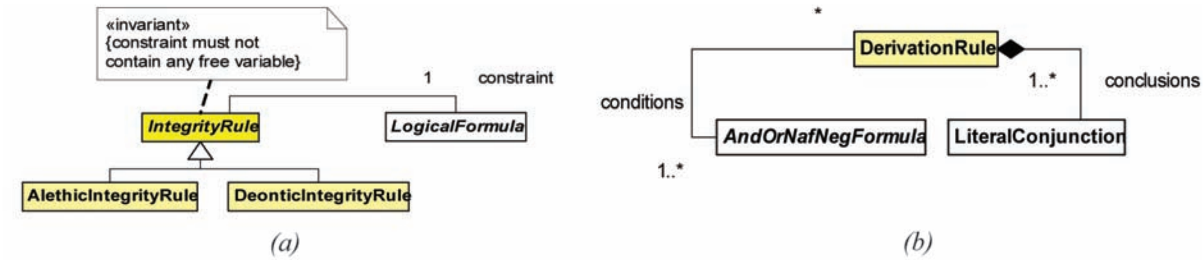
The vocabulary can be defined as a combination of *Basic Content Vocabulary*, *Relational Content Vocabulary*, and *Functional Content Vocabulary.* Basic Content Vocabulary allows the user to specify the basic elements of the domain such as individual objects and data values, classes and data types, and object and data variables. Relational Content Vocabulary helps to associate different objects from different classes through defining n-ary association and association classes. Finally, Functional Content Vocabulary assists with defining functors that correspond to the standard logic of functions. The functions can be data operations to manipulate data values, they can be object operation functions that define object-value operations, or they can be role functions which correspond to functional association (binary association) of the class elements. In (Bradshaw, Dutfield, Benoit, & Woolley, 1997), authors showed how the basic constructs and elements of the OWL language can be transferred and modeled by R2ML atoms and elements. For example, *sameAs* in OWL is equivalent to an *EqualityAtom* in R2ML and *oneOf* in OWL carries the same meaning as *Disjunction* of a set of atoms in R2ML. This means any language with its concepts defined based on OWL (including KAoS and Rei) can be modeled with R2ML constructs elaborately.

Having the objects and concepts of a domain defined, R2ML makes the definition and harmonization of rules over these concepts possible through the use of four different types of rules: *Integrity Rules*, *Derivation Rules*, *Reaction Rules*, and *Production Rules*. Since in this paper we are limited in space, we only review the first two rules and more information about the other rules and constructs of the language can be found in (Wagner, Giurca, & Lukichev, R2ML: A General Approach for Marking-up Rules, 2005).

R2ML integrity rules, also known as (integrity) constraints, consist of a constraint assertion, which is a sentence in a logical language such as first-order predicate logic or OCL (see Figure 3a). R2ML supports two kinds of integrity rules: the *alethic* and the *deontic* ones. The alethic integrity rule can be expressed by a phrase, such as *"it is necessarily the case that"* and the deontic one can be expressed by phrases, such as *"it is obligatory that"* or *"it should be the case that"*. A LogicalStatement is a LogicalFormula that has no free variables, i.e., all the variables from this formula are quantified. In terms of policy languages, integrity rules can be considered as constraints that must hold consistently especially in the level of rule enforcement, e.g. *"it is necessary to give a higher priority to the commands of the administrator than to the commands of the regular users on a system."*

An R2ML derivation rule has conditions and a conclusion (see Figure 3b) with the ordinary mean-

Figure 3. The metamodel for (a) the integrity rule, and (b) the derivation rule in R2ML



ing that the conclusion can be derived whenever the conditions hold. While the conditions of a derivation rule are instances of the *AndOrNafNegFormula* class, representing quantifier-free logical formulas with conjunction, disjunction and negation; conclusions are restricted to quantifier-free disjunctive normal forms without *NAF* (Negation as Failure, i.e. weak negation). In the context of policies, we consider each deontic policy rule as a single derivation rule with the constraints making the conditions of the derivation rule and the policy decision forming the conclusion of the rule, e.g. *"If the user is from Simon Fraser University with a valid student ID then give her the permission to enter the area of the university."* It may sound more expressive to define deontic policy rules with deontic integrity rules in R2ML. However, our attempts in doing so showed that deontic rules in the context of policies carry a different meaning from their interpretation in R2ML. In R2ML, a deontic integrity rule represents a constraint that should be satisfied or must hold with a concrete proof for its truthfulness, though a doentic policy demonstrates concerns over performing a duty or obligation as a result of satisfying a series of related conditions.

Atoms are the basic logical constituents of a rule which are compatible with the concepts of OWL, RuleML, and SWRL. Atoms connect objects to values, classes to instances, and objects to objects, put restrictions on the objects and data values, and so on. Here, we briefly represent some of the atoms that are relevant to our purpose of representing policy languages. *ReferencePropertyAtoms* (see Figure 4) associate object terms as subjects with other terms (objects or data values) as objects. A *ReferencePropertyAtom* in R2ML corresponds to an OWL (and similarly a KAoS) object property, or to the OWL concept of value for an individual-valued property. *ObjectDescriptionAtoms* (see Figure 5) are another class of useful atoms for our purpose. They refer to a class as a base type and to zero or more classes as categories, and consist of a number of property/term pairs (i.e., attribute data term pairs and reference property object term pairs). Any instance of

Figure 4. R2ML's ReferencePropertyAtom

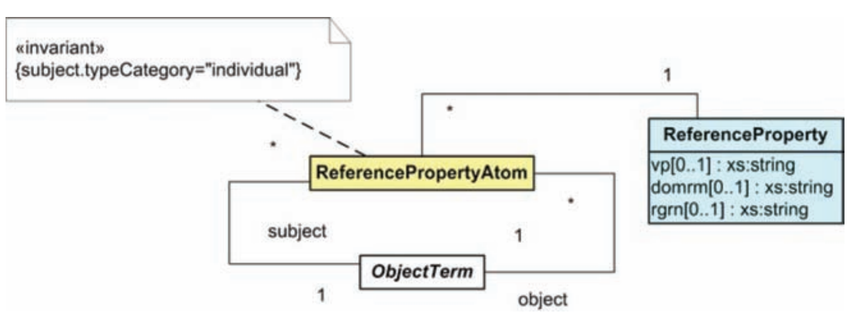
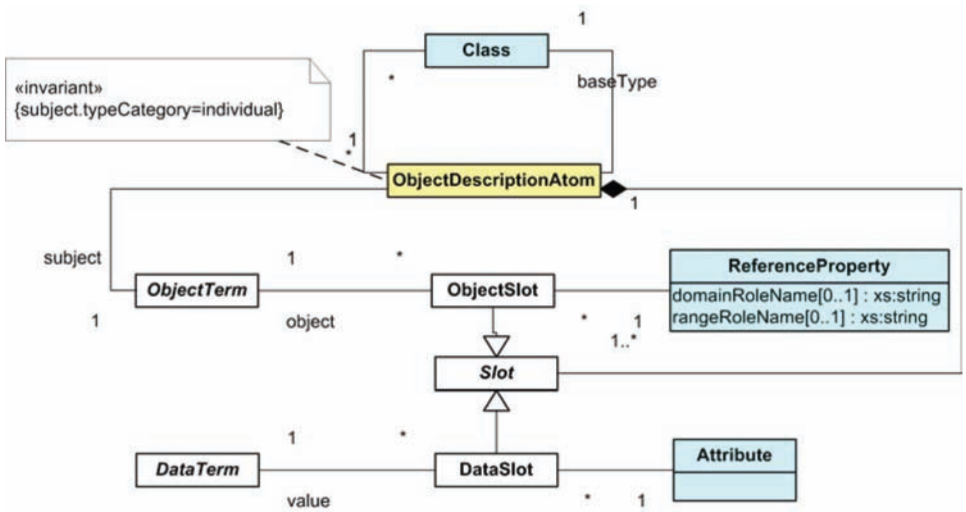


Figure 5. R2ML's ObjectDescriptionAtom



such atom refers to one particular object that is referenced by an *objectID*, if it is not anonymous. This atom corresponds to the instantiation of an object from a class in OWL.

Policy Languages

Policies can be considered as building blocks in enhancing the security, privacy, and usability of a system (Bonatti, et al., 2006). For policies to work effectively, they require to be defined over the domain knowledge that is provided by the domain experts. The process of representing the domain knowledge and defining policy rules over this knowledge, in a machine readable form, is handled by using policy languages. A policy language enables combination of constraints and capabilities, implying rules, over resources that represent policies.

Thus far, there have been various designed policy languages (such as XACML (Godik & Moses, 2003), Ponder (Damianou, Dulay, Lupu, & Sloman, 2001), Protune (Bonatti & Olmedilla, 2005), PeerTrust (Nejdl, Olmedilla, & Winslett, 2004), KAoS (Uszok, et al., 2003), and Rei (Kagal, 2002)) aiming to enable policy engineers to represent their organizational policies in a machine understandable format. In this chapter, we focus on KAoS policy language mainly because *i) it follows a markup format for defining policies*, and *ii) it uses ontologies to represent the domain knowledge*. The first property helps with easier mapping of policies from the KAoS policy languages to a Web rule language, because the source and the target language more or less follow a similar syntactical format. The second property helps with easier transformation of the domain knowledge from the source language to the target language. There have been already efforts on providing accurate mappings between two domain ontologies (Kalfoglou & Schorlemmer, 2003). Such efforts assist with the conversion of the domain knowledge from one conceptual model to a different conceptual model, thus making the domain knowledge easy to be shared across various policy languages and their corresponding reasoning engines.

Rei (Kagal, 2002) and KAoS (Uszok, et al., 2003) are two semantically enriched Web policy languages that use Semantic Web ontologies to define the resources, the behavior, and the users of a domain. Using ontologies enables these two languages to easily adjust themselves to the target system regardless of the number of resources and users in act. KAoS describes the entities and concepts of its world using OWL, while Rei can understand and reason over a domain of concepts defined in either RDF or OWL.

In terms of available policy rules, both KAoS and Rei have four main types. *Permission*, *Prohibition*, *Obligation*, and *Dispensation* in Rei are respectively equivalent to *PosAuthorizationPolicy*, *NegAuthorizationPolicy*, *PosObligationPolicy*, and *NegObligationPolicy* in KAoS. The defined policy rules in each of the languages are then sent to a reasoner that performs the process of conflict resolution and decision making for the rules that match the current state of the world. This task is done by using Stanford's Java Theorem Prover (JTP) in KAoS and a Prolog engine in Rei version 1.0. Rei version 2.0 has extended its reasoning engine to use F-OWL, an ontology inference engine for OWL, based on Flora and XSB Prolog (XSB, 2007). Although these two policy languages have a lot in common there are dissimilarities between them as well. The main difference between KAoS and Rei is the underlying formalism of the languages. KAoS follows description logic coded in the form of OWL expressions with additional constraints that help with defining policy elements and rules. On the other hand, Rei uses its own language that defines policy rules in terms of Prolog predicates expressed as RDF triples. This way Rei follows semantics close to Prolog's semantics which is itself built on top of the concepts of logic programs.

The process of rule enforcement in KAoS is done by extending its enforcement engine depending on the domain it is going to be used in. In Rei, however, there is no rule enforcement engine. Yet, due to the deterministic properties of declarative logic, reasoning over dynamically determined values in Rei policies is more accurate than KAoS in which chances of dealing with unknown situations are likely to happen. In order for processes and services to communicate remotely, Rei relies on a rich set of Speech Acts. In Rei, Speech Acts are used by a sender to express the request for performing one of the actions: *Delegation*, *Revocation*, *Request*, *Cancel*, *Command*, and *Promise* by the receiver. Conversely, in KAoS the remote communication procedure is done through the message passing of the underlying platform.

Defining KAoS policies as OWL expressions gives the language more flexibility to maneuver over the concepts of the world. Different quantifying expressions, inheritance relationships, cardinality restrictions, etc. can be explicitly expressed in KAoS thanks to the constructs of OWL. It also enables KAoS to perform static conflict resolution and policy disclosure. KAoS has its classes and properties already defined in OWL ontologies, referred to as KAoS Policy Ontologies (KPO) (Uszok, et al., 2003), which are accessible from (Uszok & Bradshaw).

Looking back to the similarities and differences discussed, providing meaningful transformations from/to a web rule language to/from a policy language is not a trivial goal. Aside from all the difference in the syntax of the source and the target language, the transformations should care about the underlying logic that each policy language adheres to. As we discussed above, Rei and KAoS follow different logical formalisms, and thus different mapping considerations require to be done in choosing the appropriate elements in the target language to model the elements of the source language. This helps with capturing the logical semantics of the transformed rules. Furthermore, the concepts that might be missed in the procedure of transformation, due to a logical or conceptual mismatch between the policy language and the web rule language, need to be considered (we discuss it further in the next section). Web rule languages have been design with a broad consideration on the supported logic and required logical elements such that they can capture all various types of rules. However, the mismatch in the logics or the concepts may happen when transferring a rule from a web rule language to a policy language. The concepts that result in a logical or conceptual mismatch must be carefully monitored during the process of transformation. Having the lost information during the process of transformation found, the importance of such elements and the harms and threats that may happen to the resources due to information loss might be taken into consideration.

Relations between Policies, Rules, and First Order Logic (FOL)

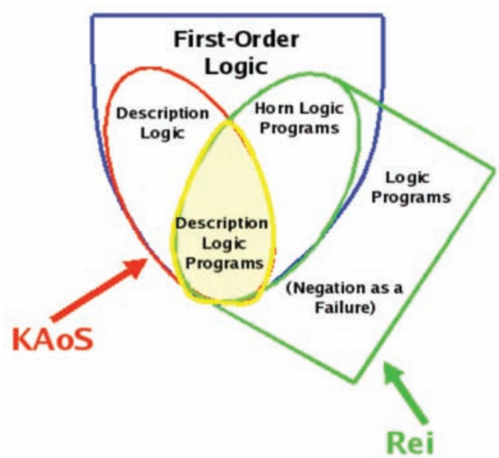
As already mentioned, rules are among the most frequent techniques for representing knowledge. Different knowledge representation methods (including rules in a broader and policies in a narrower sense) typically follow different logical formalisms. A clear use of Web rule languages in place or in conjunction with policies requires a clear understanding of the logical formalisms behind defining rules and policies. There are two major fragments to FOL that are mainly used for representing knowledge, namely, *description logic (DL)* and *computational logic programs (LP)*. Description Logic is a subset of the well-known First Order Logic (FOL) without function symbols (Grosof, Horrocks, Volz, & Decker, 2003), which is similar to a fragment of Horn FOL, *def-Horn*, that also contains no function symbols. LP, however, represents a set of logical constructs that neither includes nor is fully included by FOL but only intersects with it. That is to say, although LP and FOL share some logical constructs, there are concepts in either of the two sets that are missing or do not have any exact equivalent in the other set. For example, FOL can express positive disjunctions (see Grosof, Horrocks, Volz, & Decker, 2003, for details), which are inexpressible in LP. Additionally, various important features and logical elements of LP, which are frequently used in practical rule-based applications, are inexpressible in FOL (e.g. Negation-as-Failure) (Grosof, Horrocks, Volz, & Decker, 2003).

As we mentioned earlier, transformation of rules (or policies) from one language to another requires an accurate syntactic and semantic mapping of the concepts of the source language to the concepts of the target language. Not only does this require a syntactic analysis of the two languages, but also requires capturing the semantics of the rules (or policies), the representation of the domain knowledge, and also the underlying logic of each rule (or policy) language. We have already described that policy languages may come with different logical formalisms, for example KAoS is based on DL while Rei is based on LP. The set of logical constructs for web rule languages is rich enough to capture the concepts of a policy language regardless of whether it follows DL or LP. In such cases, the obtained web rule would admit to the same logical formalisms that the source policy language follows. Nonetheless, when transforming rules from a web rule language to a policy language, it becomes very important to identify the differences in the underlying logic of the source web rule language and the target policy language.

When transforming rules from a DL-based language to an LP-based one, there is no possibility to cover all the concepts. Yet, most of the concepts of these languages are transformable from one to another due to the similarities between DL and def-Horn[1] shown as the intersection of these two sets in Figure 6. Rule transformations, as the major objective followed by RIF and Web rule languages, are not only concerned about a syntactic representation of the elements in a rule language, but they also care about the semantics of the rules. Rule languages can capture the semantics of a domain, first by modeling the domain knowledge using the concepts of the rule language and then by benefiting from the underlying constructs that carry different logical semantics ranging from descriptive elements to declarative ones.

Grosof, Horrocks, Volz, and Decker (2003) provide an elaborate method of mapping the basic elements of description logic to declarative logic programs. OWL as a subset of RDFS corresponds to a fragment of classical FOL. It is shown by Grosof, Horrocks, Volz, and Decker (2003) that

Figure 6. Expressive overlap of description logic and logic programs and where Rei and KAoS sit in this classification



OWL elements are convertible to definite Horn FOL elements which in turn are convertible to definite Datalog Logic Programs as a restricted model of Logic Programs (LPs). For example, classes and class expressions are equivalent to FOL formulae with one free variable, and properties (and property expressions when supported by description logic) are equivalent to FOL formulae with two free variables. Classes and property inclusion axioms are also considered as FOL sentences consisting of an implication between two formulae with the free variables universally quantified at the outer level. Table 1 shows a selection of OWL constructs with their corresponding description logic syntax and FOL expressions. Details of the definitions can be found in (Grosof, Horrocks, Volz, & Decker, 2003).

Model Driven Engineering

The main goal of MDE is to switch the focus from low-level implementation details to problem-specific concepts (Schmidt, 2006). The core activity is then to define languages for particular problem domains (*in our case policies*). Metamodeling is an approach used in MDE for defining languages. A metamodel is a model of a modeling language, that is, a metamodel defines a set of sentences that can be expressed in a modeling language (Kühne, 2006). As such, a metamodel can be regarded as an abstract syntax of a language. Since the goal of MDE is also to provide a common framework for processing different modeling languages, it also defines a technical context in which different modeling languages are defined and used, and on top of which different modeling tools are built.

Metamodeling architectures are used as a solution to this issue. Usually, a metamodeling architecture (e.g., OMG's Model-Driven Architecture) is organized in a layered fashion, where typically there are four layers. The top most layer is called metametamodel (and tagged with M3 or

Table 1. Some of the OWL constructors and the equivalent description logic and FOL expressions

OWL Constructor	DL Syntax	FOL Expressions
subClassOf	$C \subseteq D$	$D \leftarrow C$
transitiveProperty	$P^+ \subseteq P$	$\forall\ x, y, z(P(x, y) \wedge (P(y, z)) \rightarrow P(x, z)$
inverseOf	$P \equiv Q^-$	$\forall\ x, yP(x, y) \Leftrightarrow Q(y, x)$
intersectionOf	$C_1 \cap ... \cap C_n$	$C_1(x) \wedge ... \wedge C_n(x)$
unionOf	$C_1 \cup ... \cup C_n$	$C_1(x) \vee ... \vee C_2(x)$
complementOf	$\neg C$	$\neg C(x)$
one of	$(a_1,...,a_n)$	$x = a_1 \vee ... \vee x = a_n$
hasClass	$\exists$P.C	$\exists y(P(x, y) \wedge C(y))$
toCass	$\forall$P.C	$\forall y(P(x, y) \rightarrow C(y))$

L3), and on this layer a metamodeling languages is defined. Most commonly, this language is defined by itself. In a metamodeling architecture, there is typically one and only one metamodeling language defined. Examples are – OMG's standard Meta-Object Facility (OMG), 2008) and Eclipse's Ecore (Budinsky, Brodsky, & Merks, 2003). While the abstract syntax of two languages is slightly different (i.e., it is a reduced part of UML elements related to class models), the graphical concrete syntax of both languages is borrowed from UML class models. The rationale for having only one M3 language is to have a common space for defining modeling languages on the M2 layer. Thus, various modeling languages can be processed in the same way using the same tool set (e.g., APIs, constraint, query, and transformation languages).

Model transformations are also an intrinsic part of MDE, as the idea is to transform M1 models to different platforms or to allow translation for one type of model (e.g., UML) to another (e.g., ODM). MOF Query/View/Transformation (QVT) is the OMG's official standard for model transformations (OMG, 2005).

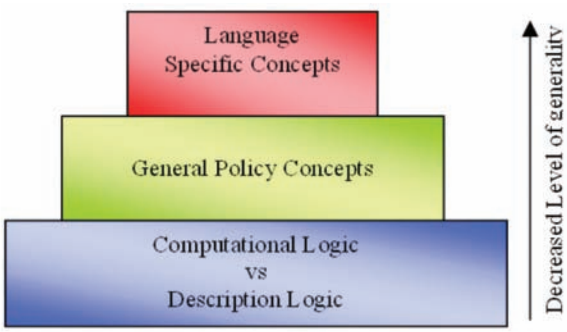
POLICY INTERCHANGE FRAMEWORK

Interchanging the rules in general, and the policies in particular, between different business enterprises is a goal followed by RIF working group (Ginsberg, Hirtle, McCabe, & Patranjan, 2006). We already discussed it in the introduction section why interchanging the policies between different business partners is important to achieve. Nonetheless, the variety of policy languages that have been developed so far, the lack of a standard for defining policies, and the limited number of experts in each of the existing policy languages, made us think about designing a policy interchange framework that can easily expand to cope with different policy languages, especially the ones that have not been developed yet. It is worth noting that while interchanging the policies, we deal with their logic, their abstract syntax, their concrete syntax, the domain in which these policies are deployed, and also the semantics that each term or concept in each language or domain carries. Our studies led us to a point where we realized that, in order to have such a framework, we need to start from the logic that the language follows, to the abstract syntax that it represents, having the semantics of its domain concepts in mind.

One way to represent the domain knowledge is to provide an ontological representation of the domain and form the policy rules around these concepts. The ontological representation of domain concepts brings in all the benefits of Semantic Web into the implementation of policy rules and enables agents from different domains to have shared and common understanding of the concepts of each domain. Consequently, the problem of exchanging policies from one domain to another is divided to two sub-problems of i) transferring the concepts of a domain to the concepts of another domain, and ii) capturing and transferring the logic and the semantics of the language from the source language to the target language. The first sub-problem has already undertaken a lot of research in the field of ontology mapping (Kalfoglou & Schorlemmer, 2003). In this section, we basically illustrate how the second sub-problem can be addressed using web rule languages and this is the second sub-problem that the community of RIF is targeting.

The concrete syntax of the languages is probably the least challenging issue to deal with, as there are powerful transformation tools and languages, such as QVT/ATL (OMG, 2005) and XSLT (Clark, 1999), that can mine through the concrete definition of the languages and extract the appropriate concepts. For the abstract syntax of the language, we normally encounter the general concepts that are shared between the languages with similar purposes, and the language-specific concepts that are not shared between all the languages of one single category. The language-

Figure 7. The layered policy interchange framework architecture



specific concepts take more important roles in languages that are in a lower level of abstraction. Policy languages, due to their characteristics in addressing low level domain specific concepts and the differences in the domains that they need to be deployed in, may have various dissimilar concepts that are specific for each language. A powerful policy interchange framework should be able to clearly distinguish between the logic, the language-independent, and the language-specific concepts and try to have general definitions for as many of these concepts as possible.

Figure 7 shows our proposed architecture for a policy interchange framework. As the figure shows, the policy framework, first starts with identifying the similarities and dissimilarities between the underlying logics of the policy languages. The two main logics that are widely used in defining rules and representing the knowledge of a domain are computational logic and descriptive logic and there has been already research on how to map the concepts of these two logics (Grosof, Horrocks, Volz, & Decker, 2003) (Kolovski, Hendler, & Parsia, 2007).

The next level in the architecture is to identify the general concepts that are common across policy languages. Once these concepts are recognized, the rules from one source language to one target language can be mapped to some degree, having the domain knowledge plus the language vocabulary identified and mapped. For the language-specific concepts of a source policy language, either it is possible to define the concepts through using a series of concepts from the target language, or simply the concepts are not convertible. Depending on the importance of the meanings that the unmappable concepts carry, the transformation to the target policy language would be successful or unsuccessful. As Figure 7 shows, by moving to the upper levels of the architecture, the generality of the mappings is significantly decreased, such that reusing it for other policy languages becomes impossible in the top most level. Nonetheless, it is possible for the concepts in the lower levels to be shared between different policy languages. To plug a new policy language to the framework, first and foremost, the underlying logic of the policy language should be identified and mapping rules to cover the corresponding concepts should be developed. Moving to the upper level in the architecture, the common policy concepts are identified and mapped to the concepts of the general policy model used. For the top most level, a detailed review of the source policy language by its experts is required to work around the concepts that are not mappable.

In the following sections of this chapter, we try to extract the abstract syntax of KAoS by providing its metamodel, extract the language-independent and language specific concepts that the KAoS policy language covers, and argue how the mapping between this policy language and R2ML can happen without facing serious information loss.

METAMODELING FOR POLICIES AND RULE LANGUAGES

Concept modeling is considered a critical step in better understanding and comprehending the constituents of a system in general, and a language in particular. One of our major intentions in providing the metamodel for KAoS is to hide the low-level details of the language in order to focus on its conceptual characteristics. In this section, we present the UML metamodel for KAoS and contrast it against the common concepts in R2ML that can carry the equivalent meanings for those concepts.

As we mentioned earlier, KAoS exploits the use of Web Ontology Language (OWL) to define its concepts. Thus, providing a meta-model for the language can be done through transforming the OWL definition of the languages to a UML metamodel. There have been several efforts in defining standard mappings from ontology languages to UML. In particular, the Ontology Definition Metamodel (ODM) (ODM, 2003) initiative follows the goal of using OMG's Meta Object Facility (MOF) as a metamodeling language to develop a linguistic metamodel to represent the ontology languages. Although ODM, itself, is still undergoing modifications, we chose to use the current state of ODM to represent Rei and KAoS. However, at some points, based on the needs and the real meaning of the concepts in each language, we had to slightly modify the ODM definition for getting a better reflection of the concepts through using UML constructs.

The Metamodel for KAoS Policy Language

KAoS as a policy language has elements to define and distinguish different types of policies over different resources. Due to space limits, here we only focus on the major concepts in KAoS and show how they can be transformed to their equivalent counterparts in R2ML. Figure 8 represents the types of policies that are covered in KAoS. As Figure 8 shows, a policy in KAoS is either an *AuthorizationPolicy* or an *ObligationPolicy* with each of them further specialized to *PosAuthorizationPolicy*, *NegAuthorizationPolicy*, *PosObligationPolicy*, and *NegObligationPolicy*. It can also be a policy over the policies that conflict, *ConflictedPolicies*, which usually is not considered a policy type, but is just a subclass of the class Policy. An *AuthorizationPolicy* introduces the set of permissions and prohibitions for an actor while dealing with a context. However, KAoS is a context-based policy language that defines permissions and prohibitions not over the roles of the actors but over the context of interactions.

In KAoS, each policy element stands for exactly one policy rule. As shown in Figure 9, a KAoS policy can have a control action, a triggering action, an oblige action, and a series of conditions. It also can define a site to which the policy rule should be applied. As we discussed earlier, KAoS is a context based policy language. *This means that the constraints are defined over the actions that are going to be taken on a resource, rather than the users (or roles) that are going to take these actions.* This can be considered as a major difference between the role-based access control policies and the context-based access

Figure 8. Different types of policies in KAoS

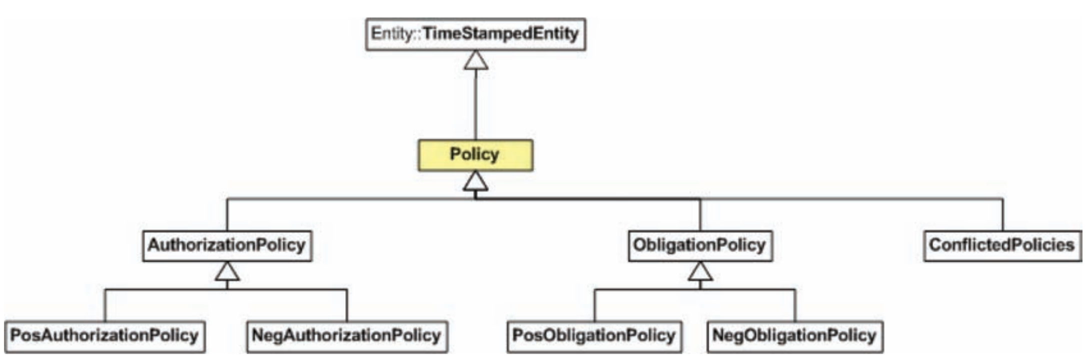
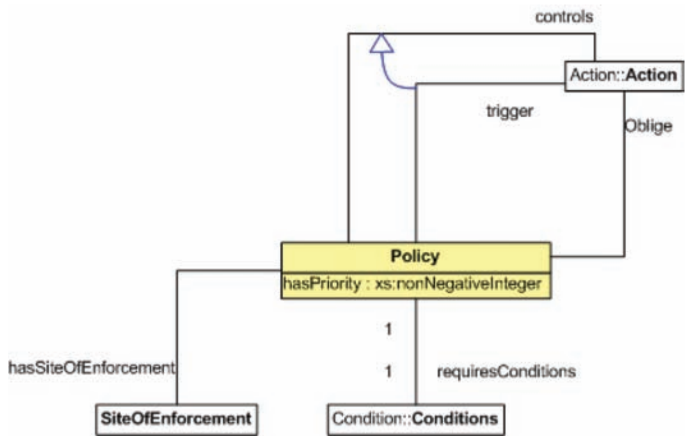


Figure 9. The policy metamodel for KAoS



control policies.

An action in KAoS, as shown in Figure 10, has a context to which the action is applied, which either is a *dataContext* of any type (i.e. of type OwlThing), or an *objectContext* of type Target, which is itself an entity. A KAoS action also has one and exactly one Actor, referred to by the *performedBy* property in the definition of the Action. Considering the fact that in context-based policy management Actions are constrained, as opposed to roles in role-based access control, a rich definition for the class Action, its properties, and its attributes helps with having accurate definitions for the constraints and conditions. Having a look at the metamodel in Figure 10, it is clear that KAoS also follows the same objective with defining its actions.

KAoS provides a rich set of actions, inherited from the class Action, that enable better classification and definition of the actions which are going to be performed. This brings more expressivity to the language, enabling a better semantic modeling of each policy. Figure 11 provides a clear classification of all sets of actions that are defined in KAoS. A user can extend any of these classes to make them cope with what she intends to express. *ObligateAction* is one of these important actions in KAoS. An *ObligateAction* makes the active actor

Figure 10. The metamodel for KAoS Action

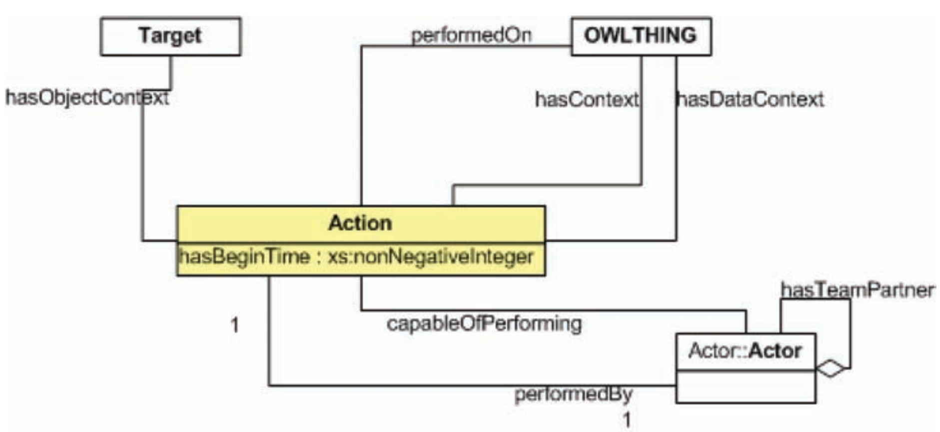


Figure 11. The metamodel for various KAoS action types

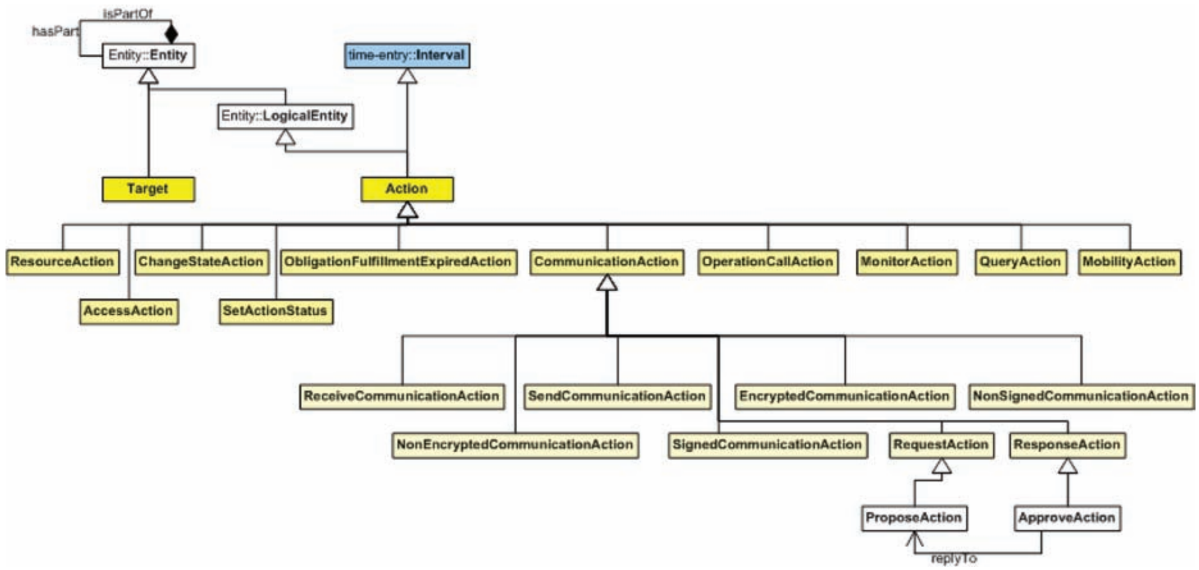
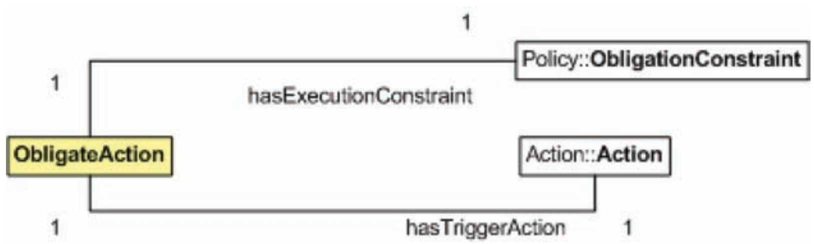


Figure 12. The metamodel for ObligateAction



to execute the *ObligateAction* once the triggering action happens and the execution constraints for the obligation action are met (see Figure 12).

In KAoS, similar to the hierarchy of actions, a detailed hierarchy of entities and actors have also been defined. Actors in KAoS, represent a complete set of possible agents that might interact with the system (see Figure 13).

Different physical and artificial actors have been defined by the system which helps with precise classification of the actors while working with policies. The only problem with the obtained model, as we have highlighted in Figure 13, is the redundant definition of both Human and Person which seem to be identical. The reason behind distinguishing between these two concepts by the developers of KAoS is not clear to the authors of this article. Furthermore, according to the ontology definition of KAoS for the concept Human, it inherits from both Person and *PhysicalActor*. Since Person itself has been already defined as a subclass of *PhysicalActor*, the inheritance from *PhysicalActor* by Human seems to be redundant.

An actor in KAoS can be controlled by a set of policies, and is capable of performing a series of actions. It can also cooperate with other actors as a team (see Figure 14).

For a policy to be able to control the behaviour of a system, conditions are required to be place on either potential roles (in case of role-based access control) or potential actions (in case of context-based access control) in a system. A condition in

Figure 13. The metamodel for class Actor in KAoS

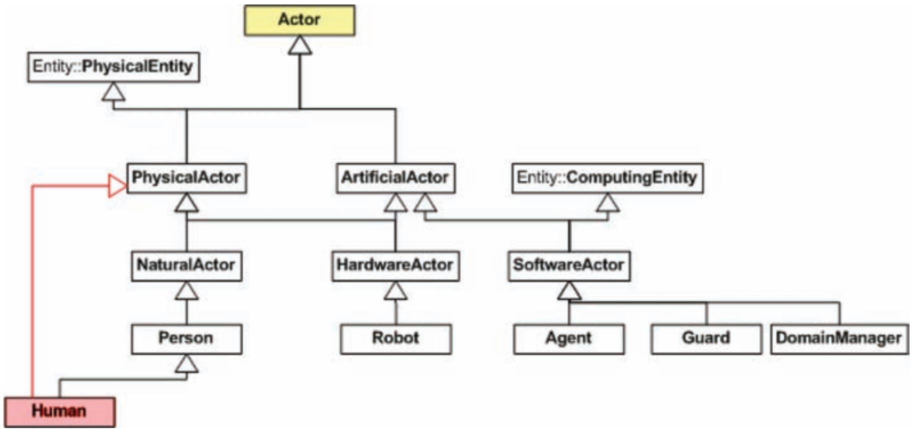
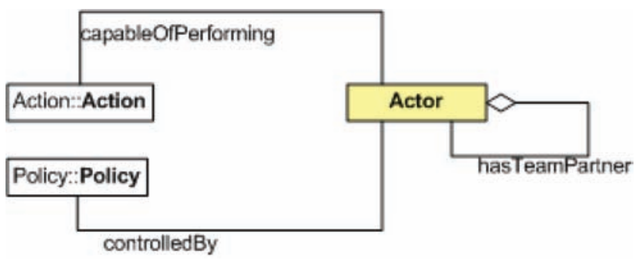


Figure 14. The metamodel for the class Actor in KAoS



KAoS represents a situation based on the current State and the history of the events (see Figure 15). As Figure 16 shows, the EventHistory keeps the time-stamped events that are mostly in the form of occurred actions, represented in the *Action-History* class. To each *ActionHistory*, relates a *hasRegisteredAction* which represents the current action occurred in each time interval, the actor of the action, and the context to which the action has been applied.

Conceptual Mapping of KAoS and R2ML

Given a rather comprehensive representation of KAoS and R2ML (both syntactically and semantically), we can take the main step in defining the mappings between these two languages, aiming at providing an exchange method between different concepts of these languages. It should be noted that although we deal with the syntactical mapping of the elements, the process of mapping considers semantic similarities as well. In this section, we show that, despite having several possibilities to map the elements of one language to the elements of another language, we choose those elements which are most similar semantically. Our mappings in this section follow the *QVT's graphical notations* (QVT, 2005) for transforming between concepts. However, we have included some dashed arrow lines to make the one-to-one mappings easier to understand.

In our introduction to R2ML, we presented the major types of rules that R2ML covers. Among all the rules, *derivation rules* seem to be better options to model the policies, mostly because they entail the meaning of: "obtaining a conclusion upon satisfaction of a set of conditions (or

Figure 15. The metamodel for conditions in KAoS

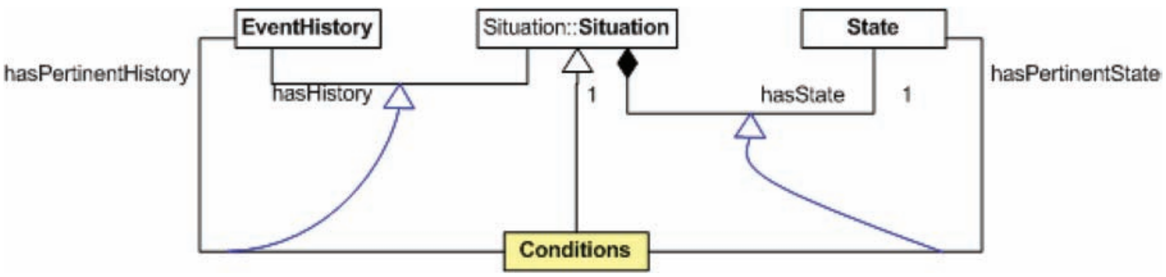
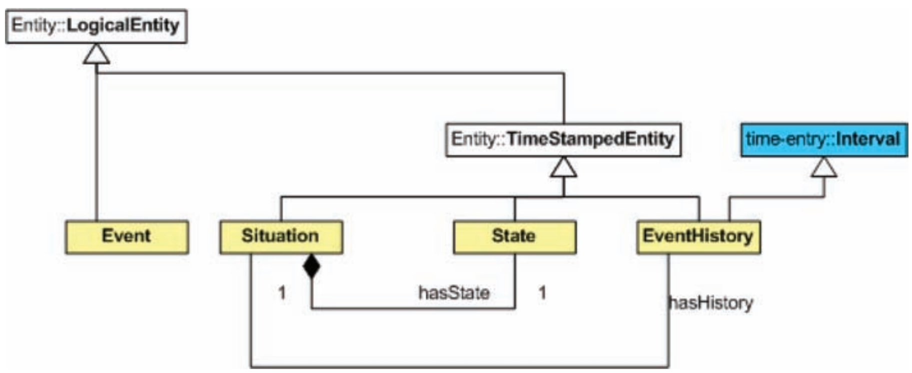


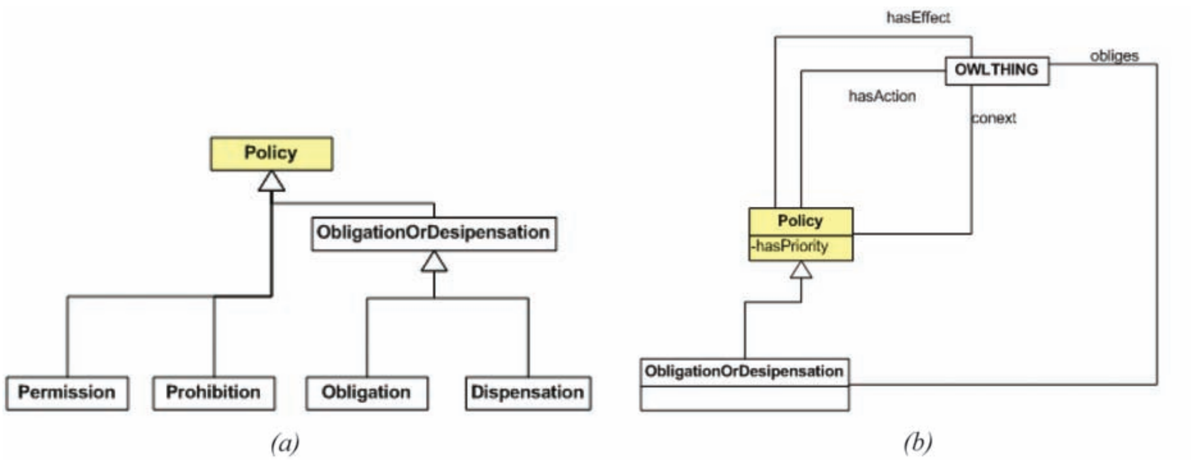
Figure 16. The metamodel for KAoS Situation



credentials)". Using *implication* in *integrity rules* makes them quite close to our intentions too, but as we already mentioned, integrity rules are generally used where the set of conditions of the rule are persistent over time. *Reaction rules* can be used to reflect the pre- and post-conditions of a policy rule, especially while communicating with Web services. Nonetheless, we believe the intention in using policies is closer to the meanings expressed by derivation rules, i.e. deriving the conclusions once the conditions are met. Thereupon, we choose to use *derivation rules* in order to model policy languages. Some other research works covering this same area are available in (Kaviani et al., 2007-1) and (Kaviani et al., 2007-2).

In our representation of different types of policies for KAoS (see Figure 8), we showed that there are generally four main types of policy rules that these languages support, i.e. *PosAuthorizationPolicy*, *NegAuthorizationPolicy*, *PosObligationPolicy*, and *NegObligationPolicy*. To be able to model these concepts in R2ML, we decided to define some general classes for these types of policies using R2ML. Due to the simplicity of the naming for the policies in Rei, we chose the same set of names to define our policy classes in R2ML. We have also mentioned that R2ML has its own mechanisms for defining vocabularies (also called R2ML vocabulary), so it is possible to exploit R2ML vocabulary to define these policy concepts. Figure 17 show the metamodels for the class *Policy* and its subclasses that we have defined in R2ML. As it can be seen in the models, these classes have simple definitions at the moment which currently seem satisfactory for our purpose. However, the metamodels for these concepts are developed in our recent work, aiming at providing a Policy Modeling Language by using R2ML (Wagner, Giurca, & Lukichev, 2005). Again it should be noted that the (Tonti,

Figure 17. The metamodels for our R2ML Policy classes



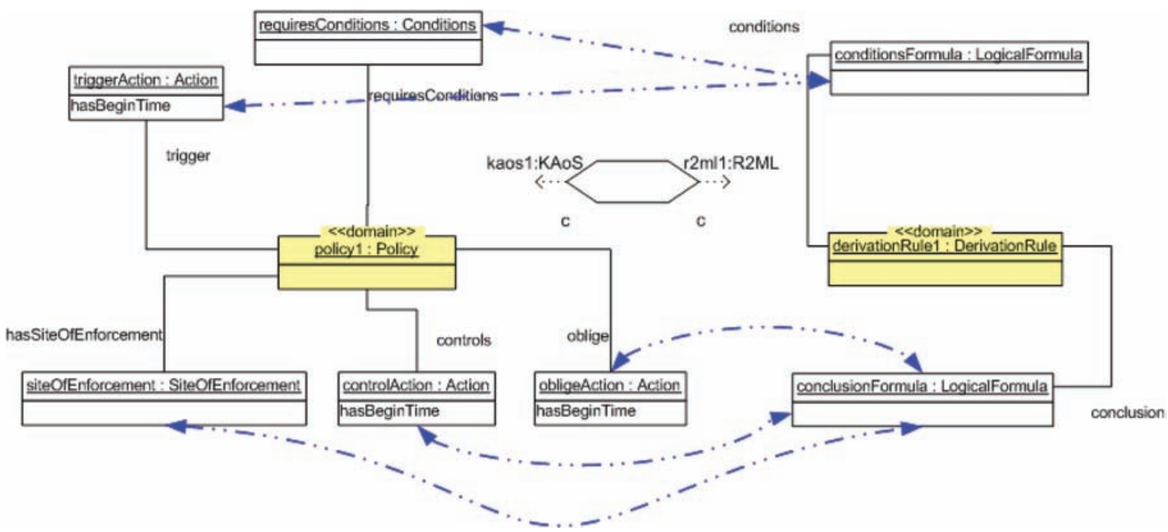
Bradshaw, Jeffers, Montanari, Suri, & Uszok, 2003) properties of these general R2ML-Policy classes do not have any range and they refer to any available resource. This is mainly to increase the flexibility of our R2ML classes to work with different actions and policies from different policy languages. We may decide to make them more restricted in the future.

While transforming each single policy rule from KAoS to R2ML, we decide about how to map these policies to R2ML policies by choosing the same naming conventions that we introduced at the beginning of this section. Next, we describe how the concepts of R2ML along with the classes that we defined above, are jointly used to provide a meaningful mapping between the concepts of the two policy languages.

We have previously mentioned that a KAoS policy is an object of the Policy class in KAoS Policy Ontology (KPO) (Tonti, Bradshaw, Jeffers, Montanari, Suri, & Uszok, 2003) with its attributes instantiated to a set of users, events, and resources that make the policy fire. Considering the KAoS policy element as a rule, the *control* element is executed upon the occurrence of the events described in the *requiresConditions* element (Figure 9). The *control* element in KAoS refers to an action. It can itself place a series of constraints on the definition of the action that can be executed by the policy. The main advantage in using KAoS is its flexibility of expansion. The *Action* class in KAoS can be easily expanded thanks to the use of OWL, to fully capture the meaning of the final action that is desired to be executed. Placing constraints on different properties of this action can also specialize the meaning that is desired to be transformed.

Once a KAoS policy rule is fired, the decision over whether or not to perform the action of the *control* element is made, an obligation of execution is placed over the *obligation* action, and the *effect* of the action is enforced to the current state of the system. This means, to model a KAoS policy with a derivation rule, we need to place the content of the *controls* element, the *oblige* element, the *hasSiteOfEnforcement* element, and the *effect* element (if there is any) in the conclusion part of the derivation rule. Whatever else that can lead to making such a decision should be placed in the condition part of a derivation rule. This includes the content of the *requiresConditions* element, the *triggerAction*, and also the variables that initialize the values for the elements in the conclusion of a rule (see Figure 18). Figure 18 shows the transformation of a KAoS policy rule to a derivation rule in its most abstract model.

Figure 18. Mapping rule of a KAoS Policy to a R2ML derivation rule



The main issue is to decide how and using what R2ML elements this process of transformation can be best achieved. The conclusion of a derivation rule should be only one R2ML *Atom*, now the question is how to manage all the actions in one single *Atom*.

The first step in transforming a KAoS policy rule to a R2ML derivation rule is to transform those elements of the KAoS *Policy* class that are placed in the conclusion part of our R2ML rule, represented with a policy *objectDescriptionAtom*. An instantiation of the derived R2ML policy class will be placed as a conclusion in the derivation rule. Figure 19 shows how this mapping can happen. The *ObjectDescriptionAtom* that captures the meaning of a policy element has four slots. One is the *hasPriority* element which if of type integer and represents a numeric value for the priority of the rule. The other three slots represent object elements that show the action that the policy controls (i.e. *hasAction*), the context to which the policy is applied, and the optional action that the policy may oblige the actor to do.

As earlier argued, an R2ML ObjectDescriptionAtom describes a set of properties and attributes, referred to as *slots*, for an object called the *subject*. As for the policy rules, the conclusion of a policy rule is a policy element (or object) with the values for the appropriate action to be taken, the actor to which giving the permission is allowed or denied, the context to which the policy is applied, and the priority of the policy rule. As a result, an *ObjectDescriptionAtom*, with a policy object as its subject is the best element to model the conclusion of a derivation rule. Figure 20 shows how we instantiate a policy object from our *R2ML Policy* class. It illustrates how the policy object is mapped to the subject of the R2ML *ObjectDescriptionAtom*, the corresponding class for this atom is set to the *Policy* class and different attributes and properties for a policy object are mapped to corresponding *ObjectTerms* or *DataTerms* depending on whether the attribute is a *DataAttribute* or *ObjectAttribute*. It has also been shown in Figure 20 that an *ObjectDescriptionAtom* can have several slots to cover various of data or object attributes for a policy. It makes it the appropriate R2ML element, both syntactically and semantically, to be used for the purpose of transforming between the objects of the policy rules. It should be taken into consideration that, in Figure 20, we are just showing the transforma-

Figure 19. Mapping a KAoS Policy class to an R2ML policy model

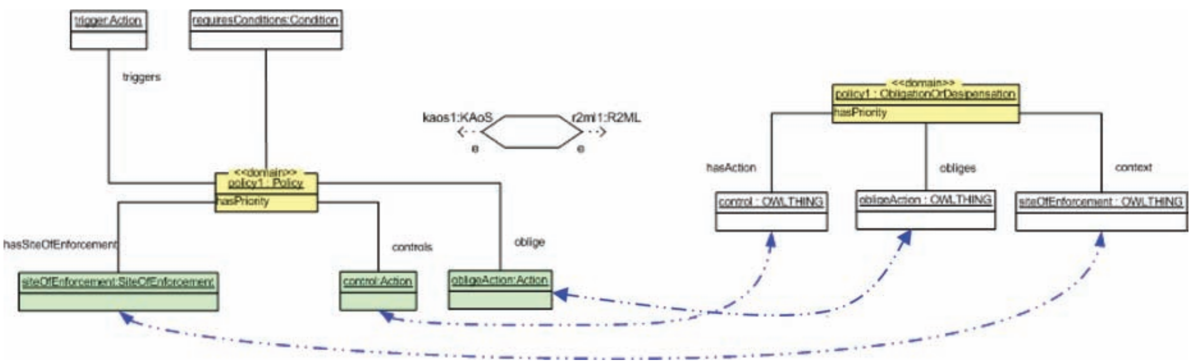
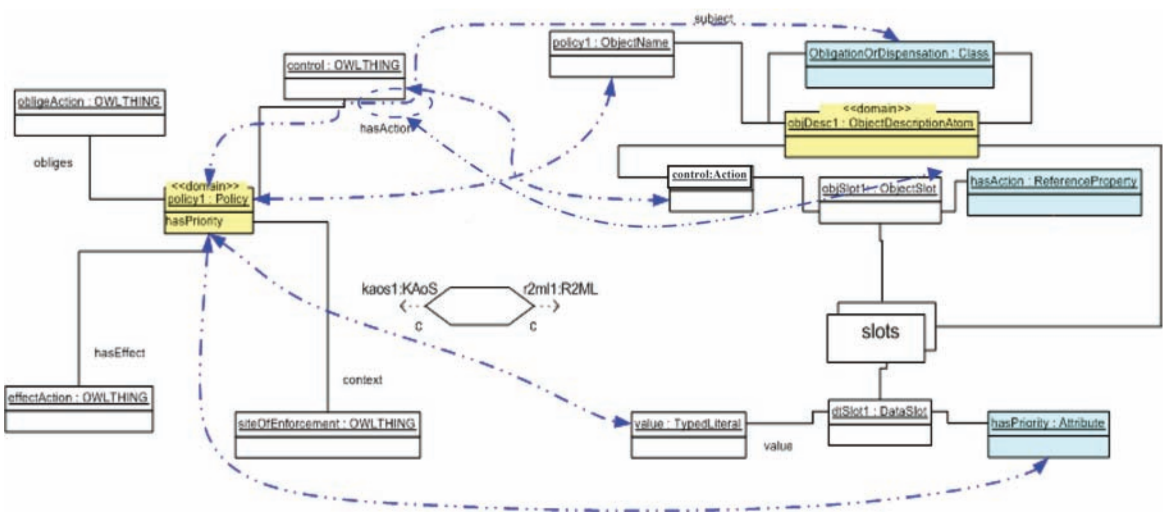


Figure 20. Mapping a KAoS Policy class to R2ML ObjectDescriptionAtom



tion of the control element to a slot in the policy *ObjectDescriptionAtom* with the property name *hasAction* represented as a *ReferenceProperty* and the corresponding object to this property (*control* in Figure 20) placed in the body of the slot.

However, referring to Figure 8, one should note that the class Policy is an *abstract* class, and so is our R2ML Policy definition of Figure 17. Thus, in Figure 20, the class Policy is usually replaced with one of its concrete sub-classes, i.e. *Permission*, *Prohibition*, *Obligation*, or *Dispensation*

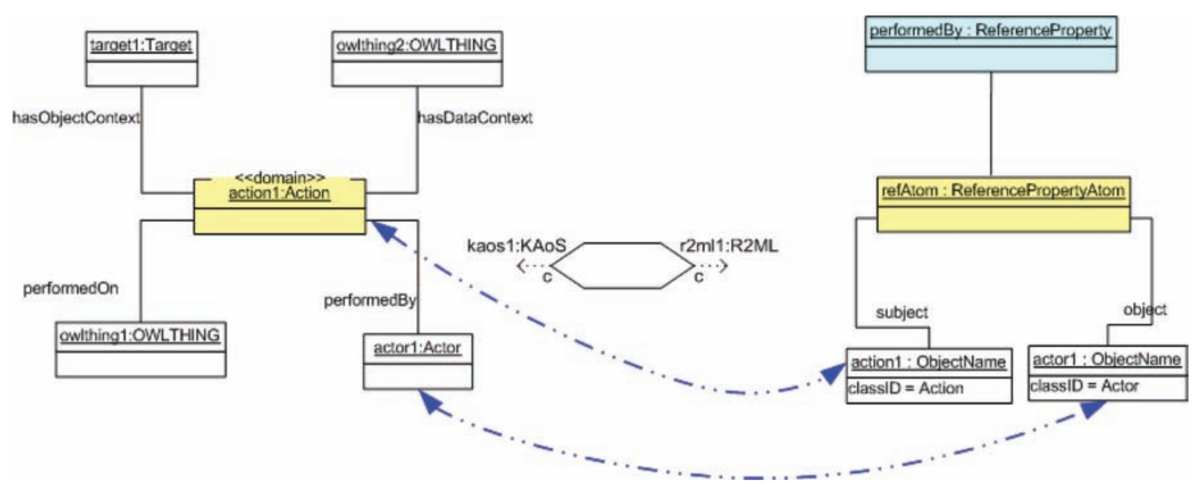
Having the conclusion of our derivation rules constructed, we need to extract its condition part from KAoS policy rules as well. The condition part of a policy derivation rule contains the definition of the variables, actions, actors, conditions, and any other construct that result in deriving the final policy object. Although the conditions of a policy rule could be modeled with *ObjectDescriptionAtoms*, we chose *ReferencePropertyAtoms* mainly to be compliant with other R2ML transformations (e.g., transformations between F-Logic and R2ML also have *ReferencePropertyAtom* in the condition part). It simplifies the later conversions of the policies to other rule languages for which we have R2ML transformations already defined (e.g., Rei, F-Logic, and RuleML). On the other hand, a *ReferencePropertyAtom* triple models a binary predicate. A set of *ReferencePropertyAtoms* with the same subject element can always be combined and converted to any element of higher arity (e.g. *ObjectDescriptionAtom*), and thus using *ReferencePropertyAtom* does not contradict the use of *ObjectDescriptionAtom*. Furthermore, in our case, *ReferencePropertyAtoms* carry even a better semantic meaning for the transformations. Semantically, they are equivalent to an OWL object property, and as KAoS is semantically very close to OWL, they model object properties of KAoS too. Figure 21 shows the conversion of a KAoS property (here, the *performedBy* property) to a *ReferencePropertyAtom* in R2ML. The converted element is then placed in the *conditions* part of a derivation rule.

In some cases, it happens that all the objects are not named-objects, but rather, they refer to *ObjectVariables*. In KAoS, due to the impossibility of representing variables in OWL, a *role-value-map* method is used in which a class representing the set of possible elements for a slot replaces the object variable in order to show all the possible options an element can take. The transformation procedure should recognize these classes and convert them to appropriate R2ML elements. This happens in two steps during our mappings. Once the role-value-mapped class is determined, it is first mapped to an *ObjectClassificationAtom* with a variable name assigned to it, and then the variable is used in place of the required elements for the class, for example in a *ReferencePropertyAtom*. Figure 22 shows the conversion of a role-value-map class to a R2ML *ObjectClassificationAtom*. In this figure, the generated variable *X* shows the variable that later can be used in other places in the rule.

For example, in case we need to convert it to the *ReferencePropertyAtom* of Figure 21, the only change would be to replace *actor1* with *X*. It should be highly noted that, an *ObjectClassificationAtom*, representing a variable, always needs to be placed in the condition part of the derivation rule, even if the variable needs to be used in the conclusion part. Figure 23 also shows another example of how the combination of *ObjectClassificationAtom* and *ReferencePropertyAtom* can be used to define a class of actors in KAoS

A KAoS policy might also have a *trigger action*. To the best of our knowledge, this element is only used with *NegObligation*- and *PosObligation-Policies* showing a set of events that trigger the occurrence of an action. In our *R2ML Policy* model, we have considered a slot for these actions.

Figure 21. The abstract transformation of a KAoS property to a R2ML ReferencePropertyAtom



Furthermore, the detailed information about these actions is defined as *ReferencePropertyAtoms* in the condition part of a derivation rule. Consequently, the same process shown in Figure 21 is also applied to the trigger element, which itself refers to an action. The *obligeAction* element of a KAoS policy is also considered an *ObjectSlot* in the *ObjectDescriptionAtom* of the R2ML Policy object. The *ObjectSlot* refers to the action that should be performed by the actor as an obligation,

Figure 22. The abstract transformation of a KAoS role-value property to a R2ML ObjectClassificationAtom

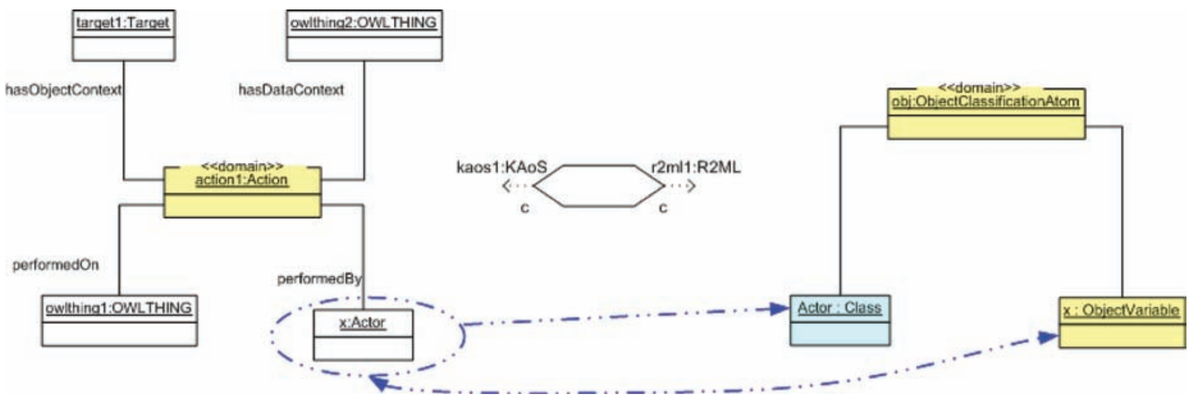


Figure 23. Using ObjectClassificationAtom and ReferencePropertyAtom to map KAoS variables

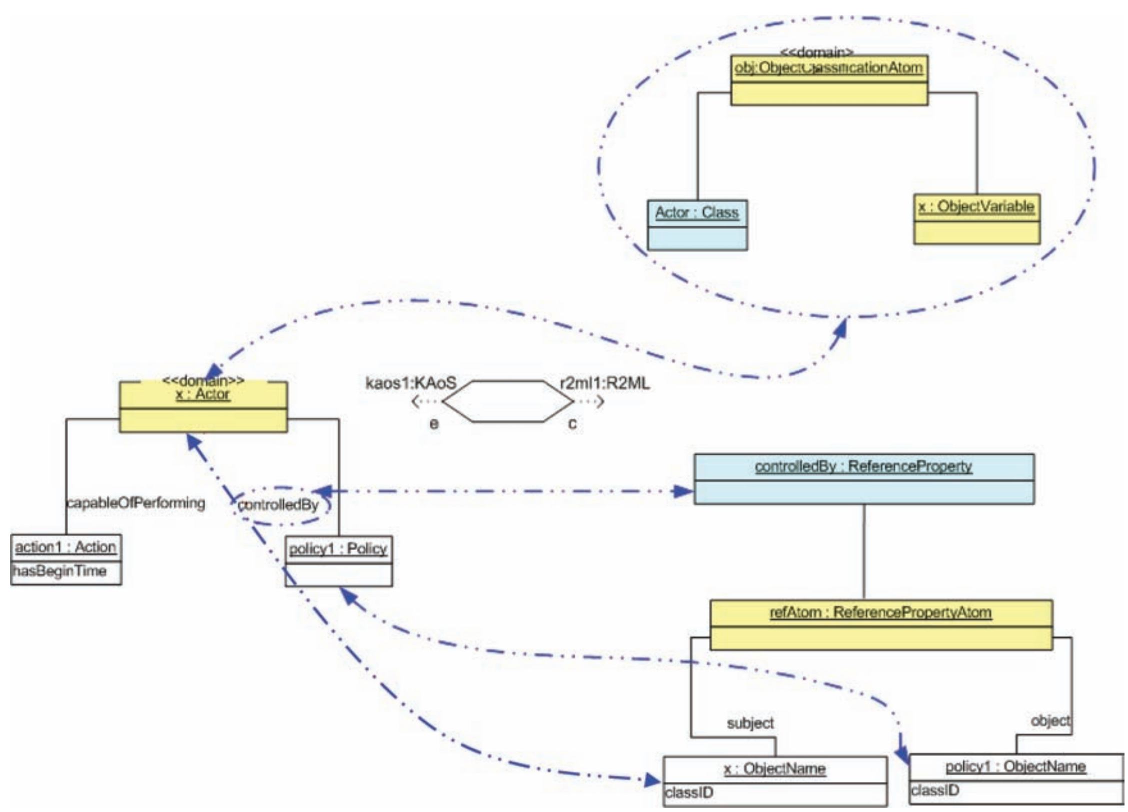


Figure 24. The concrete XML-based conversion of some of the KAoS elements to R2ML elements

	KAoS Element	R2ML Element
Policy Element	<policy:**PosAuthorizationPolicy** rdf:ID="Policy_CommunicationCertificate"> < policy:requiresConditions rdf:resource="#Policy_ TrustedEntity "/> <policy:control s rdf:resource="#Policy_CommunicationCertificate_Action"/> </policy:PosAuthorizationPolicy>	<r2ml:DerivationRule> <r2ml:conditions> <!-- The mappings for the condition part goes here --> </r2ml:conditions> <r2ml:conclusion> <r2ml:ObjectDescriptionAtom r2ml:classID="**Permission**"> <!-- The mappings for the conlcusion part goes here --> </r2ml:ObjectDescriptionAtom> </r2ml:conclusion> </r2ml:DerivationRule>
Control Element	<owl:Class rdf:ID="Policy_CommunicationCertificate_Action"> <owl:intersectionOf> <!-- the set of constraints on the action in KAoS are defined as intersection of a series of restrictions in OWL --> </owl:intersectionOf> </owl:Class>	<r2ml:ObjectDescriptionAtom r2ml:classID="**Permission**"> <!-- The mappings for the conlcusion part goes here --> </r2ml:ObjectDescriptionAtom>
	<owl:Class> <owl:Restriction> <owl:onProperty rdf:resource="http://ontology.ihmc.us/ Action.owl#performedBy"/> <owl:allValuesFrom> <!-- Defines the set of actors that are responsible for the action--> </owl:allValuesFrom> </owl:Restriction> </owl:Class>	<r2ml:ObjectSlot r2ml:referencePropertyID=" http://ontology.ihmc.us/Action.owl#performedBy"> <!--We can use either r2ml:ObjectName or r2ml:ObjectVariable to define the actors --> </r2ml:ObjectSlot>
	<owl:Restriction> <owl:onProperty rdf:resource="http://ontology.ihmc.us/Policy.owl#hasPartner"/> <owl:allValuesFrom rdf:resource="#TrustedServiceProvider"/> </owl:Restriction>	<r2ml:ObjectSlot r2ml:referencePropertyID="context-N10058"> <r2ml:object> <r2ml:ReferencePropertyFunctionTerm r2ml:referencePropertyID=" http://ontology.ihmc.us/Policy.owl#hasPartne"> <r2ml:contextArgument> <r2ml:ObjectName r2ml:name="#TrustedServiceProvider"/> </r2ml:contextArgument> </r2ml:ReferencePropertyFunctionTerm> </r2ml:object> </r2ml:ObjectSlot>

once the policy rule is fired.

Having the most important elements of a KAoS policy explained here, we present some of the concrete, XML syntax-based formalization of our transformation rules in Figure 24.

Other elements of the KAoS policy language are placed in the body of a derivation rule. As a result, we perform a conversion to *ReferencePropertyAtom* for those elements. The main reason is to be compliant with Rei. Moreover, once we have the relations defined in the form of RDF triples, then conversion to the predicates of higher arity is a lot simpler, as for *n* triples with similar *subject* we can make one ObjectDescriptionAtom of arity ***n*** by combining all these elements.

AN EXAMPLE OF APPLYING THE TRANSFORMATION RULES TO A KAOS POLICY

To this point, we have conceptualized and formalized our policy transformation framework. Now that the formalization of our transformations is complete, we can apply our definitions to a real example of the KAoS policy taken from (Toninelli,

Figure 25. (a) A KAoS policy restricting access to a printer only to the community of sky team customers and, (b) its equivalent policy as an R2ML derivation rule

```
<owl:Class rdf:ID=”SkyTeamCustomer”>
    <owl:subClassOf
                  rdf:resource=”&some-
                  ontology;Customer”/>
    <owl:subClassOf>

       <owl:Restriction>
           <owl:onProperty
                  rdf:resource=”&some-
                      ontology;firm”/>
           <owl:allValuesFrom
                 rdf:resource=”&some-
           ontology;SkyTeamAlliance”/>
       </owl:Restriction>
    </owl:subClassOf>
</owl:Class>

<owl:Class rdf:ID=”SkyTeamGate31-
               57PrinterAccessAction”>
    <owl:intersectionOf
           rdf:parseType=”Collection”>

 <owl:Class
           rdf:about=
              “&action;AccessAction”/>
       <owl:Restriction>
            <owl:onProperty
                rdf:resource=
               “&action;performedBy”/>
            <owl:allValuesFrom
                   rdf:resource=
                  “#SkyTeamCustomer”/>
       </owl:Restriction>
       <owl:Restriction>
            <owl:onProperty
            rdf:resource=
            “&action;accessedEntity”/>
            <owl:allValuesFrom
                      rdf:resource=
                      “#Printer31-57”/>
       </owl:Restriction>
    </owl:intersectionOf>
</owl:Class>


<policy:PosAuthorizationPolicy
                rdf:ID=”SkyTeamGate31-
                     57PrinterAccess”>
    <policy:controls
        rdf:resource=”#SkyTeamGate31-
        57PrinterAccessAction”/>
    <policy:hasSiteOfEnforcement
        rdf:resource=”&some-
        ontology;TargetSite”/>
    <policy:hasPriority> 10
                 </policy:hasPriority>
</policy:PosAutihorizationPolicy>
```

```
<r2ml:DerivationRule>
  <r2ml:conditions>
     <r2ml:ObjectClassificationAtom
      r2ml:classID=”&kaos_action:AccessAction”>
        <r2ml:ObjectVariable r2ml:name=”X”/>
     </r2ml:ObjectClassificationAtom>

     …

     <r2ml:ReferencePropertyAtom
      r2ml:referencePropertyID=
      “kaos_action:performedBy”>

         <r2ml:subject>

      <r2ml:ObjectVariable
       r2ml:name=”X”/>

         </r2ml:subject>

         <r2ml:object>

      <r2ml:ObjectVariable
       r2ml:name=”Y”/>

         </r2ml:object>

     </r2ml:ReferencePropertyAtom>
       …
     <r2ml:ObjectDescriptionAtom
      r2ml:classID=”&some-ontology;Customer”>
  <r2ml:subject>
          <r2ml:ObjectVariable
           r2ml:name=”Y”
           r2ml:classID=”&some-
               ontology;SkyTeamCustomer”/>
         </r2ml:subject>
        <r2ml:ObjectSlot
          r2ml:referencePropertyID=”&some-
                              ontology;firm”>
          <r2ml:object>
             <r2ml:ObjectVariable
              r2ml:name=”W”
         r2ml:classID=”&some-
          ontology;SkyTeamAlliance”/>
    </r2ml:Object>
 </r2ml:ObjectSlot>
     </r2ml:ObjectDescriptionAtom>
  </r2ml:conditions>


  <r2ml:conclusion>
     <r2ml:ObjectDescriptionAtom
      r2ml:classID=”r2ml_policy:Permission”
        <r2ml:subject>
          <r2ml:ObjectName
           r2ml:objectID=”SkyTeamGate31-
                           57PrinterAccess”/>
        </r2ml:subject>
        <r2ml:ObjectSlot
r2ml:referencePropertyID=”r2ml_policy:hasAction”>
   <r2ml:object>
        <r2ml:ObjectVariable
```

Bradshaw, Kagal, & Montanari, 2005). Let us consider a travel agency, named Sky Team that plans to promote its services by giving free printing access to those of its customers located in any of the gates of 31 to 57 at the local airport. Figure 25a shows the KAoS representation of this policy. The policy is a PosAuthorizationPolicy and instantiates hasSiteOfEnforcement and hasPriority.

As already stated, we define four main policy classes for our policy rules, including, *Permission*, *Prohibition*, *Obligation*, and *Dispensation*. The first step in providing the mappings is to determine the appropriate R2ML classes and elements that can be used according to our definitions in the previous section. A *PosAuthorizationPolicy* in KAoS is equivalent to our definition of the *Permission* class in R2ML. Having a KAoS policy element necessitates the definition of an *ObjectDescriptionAtom* to describe its properties, which should be placed in the conclusion part of a derivation rule. The *ObjectDescriptionAtom* takes the name of the policy instance as its *subject* and is considered as an instance of the *Permission* class that we defined in the previous chapter. The *hasSiteOfEnforcement* property from a KAoS policy is modeled as an *ObjectSlot* for our *ObjectDescriptionAtom* with its *referencePropertyID* referring to *context* according to our definition of R2ML Policy class. The *hasPriority* attribute in KAoS has an equivalent *hasPriority* element in our R2ML Policy model which is considered as a *DataSlot*.

In our example, we have a PrinterAccessAction as a subclass of AccessAction for which we have the metamodel illustrated in Figure 26. As the figure shows, accessedEntity is the subProperty of hasObjectContext and thus the class that this property refers to (i.e. Printer31-57 in our example) plays as the context for the PrinterAccessAction.

Having a look at the SkyTeamGate31-57PrinterAccessAction class, as the to-be-controlled class, we realize that SkyTeamCustomer, Printer31-57, and AccessAction are OWL classes. To be able to place these elements in the conclusion part of a derivation rule, we need to associate the classes with variables which show the extensional meaning of the classes. The object for *hasAction* refers to variable *X*. *ObjectClassificationAtoms* seem to be appropriate R2ML constructs to connect a variable to its class name. However, they need to be placed in the condition part of the derivation rule. Getting to this point, we have mapped almost everything in Table 3a to R2ML, but Table 3 provides a description for the class SkyTeamCustomer, considering it as a subclass of the class Customer with its firm property only limited to the class SkyTeamAlliance. As it is a descriptive class, it should be placed in the condition part of a derivation rule, and *ObjectDescriptionAtom* is probably the best element to map this class to.

Figure 26. The KAoS metamodel for AccessAction

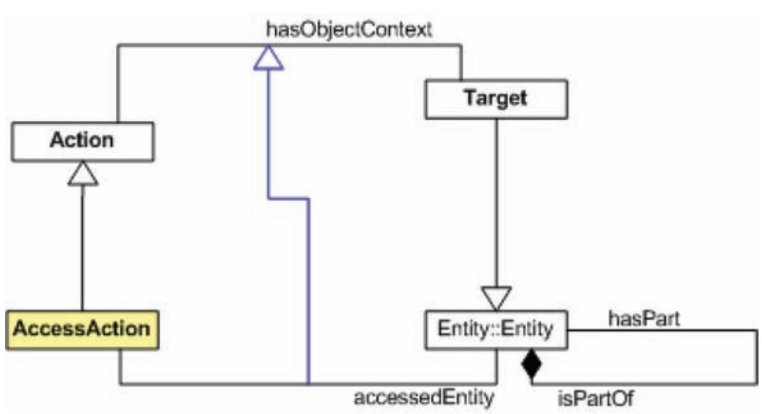


Table 3b shows the full transformation of this KAoS policy to its equivalent R2ML code based on the provided descriptions

DISCUSSION AND CONCLUSION

In this last section, we analyze the transformations that we have explained so far, we show how these transformations can be used together with other transformations that have been already developed for R2ML based on derivation rules, and we discuss the possibility of using them to convert our policy-driven R2ML rules to the rule languages that are supported by those policy languages

Compatibility of our Transformations with other R2ML Transformations

As we already stated, R2ML is supposed to be an intermediary language, to/from which other policy languages can be transformed. To achieve this goal several transformations have been defined and developed between R2ML and other rule languages by mapping the rules from these languages to one of four main types of rules that R2ML supports, i.e. *derivation rules*, *integrity rules*, *production rules*, and *reaction rule*s. Table 2 shows the current transformations that have been developed between R2ML and other rule languages as well as their corresponding R2ML rules which have been used. As our transformations of policy rules to R2ML are converted to its subset of derivation rules, we can use the currently existing transformation of derivation rules to convert our policies to the rules in any of the target languages (i.e., RuleML, Jess, Jena, and F-Logic).

Information Loss During Policy Transformation

Information loss happens in most transformations between different policy languages due to the differences in their underlying concepts. However, in our model, regardless of the policy language that is being used, the concepts can be mapped to R2ML thanks to the rich set of elements that R2ML supports for different logical bases. For example, the concept of *Domain* from KAoS can be covered by extending our R2ML Policy model of Figure 17, yet it should be investigated how these concepts are effective and useful when they are shared between different policy languages. Extending R2ML by the concepts from different policy languages under study does not make the whole model more generic, but it reduces the expressivity when coping with the abstract notation of different policy languages. This implies that the more we try to adjust our R2ML policy model to one policy language, the more it will be diverged

Table 2. Transformations between R2ML and other Rule Languages. Right arrows explain transformations from R2ML to these languages while double head arrows explain transformations between R2ML and the language in both directions

R2ML	**RuleML**	**Jess**	**F-Logic**	**JBoss**	**Jena**	**SWRL**	**OCL**
Derivation	→	→	→		→		
Integrity						↔	↔
Reaction							
Production	→	→		→	→		
Transformation Language	XSLT					QVT/ATL	

from adjusting to the concepts of other policy languages. Extensions to a policy model should be done carefully and after deeply reviewing the pros and cons of the newly added concepts. There is a tradeoff between supporting the semantics of one policy language and being able to cope with the semantic of other policy languages, which if not carefully designed, would render the R2ML policy model useless.

To be able to check the feasibility of the transformations in protecting the resources, a deeper analysis should be conducted. The most promising approach would be to deploy the policies on different resources and broker agents and get them to communicate. The level of jeopardizing the resources and the to-be-protected contents should be carefully examined. Policies are the critical rules in a business system which regulates the behavior of the system. Any flaw or miss-interpretation of the policies may result in the non-compostable loss of data and information. This further necessitates a detailed examination of the policies while working on the level of policy exchange. The practical analysis of the transformed policies is part of the upcoming research project that will be conducted in the Laboratory for Ontological Research (LORe) at Simon Fraser University

This book chapter demonstrated the possibility to exchange the policy languages by *1) high level metamodel representation of their concepts*, *2) capturing the semantics of the concepts*, *3) identifying the similarities and dissimilarities of the modeled concepts*, and *4) applying the transformations to the concrete syntax of the policy language, benefiting from the QVT representation of the transformations*. All the above steps are compliant with the required steps in software design using MDE techniques (Schmidt, 2006). Using MDE techniques to provide the transformations significantly assists with having valid models for the source and the target languages. Furthermore, while using MDE, we are able to detect the inconsistent or missing constructs of different languages. During working with policies we discovered some points of improvement for R2ML which can add to its efficiency. First of all, we found that it is important for rules to have a value as an indicator for their priority. Having a set of rules, as in R2ML derivation or integrity rule sets, we need to provide an indicator of how one rule in this series can be chosen over the other ones. It especially helps with cases where more than one rule might be applicable to a certain situation. Although this issue matters mainly at the level of rule enforcement, where the enforcement engine needs to select among a set of applicable rules, and knowing that the current versions of R2ML are not supposed to be used at any enforcement level, dealing with rule exchange for cases where the source or the target rules need to have priorities makes it necessary to have priority indicators for the rules.

Another point that seems to be open to extend in R2ML is the possibility to use quantifiers to express cardinality. While working with KAoS, we realized that this language supports *min_cardinality* and *max_cardinality*. However, these quantified formulas are not possible to be used in the condition of a derivation rule set as the elements of the condition part are considered quantifier free and should be universally quantified. An extension to R2ML such that cardinalities can be supported in the condition of the derivation rules (and possibly other types of rules) has been also considered as a potential improvement in the next version of R2ML. Further to this, there have been long discussions on whether derivation rules or integrity rules should be used to model the policies. Looking back at the definitions of integrity and derivation rules, we see that integrity rules define that something *must necessarily* hold, or it *should* hold, while derivation rules carry a derivative meaning representing a set of new conclusions based on the presented facts. Both of these types of rules can be used to define the policies depending on the type and the purpose of the policy (i.e. authorization, authentication, and Quality of Service).

REFERENCES

Boley, H., Tabet, S., & Wagner, G. (2001). Design rationale of RuleML: A markup language for Semantic Web rules. In *International Semantic Web Working Symposium (SWWS)* (pp. 381-402). CA.

Bonatti, P., Duma, C., Fuchs, N., Nejdl, W., Olmedilla, D., Peer, J., et al. (2006). *Semantic Web policies—A discussion of requirements and research issues.* Paper presented at 3rd European Semantic Web Conference (ESWC).

Bonatti, P., & Olmedilla, D. (2005). Driving and monitoring provisional trust negotiation with metapolicies. In *IEEE 6th International Workshop on Policies for Distributed Systems and Networks* (pp. 14-23). Stockholm.

Bradshaw, J., Dutfield, S., Benoit, P., & Woolley, J. (1997). KAoS: Toward an industrial-strength open agent architecture. *Software Agents*, 375-418.

Budinsky, F., Brodsky, S., & Merks, E. (2003). *Eclipse modeling framework.* Pearson Education.

Clark, J. (1999, November 16). XSL transformations (XSLT). *W3C Recommendation.*

Damianou, N., Dulay, N., Lupu, E., & Sloman, M. (2001). *The ponder policy specification language* (LNCS, pp. 18-38.

Fallside, D., & Walmsley, P. (2004, October 28). XML schema part 0: Primer second edition. *W3C Recommendation* .

Ginsberg, A., Hirtle, D., McCabe, F., & Patranjan, P.-L. (2006, July 10). RIF use cases and requirements. *W3C Working Draft* .

Godik, S., & Moses, T. (2003). eXtensible Access Control Markup Language (XACML). *Specification, OASIS.*

Grosof, B., Horrocks, I., Volz, R., & Decker, S. (2003). Description logic programs: Combining logic programming with description logic. In *WWW2003 Conference.* Budapest: ACM.

Horrocks, I., Patel-Schneider, P., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004, May 21). SWRL: A Semantic Web rule language combining OWL and RuleML. *W3C Member Submission* .

Kagal, L. (2002). *Rei: A policy language for the Me-Centric Project* (Tech. Rep).HP Labs.

Kalfoglou, Y., & Schorlemmer, M. (2003). Ontology mapping: The state of the art. *The Knowledge Engineering Review*, *18*(1), 1–31. doi:10.1017/S0269888903000651

Kaviani, N., Gasevic, D., & Hatala, M. (2007-1). Web rule languages to carry policies. In *Proceedings of the 2007 IEEE Workshop on Policies for Distributed Systems and Networks (POLICY2007), Bologna, Italy* (pp. 188-192).

Kaviani, N, Gasevic, D., Hatala, M., Clement, D., & Wagner, G. (2007-2). Integration of rules and policies for Semantic Web services. *International Journal of Advanced Media and Communication (IJAMC)*, *1*(4), 404-423.

Kolovski, V., Hendler, J., & Parsia, B. (2007). Analyzing Web access control policies. *WWW* (pp. 677-686).

Kühne, T. (2006). Matters of (meta-)modeling. *Software and System Modeling*, *5*(4), 369–385. doi:10.1007/s10270-006-0017-9

Lassila, O., & Swick, R. (1999). Resource Description Framework (RDF) model and syntax specification. *W3C Recommendation*.

Miller, J., & Mukerji, J. (2003). MDA guide version 1.0.1. *Object Management Group*.

MOF. (2008, February 20). *Meta Object Facility (MOF) specification 2.0 facility and object lifecycle*.

Nejdl, W., Olmedilla, D., & Winslett, M. (2004). *PeerTrust: Automated trust negotiation for peers on the Semantic Web*. Paper presented at Workshop on Secure Data Management in a Connected World (SDM'04).

ODM. (2001, August 5). *ODM: The Object Management Group ontology definition metamodel*. Retrieved from http://www.omg.org/cgi-bin/apps/doc?ad/05-08-01.pdf

ODM. (2003, March 4). Ontology definition metamodel. *Third Revised Submission to OMG/RFP* (pp. 1-314).

QVT. (2005, June). MOF QVT final adopted specification. *OMG* .

Schmidt, D. (2006). Model-driven engineering. *IEEE Computer*, *39*(2), 25–31.

Smith, M., Welty, C., & McGuinness, D. (2004, February 10). OWL Web ontology language guide. *W3C Recommendation* .

Toninelli, A., Bradshaw, J., Kagal, L., & Montanari, R. (2005). *Rule-based and Ontology-based policies: Toward a hybrid approach to control agents in pervasive environments*. Paper preseneted at the Semantic Web and Policy Workshop, Ireland.

Tonti, G., Bradshaw, J., Jeffers, R., Montanari, R., Suri, N., & Uszok, A. (2003). Semantic Web languages for policy representation and reasoning: A comparison of KAoS, Rei, and Ponder. In *The International Semantic Web ConferenceI (ISWC)* (pp. 419-437).

Uszok, A., & Bradshaw, J. (n.d.). IHMC ontology and policy management. Retrieved from http://ontology.ihmc.us/

Uszok, A., Bradshaw, J., Jeffers, R., Suri, N., Hayes, P., Breedy, M., et al. (2003). KAoS policy and domain services: Toward a description-logic approach to policy representation, deconfliction, and enforcement. In *IEEE 4th International Workshop on Policies for Distributed Systems and Networks* (pp. 93-96).

Wagner, G., Damasio, C., & Antoniou, G. (2005). Towards a general Web rule language. *International Journal of Web Engineering and Technology*, *2*(2), 181–206. doi:10.1504/IJWET.2005.008483

Wagner, G., Giurca, A., & Lukichev, S. (2005). A general markup framework for integrity and derivation rules. In *Principles and practices of Semantic Web reasoning*. Dagstuhl.

Wagner, G., Giurca, A., & Lukichev, S. (2005). R2ML: A general approach for marking-up rules. In *principles and practices of Semantic Web reasoning*.

Wagner, G., Giurca, A., & Lukichev, S. (2006). A *usable interchange format for rich syntax rules integrating OCL, RuleML and SWRL*. Paper presented at Reasoning on the Web Workshop at WWW2006, Edinburgh, UK.

XSB. (2007, October 6). *XSB Prolog*. Retrieved from http://xsb.sourceforge.net/

KEY TERMS AND DEFINITIONS

KAoS: KAoS language is a multi-paradigm specification language with a two-level structure: an outer semantic net layer for declaring concepts, their attributes and links to other concepts and an inner formal assertion layer for formally defining the concept.

MDA: an approach to application design and implementation that encourages efficient use of system models in the software development process, and it supports reuse of best practices when creating families of systems.

MOF 2: a metadata management framework, and a set of metadata services that enable the development and interoperability of model- and metadata-driven systems.

OWL: Web Ontology Language designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. This representation of terms and their interrelationships is called an ontology.

QVT: a standard for model transformation languages

R2ML: a MOF-defined general rule language that captures integrity, derivation, production, and reaction rules. It is a well-known proposal for RIF.

RIF: Rule Interchange Format is W3C initiative that should define an intermediary language between various rule languages, but it should not provide a formally defined semantic foundation for reasoning on the Web such as OWL for ontologies.

ENDNOTE

1 The fragment of Horn FOL with no function symbols (Grossof et al. 2003).

This work was previously published in the Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches, edited by A. Giurca, D. Gasevic, and K. Taveter, pp. 564-595, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 7.19
Scalable Authoritative OWL Reasoning for the Web

Aidan Hogan
National University of Ireland, Ireland

Andreas Harth
National University of Ireland, Ireland

Axel Polleres
National University of Ireland, Ireland

ABSTRACT

In this article the authors discuss the challenges of performing reasoning on large scale RDF datasets from the Web. Using ter-Horst's pD* fragment of OWL as a base, the authors compose a rule-based framework for application to web data: they argue their decisions using observations of undesirable examples taken directly from the Web. The authors further temper their OWL fragment through consideration of "authoritative stheirces" which counter-acts an observed behavitheir which we term "ontology hijacking": new ontologies published on the Web re-defining the semantics of existing entities resident in other ontologies. They then present their system for performing rule-based forward-chaining reasoning which they call SAOR: Scalable Authoritative OWL Reasoner. Based upon observed characteristics of web data and reasoning in general, they design their system to scale: the system is based upon a separation of terminological data from assertional data and comprises of a lightweight in-memory index, on-disk sorts and file-scans. The authors evaluate their methods on a dataset in the order of a hundred million statements collected from real-world Web stheirces and present scale-up experiments on a dataset in the order of a billion statements collected from the Web.

INTRODUCTION

Information attainable through the Web is unique in terms of scale and diversity. The Semantic Web movement aims to bring order to this information by providing a stack of technologies, the core of which is the Resource Description Framework

(RDF) for publishing data in a machine-readable format: there now exists millions of RDF data-sources on the Web contributing billions of statements. The Semantic Web technology stack includes means to supplement instance data being published in RDF with ontologies described in RDF Schema (RDFS) (Brickley and Guha 2004) and the Web Ontology Language (OWL) (Bechhofer, van Harmelen et al. 2004; Smith, Welty et al. 2004), allowing people to formally specify a domain of discourse, and providing machines a more sapient understanding of the data. In particular, the enhancement of assertional data (i.e., instance data) with terminological data (i.e., structural data) published in ontologies allows for deductive reasoning: i.e., inferring implicit knowledge.

In particular, our work on reasoning is motivated by the requirements of the Semantic Web Search Engine (SWSE) project: http://swse.deri.org/, within which we strive to offer search, querying and browsing over data taken from the Semantic Web. Reasoning over aggregated web data is useful, for example: to infer new assertions using terminological knowledge from ontologies and therefore provide a more complete dataset; to unite fractured knowledge (as is common on the Web in the absence of restrictive formal agreement on identifiers) about individuals collected from disparate sources; and to execute mappings between domain descriptions and thereby provide translations from one conceptual model to another. The ultimate goal here is to provide a "global knowledge-base", indexed by machines, providing querying over both the explicit knowledge published on the Web and the implicit knowledge inferable by machine. However, as we will show, complete inferencing on the Web is an infeasible goal, due firstly to the complexity of such a task and secondly to noisy web data; we aim instead to strike a comprise between the above goals for reasoning and what is indeed feasible for the Web.

Current systems have had limited success in exploiting ontology descriptions for reasoning over RDF web data. While there exists a large body of work in the area of reasoning algorithms and systems that work and scale well in confined environments, the distributed and loosely coordinated creation of a world-wide knowledge-base creates new challenges for reasoning:

- the system has to perform on web-scale, with implications on the completeness of the reasoning procedure, algorithms and optimisations;
- the method has to perform on collaboratively created knowledge-bases, which has implications on trust and the privileges of data publishers.

With respect to the first requirement, many systems claim to inherit their scalability from the underlying storage—usually some relational database system—with many articles having been dedicated to optimisations on database schemata and access; c.f. (Pan and Heflin 2003; Theoharis, Christophides et al. 2005; Zhou, Ma et al. 2006; Hondjack, Pierra et al. 2007). With regards the second requirement, there have been numerous papers dedicated to the inter-operability of a small number of usually trustworthy ontologies; c.f. (Ghilardi, Lutz et al. 2006; Lutz, Walther et al. 2007; Jiménez-Ruiz, Grau et al. 2008). We leave further discussion of related work to Section 6, except to state that the combination of web-scale and web-tolerant reasoning has received little attention in the literature and that our approach is novel.

Our system, which we call "Scalable Authoritative OWL Reasoner" (SAOR), is designed to accept as input a web knowledge-base in the form of a body of statements as produced by a web-crawl and to output a knowledge-base enhanced by forward-chaining reasoning over a given fragment of OWL. In particular, we choose forward-chaining to avoid the runtime complexity of query-rewriting associated with backward-chaining approaches: in the web search scenario, the requirement for low query response times and resource usage preclude the applicability of query-rewriting for many reasoning tasks.

SAOR adopts a standard rule-based approach to reasoning whereby each rule consists of (i) an 'antecedent': a clause which identifies a graph pattern that, when matched by the data, allows for the rule to be executed and (ii) a 'consequent': the statement(s) that can be inferred given data that match the antecedent. Within SAOR, we view reasoning as a once-off rule-processing task over a given set of statements. Since the rules are all known *a-priori*, and all require simultaneous execution, we can design a task-specific system that offers much greater optimisations over more general rule engines. Firstly, we categorise the known rules according to the composition of their antecedents (e.g., with respect to arity, proportion of terminological and assertional patterns, etc.) and optimise each group according to the observed characteristics. Secondly, we do not use an underlying database or native RDF store and opt for implementation using fundamental data-structures and primitive operations; our system is built from scratch specifically (and only) for the purpose of performing pre-runtime forward-chaining reasoning which gives us greater freedom in implementing appropriate task-specific optimisations.

This article is an extended version of (Hogan, Harth et al. 2008), in which we presented an initial modus-operandi of SAOR; we provided some evaluation of a set of rules which exhibited linear scale and concluded that using dynamic index structures, in SAOR, for more complex rulesets, was not a viable solution for a large-scale reasoner. In this article, we provide extended discussion of our fragment of OWL reasoning and additional motivation for our deliberate incompleteness in terms of computational complexity and impediments posed by web data considerations. We also describe an implementation of SAOR which abandons dynamic index structures in favour of batch processing techniques known to scale: namely sorts and file-scans. We present new evaluation of the adapted system over a dataset of 147m triples collected from 665k web sources and also provide scale-up evaluation of our most optimised ruleset on a dataset of 1.1b statements collected from 6.5m web sources.

Specifically, we make the following contributions in this article:

- We discuss and apply a selected rule-based subset of OWL reasoning, i) to be computationally efficient, ii) to avoid an explosion of inferred statements, iii) to be tolerant to noisy web data and iv) to protect existing specifications from undesirable contributions made in independent locations. That is, our system implements a positive fragment of OWL Full which has roots in ter Horst's pD* (ter Horst 2005) entailment rules and our system includes analysis of the authority of sources to counter-act the problem of ontology hijacking in web data (Section 3).
- We describe a scalable, optimised method for performing rule-based forward-chaining reasoning for our fragment of OWL. In particular, we refine our algorithm to capitalise on the similarities present in different rule antecedent patterns and the low volume of terminological data relative to assertional data. We implement the system using on-disk batch processing operations known to scale: sorts and scans (Section 4).
- We show experimentally that a forward-chaining materialisation approach is feasible on Web data, showing that, by careful materialisation through our tailored OWL ruleset, we can avoid an explosion of inferred statements. We present evaluation with respect to computation of our most expressive ruleset on a dataset of 147m statements collected from 665k sources and present scale-up measurements by applying

our most optimised ruleset on a dataset of 1.1b statements collected from 6.5m sources. We also reveal that the most computationally efficient segment of our reasoning is the most productive with regards inferred output statements (Section 5).

We discuss related work in Section 6 and conclude with Section 7.

PRELIMINARIES

Before we continue, we briefly introduce some concepts prevalent throughout the article. We use notation and nomenclature as is popular in the literature, particularly from (Hayes 2004).

RDF Term

Given a set of URI references $\mathcal{U}$, a set of blank nodes $\mathcal{B}$, and a set of literals $\mathcal{L}$, the set of *RDF terms* is denoted by $\mathcal{RDFTerm} = \mathcal{U} \cup \mathcal{B} \cup \mathcal{L}$. The set of blank nodes $\mathcal{B}$ is a set of existensially quantified variables. The set of literals is given as $\mathcal{L} = \mathcal{L}_p \cup \mathcal{L}_t$, where $\mathcal{L}_p$ is the set of plain literals and $\mathcal{L}_t$ is the set of typed literals. A typed literal is the pair $l = (s, t)$, where s is the lexical form of the literal and $t \in \mathcal{U}$ is a datatype URI. The sets $\mathcal{U}$, $\mathcal{B}$, $\mathcal{L}_p$ and $\mathcal{L}_t$ are pairwise disjoint.

RDF Triple

A triple $t = (s, p, o) \in (\mathcal{U} \cup \mathcal{B}) \times \mathcal{U} \times (\mathcal{U} \cup \mathcal{B} \cup \mathcal{L})$ is called an *RDF* triple. In a triple (s, p, o), s is called subject, p predicate, and o object.

RDF Triple in Context/RDF Quadruple

A pair (t, c) with a triple $t = (s, p, o)$ and $c \in \mathcal{U}$ is called a triple in context c (Guha, McCool et al. 2004; Harth and Decker 2005; Prud'hommeaux and Seaborne 2008). We may also refer to (s, p, o, c) as the RDF quadruple or quad q with context c.

We use the term 'RDF statement' to refer generically to triple or quadruple where differentiation is not pertinent.

RDF Graph/Web Graph

An *RDF graph* $\mathcal{G}$ is a set of *RDF t*riples; that is, a subset of $(\mathcal{U} \cup \mathcal{B}) \times \mathcal{U} \times (\mathcal{U} \cup \mathcal{B} \cup \mathcal{L})$.

We refer to a *web graph* $\mathcal{W}$ as a graph derived from a given web location (i.e., a given document). We call the pair $(\mathcal{W}, c)$ a web-graph $\mathcal{W}$ in context c, where c is the web-location from which $\mathcal{W}$ is retrieved. Informally, $(\mathcal{W}, c)$ is represented as the set of quadruples (t_w, c) for all $t_w \in \mathcal{W}$.

Generalised Triple

A triple $t = (s, p, o) \in (\mathcal{U} \cup \mathcal{B} \cup \mathcal{L}) \times (\mathcal{U} \cup \mathcal{B} \cup \mathcal{L}) \times (\mathcal{U} \cup \mathcal{B} \cup \mathcal{L})$ is called a *generalised triple*.

The notions of generalised quadruple, generalised statement and generalised graph follow naturally. Our definition of "generalised" is even more liberal than that described in (ter Horst 2005) wherein blank nodes are allowed in the predicate position: we also allow literals in the subject and predicate position. Please note that we may refer generically to a "triple", "quadruple", "graph" etc. where a distinction between the "generalised" and "RDF" versions is not pertinent.

Merge

The merge $M(\mathcal{S})$ of a set of graphs $\mathcal{S}$ is the union of the set of all graphs $\mathcal{G}'$ for $\mathcal{G} \in \mathcal{S}$ and $\mathcal{G}'$ derived from $\mathcal{G}$ such that $\mathcal{G}'$ contains a unique set of blank nodes for $\mathcal{S}$.

Web Knowledge-Base

Given a set $\mathcal{S}_{\mathcal{W}}$ of RDF web graphs, our view of a web knowledge-base $\mathbb{KB}$ is taken as a set of pairs $(\mathcal{W}', c)$ for each $\mathcal{W} \in \mathcal{S}_{\mathcal{W}}$, where $\mathcal{W}'$ contains a unique set of blank nodes for $\mathcal{S}_{\mathcal{W}}$ and c denotes the *URL* location of $\mathcal{W}$.

Informally, $\mathbb{KB}$ is a set of quadruples retrieved from the Web wherein the set of blank nodes are unique for a given document and triples are enhanced by means of context which tracks the web location from which each triple is retrieved. We use the abbreviated notation $\mathcal{W} \in \mathbb{KB}$ or $\mathcal{W}' \in \mathbb{KB}$ where we mean $\mathcal{W} \in \mathcal{S}_{\mathcal{W}}$ for $\mathcal{S}_{\mathcal{W}}$ from which $\mathbb{KB}$ is derived or $(\mathcal{W}', c) \in \mathbb{KB}$ for some c.

Class

We refer to a class as an RDF term which appears in either:

- o of a triple t where p is `rdf:type`; or
- s of a triple t where p is `rdf:type` and o is `rdfs:Class` or `:Class`[1].

Property

We refer to a property as an RDF term which appears in either:

- p of a triple t; or
- s of a triple t where p is `rdf:type` and o is `rdf:Property`.

Membership Assertion

We refer to a triple t as a membership assertion of the property mentioned in predicate position p. We refer to a triple t with predicate `rdf:type` as a membership assertion of the class mentioned in the object o. For a class or property v, we denote a membership assertion as $m(v)$.

Meta-Class

A meta-class is a class of classes or properties; i.e., the members of a meta-class are either classes or properties. The set of RDF(S) and OWL meta-classes is as follows: { `rdf:Property`, `rdfs:Class`, `rdfs:ContainerMembershipProperty`, `:AnnotationProperty`, `:Class`, `:DatatypeProperty`, `:DeprecatedClass`, `:DeprecatedProperty`, `:FunctionalProperty`, `:InverseFunctionalProperty`, `:ObjectProperty`, `:OntologyProperty`, `:Restriction`, `:SymmetricProperty`, `:TransitiveProperty` }.

Meta-Property

A meta-property is one which has a meta-class as it's domain. Meta-properties are used to describe classes and properties. The set of RDFS and OWL meta-properties is as follows: { `rdfs:domain`, `rdfs:range`, `rdfs:subClassOf`, `rdfs:subPropertyOf`, `:allValuesFrom`, `:cardinality`, `:complementOf`, `:disjointWith`, `:equivalentClass`, `:equivalentProperty`, `:hasValue`, `:intersectionOf`, `:inverseOf`, `:maxCardinality`, `:minCardinality`, `:oneOf`, `:onProperty`, `:someValuesFrom`, `:unionOf` }.

Terminological Triple

We define a *terminological* triple as one of the following:

- a membership assertion of a meta-class; or
- a membership assertion of a meta-property; or
- a triple in a non-branching, non-cyclic path $t_0^r, \ldots, t_n^r$ where $t_0^r = (s_0, p_0, o_0)$ *for* $p_0 \in$ {`:intersectionOf`, `:oneOf`, `:unionOf`}; $t_k^r = (o_{k-1}$, `rdf:rest`, $o_k)$ for $1 \le k \le n$, $o_{k-1} \in \mathcal{B}$ and o_n = `rdf:nil`; or a triple $t_k^f = (o_k$, `rdf:first`, $e_k)$ with o_k for $1 \le k < n$ as before.

We refer to triples $t_0^r, \ldots, t_n^r$ and all triples t_k^f as terminological collection triples, whereby *RDF* collections are used in a union, intersection or enumeration class description.

Triple Pattern, Basic Graph Pattern

A *triple pattern* is defined as a generalised triple where, in all positions, variables from the infinite set $\mathcal{V}$ are allowed; i.e.: $tp = (s_v, p_v, o_v) \in \times (\mathcal{U} \cup \mathcal{B} \cup \mathcal{L} \cup \mathcal{V}) \times (\mathcal{U} \cup \mathcal{B} \cup \mathcal{L} \cup \mathcal{V})$. A set (to be read as conjunction) of triple patterns $\mathcal{GP}$ is also called a basic graph pattern.

We use—following SPARQL notation (Prud'hommeaux and Seaborne 2008)—alphanumeric strings preceded by '?' to denote variables in this article: e.g., ?x. Following common notation, such as is used in SPARQL and Turtle[2], we delimit triples in the same basic graph pattern by '.' and we may group triple patterns with the same subject or same subject-predicate using ';' and ',' respectively. Finally, we denote by $\mathcal{V}(tp)$ (or $\mathcal{V}(\mathcal{GP})$, resp.) the set of variables appearing in tp (or in $\mathcal{GP}$, resp.).

Instance

A triple $t = (s, p, o)$ (or, resp., a set of triples, i.e., a graph $\mathcal{G}$) is an instance of a triple pattern $tp = (s_v, p_v, o_v)$ (or, resp., of a basic graph pattern $\mathcal{GP}$) if there exists a mapping $\mu{:}\mathcal{V} \cup \mathcal{RDFTerm} \rightarrow \mathcal{RDFTerm}$ which maps every element of $\mathcal{RDFTerm}$ to itself, such that $t = \mu(tp) = (\mu(s_v), \mu(p_v), \mu(o_v))$ (or, resp., and slightly simplifying notation, $\mathcal{G} = \mu(\mathcal{GP})$.

Terminological/Assertional Pattern

We refer to a *terminological -triple/-graph pattern* as one whose instance can only be a terminological triple or, resp., a set thereof. We denote a terminological collection pattern by *?x* p *(?e*^{1}, ..., *?e*^{n}*)* . where $p \in$ {`:intersectionOf`, `:oneOf`, `:unionOf`} and ?e^k is mapped by the object of a terminological collection triple $t_k^r = (o_k$, `rdf:first`, $e_k)$ as before. An assertional pattern is any pattern which is not terminological.

Inference Rule

We define an *inference rule* r as the pair ($\mathcal{A}nte$, $\mathcal{C}on$), where the antecedent $\mathcal{A}nte$ and the *consequent* $\mathcal{C}on$ are basic graph patterns such that $\mathcal{V}(\mathcal{C}on)$ and $\mathcal{V}(\mathcal{A}nte)$ are non-empty, $\mathcal{V}(\mathcal{C}on) \subseteq \mathcal{V}(\mathcal{A}nte)$ and $\mathcal{C}on$ does not contain blank nodes.[3] In this article, we will typically write inference rules as:

$$\mathcal{C}on \Rightarrow \mathcal{A}nte \qquad (1)$$

Rule Application and Closure

We define a *rule application* in terms of the immediate consequences of a rule r or a set of rules $\mathcal{R}$ on a graph $\mathcal{G}$ (here slightly abusing the notion of the immediate consequence operator in Logic Programming: cf. for example (Lloyd 1987)). That is, if r is a rule of the form (1), and $\mathcal{G}$ is a set of RDF triples, then:

$$T_r(\mathcal{G}) = \{\mu(\mathcal{C}on) \mid \exists\mu \text{ such that } \mu(\mathcal{A}nte) \subseteq G\}$$

and accordingly $T_\mathcal{R}(\mathcal{G}) = \bigcup_{r\in\mathcal{R}} T_r(\mathcal{G})$. Also, let $\mathcal{G}_{i+1} = \mathcal{G}_i \cup T_\mathcal{R}(\mathcal{G}_i)$ and $\mathcal{G}_0 = \mathcal{G}$; we now define the *exhaustive application* of the $T_\mathcal{R}$ operator on a graph $\mathcal{G}$ as being upto the least fixpoint (the smallest value for n) such that $\mathcal{G}_n = T_\mathcal{R}(\mathcal{G}_n)$. We call $\mathcal{G}_n$ the *closure* of $\mathcal{G}$ with respect to ruleset $\mathcal{R}$, denoted as $\mathcal{C}l_\mathcal{R}(\mathcal{G})$. Note that we may also use the intuitive notation $T_\mathcal{R}(\mathbb{KB})$, $\mathcal{C}l_\mathcal{R}(\mathbb{KB})$ as shorthand for the more cumbersome $T_\mathcal{R}(\bigcup_{\mathcal{W}'\in\mathbb{KB}}\mathcal{W}')$, $\mathcal{C}l_\mathcal{R}(\bigcup_{\mathcal{W}'\in\mathbb{KB}}\mathcal{W}')$ respectively.

Ground Triple/Graph

A ground triple or ground graph is one without existential variables.

Herbrand Interpretation

Briefly, a *Herbrand interpretation* of a graph $\mathcal{G}$ treats URI references, blank nodes, typed literals and plain literals analogously as denoting their own syntactic form. As such, a Herbrand interpretation represents a ground view of an RDF graph where blank nodes are treated as *Skolem names* instead of existential variables; i.e., blank nodes are seen to represent the entities that they assert the existence of, analogously to a URI reference. Henceforth, we view blank nodes as their Skolem equivalents (this also applies to blank nodes as mentioned in the above notation) and only treat the ground case of RDF graphs.

Let us elaborate in brief why this treatment of blank nodes as Skolem constants is sufficient for our purposes. In our scenario, we perform forward-chaining materialisation for query-answering and not "real" entailment checks between RDF graphs. This enables us to treat all blank nodes as Skolem names (Hayes 2004). It is well known that simple entailment checking of two RDF graphs (Hayes 2004)—i.e., checking whether an RDF graph $\mathcal{G}_1$ entails $\mathcal{G}_2$—can be done using the ground "skolemised" version of $\mathcal{G}_1$. That is $\mathcal{G}_1 \models \mathcal{G}_2$ iff $sk(\mathcal{G}_1) \models \mathcal{G}_2$. Likewise, given a set of inference rules $\mathcal{R}$, where we denote entailment with respect to $\mathcal{R}$ as $\models_{\mathcal{R}}$, it is again well known that such entailment can be reduced to simple entailment with prior computation of the inference closure with respect to $\mathcal{R}$. That is, $\mathcal{G}_1 \models_{\mathcal{R}} \mathcal{G}_2$ iff $\mathcal{C}l_{\mathcal{R}}(sk(\mathcal{G}_1)) \models \mathcal{G}_2$; cf. (Gutiérrez, Hurtado et al. 2004; Hayes 2004). In this article we focus on the actual computation of $\mathcal{C}l_{\mathcal{R}}(sk(\mathcal{G}_1))$ for a tailored ruleset $\mathcal{R}$ in between RDFS and OWL Full.

PRAGMATIC INFERENCING FOR THE WEB

In this section we discuss the inference rules which we use to approximate OWL semantics and are designed for forward-chaining reasoning over web data. We justify our selection of inferences to support in terms of observed characteristics and examples taken from the Web. We optimise by restricting our fragment of reasoning according to three imperatives: *computational feasibility* (CF) for scalability, reduced output statements (RO) to ease the burden on consumer applications and, finally, web-tolerance (WT) for avoiding undesirable inferences given noisy data and protecting publishers from unwanted, independent third-party contributions. In particular, we adhere to the following high-level restrictions:

- we are incomplete (CF, RO, WT) - Section 3.1;
- we deliberately ignore the explosive behaviour of classical inconsistency (CF, RO, WT) - Section 3.1;
- we follow a rule-based, finite, forward-chaining approach to OWL inference (CF) - Section 3.2;
- we do not invent new blank nodes (CF, RO, WT) - Section 3.2;
- we avoid inference of extended-axiomatic triples (RO) - Section 3.2;
- we focus on inference of non-terminological statements (CF) - Section 3.2;
- we do not consider `:sameAs` statements as applying to terminological data (CF, WT) - Section 3.2;
- we separate and store terminological data in-memory (CF) - Section 3.3;
- we support limited reasoning for non-standard use of the RDF(S) and OWL vocabularies (CF, RO, WT) - Section 3.3
- we ignore non-authoritative (third-party) terminological statements from our reasoning procedure to counter an explosion of inferred statements caused by hijacking ontology terms (RO, WT) - Section 3.4.

Infeasibility of Complete Web Reasoning

Reasoning over RDF data is enabled by the description of RDF terms using the RDFS and OWL standards; these standards have defined entailments determined by their semantics. The semantics of these standards differs in that RDFS entailment is defined in terms of "if" conditions (intensional semantics), and has a defined set of complete standard entailment rules (Hayes 2004). OWL semantics uses "iff" conditions (extensional semantics) without a complete set of standard entailment rules. RDFS entailment has been shown to be decidable and in P for the ground case (ter Horst 2005), whilst OWL Full entailment is known to be undecidable (Horrocks and Patel-Schneider 2004). Thus, the OWL standard includes two restricted fragments of OWL whose entailment is known to be decidable from work in description logics: (i) OWL DL whose worst-case entailment is in NEXPTIME (ii) OWL Lite whose worst-case entailment is in EXPTIME (Horrocks and Patel-Schneider 2004).

Although entailment for both fragments is known to be decidable, and even aside from their complexity, most OWL ontologies crawlable on the Web are in any case OWL Full: idealised assumptions made in OWL DL are violated by even very commonly used ontologies. For example, the popular Friend Of A Friend (FOAF) vocabulary (Brickley and Miller 2007) deliberately falls into OWL Full since (i) in the FOAF RDF vocabulary[4], `foaf:name` is defined as a sub-property of the core RDFS property `rdfs:label` and (ii) `foaf:mbox_sha1sum` is defined as both an `:InverseFunctionalProperty` and a `:DatatypeProperty`: both are disallowed by OWL DL (and, of course, OWL Lite). In (Bechhofer and Volz 2004), the authors identified and categorised OWL DL restrictions violated by a sample group of 201 OWL ontologies (all of which were found to be in OWL Full); these include incorrect or missing typing of classes and properties, complex object-properties (e.g., functional properties) declared to be transitive, inverse-functional datatype properties, etc. In (Wang, Parsia et al. 2006), a more extensive survey with nearly 1,300 ontologies was conducted: 924 were identified as being in OWL Full. Taking into account that most web ontologies are in OWL Full, and also the undecidability/computational-infeasiblity of OWL Full, one could conclude that complete reasoning on the Web is impractical. However, again for most web documents only categorisable as OWL Full, infringements are mainly syntactic and are rather innocuous with no real effect on decidability ((Wang, Parsia et al. 2006) showed that the majority of web documents surveyed were in the base expressivity for Description Logics after patching infringements). The main justification for the infeasibility of complete reasoning on the Web is inconsistency.

Consistency cannot be expected on the Web; for instance, a past web-crawl of ours revealed the following:

- `#timbl a foaf:Person; foaf:homepage <http://w3.org/> .`
- `#w3c a foaf:Organization; foaf:homepage <http://w3.org/> .`
- `foaf:homepage a :InverseFunctionalProperty .`
- `foaf:Organization :disjointWith foaf:Person .`

These triples together infer that Tim Berners-Lee is the same as the W3C and thus cause an inconsistency.[5] Aside from such examples which arise from misunderstanding of the FOAF vocabulary, there might be cases where different parties deliberately make contradictive statements; resolution of such contradictions could involve "choosing sides". In any case, the explosive nature of contradiction in classical logics suggests that it is not desirable within our web reasoning scenario.

Rule-Based Web Reasoning

As previously alluded to, there does not exist a standard entailment for OWL suitable to our web reasoning scenario. However, incomplete (wrt. OWL Full) rule-based inference (i.e., reasoning as performed by logic progamming or deductive database engines) may be considered to have greater potential for scale, following the arguments made in (Fensel and van Harmelen 2007) and may be considered to be more robust with respect to preventing explosive inferencing through inconsistencies. Several rule expressible non-standard OWL fragments; namely OWL-DLP (Grosof, Horrocks et al. 2004), OWL⁻ (de Bruijn 2008) (which is a slight extension of OWL-DLP), OWLPrime (Wu, Eadon et al. 2008), pD* (ter Horst 2005; ter Horst 2005), and Intensional OWL (de Bruijn 2008); have been defined in the literature and enable incomplete but sound RDFS and OWL Full inferences.

In (ter Horst 2005), pD* was introduced as a combination of RDFS entailment, datatype reasoning and a distilled version of OWL with rule-expressible intensional semantics: pD* entailment maintains the computational complexity of RDFS entailment, which is in NP in general and P for the ground case. Such improvement in complexity has obvious advantages in our web reasoning scenario; thus SAOR's approach to reasoning is inspired by the pD* fragment to cover large parts of OWL by positive inference rules which can be implemented in a forward-chaining engine.

Table 1 summarises the pD* ruleset. The rules are divided into D*-entailment rules and P-entailment rules. D*-entailment is essentially RDFS entailment (Hayes 2004) combined with some datatype reasoning. P-entailment is introduced in (ter Horst 2005) as a set of rules which applies to a property-related subset of OWL.

Given pD*, we make some amendments so as to align the ruleset with our requirements. Table 2 provides a full listing of our own modified ruleset,

Table 1. Ter-Horst rules from (ter Horst 2005) in Turtle-like syntax

pD*	rule	where
	D*-**entailment rules**	
lg	?x ?P ?l . ⇒ ?v ?P _:*bl* .	?l$\in \mathcal{L}$[a]
gl	?x ?P _:*bl* . ⇒ ?x ?P ?l .	?l$\in \mathcal{L}$
rdf1	?x ?P ?y . ⇒ ?P a rdf:Property .	
rdf2-D	?x ?P ?l . ⇒ _:*bl* ?type ?t .	?l$= (s,t) \in \mathcal{L}_t$
rdfs1	?x ?P ?l . ⇒ _:*bl* a Literal .	?l$\in \mathcal{L}_p$
rdfs2	?P rdfs:domain ?C . ?x ?P ?y . ⇒ ?x a ?C .	
rdfs3	?P rdfs:range ?C . ?x ?P ?y . ⇒ ?y a ?C .	?y $\in \mathcal{U} \cup \mathcal{B}$
rdfs4a	?x ?P ?y . ⇒ ?x a rdfs:Resource .	
rdfs4b	?x ?P ?y . ⇒ ?y a rdfs:Resource .	?y$\in \mathcal{U} \cup \mathcal{B}$
rdfs5	?P rdfs:subProperty ?Q . ?Q rdfs:subProperty ?R . ⇒ ?P rdfs:subProperty ?R .	
rdfs6	?P a rdf:Property . ⇒ ?P rdfs:subProperty ?P .	
rdfs7	?P rdfs:subProperty ?Q . ?x ?P ?y . ⇒ ?x ?Q ?y .	?Q$\in \mathcal{U} \cup \mathcal{B}$
rdfs8	?C a rdfs:Class . ⇒ ?C rdfs:subClassOf rdfs:Resource .	
rdfs9	?C rdfs:subClassOf ?D . ?x a ?C . ⇒ ?x a ?D .	
rdfs10	?C a rdfs:Class . ⇒ ?C rdfs:subClassOf ?C .	
rdfs11	?C rdfs:subClassOf ?D . ?D rdfs:subClassOf ?E . ⇒ ?C rdfs:subClassOf ?E .	
rdfs12	?P a rdfs:ContainerMembershipProperty . ⇒ ?P rdfs:subPropertyOf rdfs:member .	
rdfs13	?D a rdfs:Datatype . ⇒ ?D rdfs:subClassOf rdfs:Literal .	
	***P*-entailment rules**	
rdfp1	?P a :FunctionalProperty . ?x ?P ?y , ?z . ⇒ ?y :sameAs ?z .	?y $\in \mathcal{U} \cup \mathcal{B}$
rdfp2	?P a :InverseFunctionalProperty . ?x ?P ?z . ?y ?P ?z . ⇒ ?x :sameAs ?y .	
rdfp3	?P a :SymmetricProperty . ?x ?P ?y . ⇒ ?y ?P ?x .	?y $\in \mathcal{U} \cup \mathcal{B}$
rdfp4	?P a :TransitiveProperty . ?x ?P ?y . ?y ?P ?z . ⇒ ?x ?P ?z .	
rdfp5a	?x ?P ?y . ⇒ ?x :sameAs ?x .	
rdfp5b	?x ?P ?y . ⇒ ?y :sameAs ?y .	?y$\in \mathcal{U} \cup \mathcal{B}$
rdfp6	?x :sameAs ?y . ⇒ ?y :sameAs ?x .	?y$\in \mathcal{U} \cup \mathcal{B}$
rdfp7	?x :sameAs ?y . ?y :sameAs ?z . ⇒ ?x :sameAs ?z .	
rdfp8a	?P :inverseOf ?Q . ?x ?P ?y . ⇒ ?y ?Q ?x .	?y,?Q$\in \mathcal{U} \cup \mathcal{B}$
rdfp8b	?P :inverseOf ?Q . ?x ?Q ?y . ⇒ ?y ?P ?x .	?y $\in \mathcal{U} \cup \mathcal{B}$
rdfp9	?C a :Class ; :sameAs ?D . ⇒ ?C rdfs:subClassOf ?D .	
rdfp10	?P a :Property ; :sameAs ?Q . ⇒ ?P rdfs:subPropertyOf ?Q .	
rdfp11	?x :sameAs ?_x . ?y :sameAs ?_y . ?x ?P ?y .⇒ ?_x ?P ?_y .	?_x $\in \mathcal{U} \cup \mathcal{B}$
rdfp12a	?C :equivalentClass ?D . ⇒ ?C rdfs:subClassOf ?D .	
rdfp12b	?C :equivalentClass ?D . ⇒ ?D rdfs:subClassOf ?C .	?D$\in \mathcal{U} \cup \mathcal{B}$
rdfp12c	?C rdfs:subClassOf ?D . ?D rdfs:subClassOf ?C . ⇒ ?C :equivalentClass ?D .	
rdfp13a	?P :equivalentProperty ?Q . ⇒ ?P rdfs:subPropertyOf ?Q .	
rdfp13b	?P :equivalentProperty ?Q . ⇒ ?Q rdfs:subPropertyOf ?P .	?Q$\in \mathcal{U} \cup \mathcal{B}$
rdfp13c	?P rdfs:subPropertyOf ?Q . ?Q rdfs:subPropertyOf ?P . ⇒ ?P :equivalentProperty ?Q .	
rdfp14a	?C :hasValue ?y ; :onProperty ?P . ?x ?P ?y . ⇒ ?x a ?C .	
rdfp14b	?C :hasValue ?y ; :onProperty ?P . ?x a ?C . ⇒ ?x ?P ?y .	?P$\in \mathcal{U} \cup \mathcal{B}$
rdfp15	?C :someValuesFrom ?D ; :onProperty ?P . ?x ?P ?y . ?y a ?D . ⇒ ?x a ?C .	
rdfp16	?C :allValuesFrom ?D ; :onProperty ?P . ?x a ?C; ?P ?y . ⇒ ?y a ?D .	?y$\in \mathcal{U} \cup \mathcal{B}$

[a] `_:bl` is a *surrogate blank node* given by an injective function on the literal `?l`

which we compare against pD* in this section. Note that this table highlights characteristics of the rules which we will discuss in Section 3.3 and Section 3.4; for the moment we point out that **rule′** is used to indicate an amendment to the respective pD* rule. Please also note that we use the notation rulex* to refer to all rules with the prefix **rulex.**

pD* Rules Directly Supported

From the set of pD* rules, we directly support rules **rdfs2, rdfs9, rdfp2, rdfp4, rdfp7,** and **rdfp17.**

pD* Omissions: Extended-Axiomatic Statements

We avoid pD* rules which specifically produce what we term *extended-axiomatic* statements mandated by RDFS and OWL semantics. Firstly, we do not infer the set of pD* axiomatic triples, which are listed in (ter Horst 2005) for RDF(S) and OWL respectively; according to pD*, these are inferred for the empty graph. Secondly, we do not materialise membership assertions for `rdfs:Resource` which would hold for every URI and blank node in a graph. Thirdly, we do not ma-

Table 2. Supported rules in Turtle-style syntax. Terminological patterns are underlined whereas assertional patterns are not; further, rules are grouped according to arity of terminological/assertional patterns in the antecedent. The source of a terminological pattern instance must speak authoritatively for at lest one boldface variable binding for the rule to fire.

SAOR	rule	where
	$\mathcal{R}0$: only terminological patterns in antecedent	
rdfc0	?C :oneOf (?x$_1$ … ?x$_n$) . ⇒ ?x$_1$ … ?x$_n$ a ?C .	$?C \in \mathcal{B}$
	$\mathcal{R}1$: at least one terminological/only one assertional pattern in antecedent	
rdfs2	**?P** rdfs:domain ?C . ?x ?P ?y . ⇒ ?x a ?C .	
rdfs3′	**?P** rdfs:range ?C . ?x ?P ?y . ⇒ ?y a ?C .	
rdfs7′	**?P** rdfs:subPropertyOf ?Q . ?x ?P ?y . ⇒ ?x ?Q ?y .	
rdfs9	**?C** rdfs:subClassOf ?D . ?x a ?C . ⇒ ?x a ?D .	
rdfp3′	**?P** a :SymmetricProperty . ?x ?P ?y . ⇒ ?y ?P ?x .	
rdfp8a′	**?P** :inverseOf ?Q . ?x ?P ?y . ⇒ ?y ?Q ?x .	
rdfp8b′	?P :inverseOf **?Q** . ?x ?Q ?y . ⇒ ?y ?P ?x .	
rdfp12a′	**?C** :equivalentClass ?D . ?x a ?C . ⇒ ?x a ?D .	
rdfp12b′	?C :equivalentClass **?D** . ?x a ?D . ⇒ ?x a ?C .	
rdfp13a′	**?P** :equivalentProperty ?Q . ?x ?P ?y . ⇒ ?y ?Q ?x .	
rdfp13b′	?P :equivalentProperty **?Q** . ?x ?Q ?y . ⇒ ?y ?P ?x .	
rdfp14a′	?C :hasValue ?y ; :onProperty **?P** . ?x ?P ?y . ⇒ ?x a ?C .	$?C \in \mathcal{B}$
rdfp14b′	?C :hasValue ?y ; :onProperty ?P . ?x a ?C . ⇒ ?x ?P ?y .	$?C \in \mathcal{B}$
rdfc1	?C :unionOf (?C$_1$…**?C$_i$**…?C$_n$) . ?x a ?C$_i$[a] . ⇒ ?x a ?C .	$?C \in \mathcal{B}$
rdfc2	?C :minCardinality 1 ; :onProperty **?P** . ?x ?P ?y . ⇒ ?x a ?C .	$?C \in \mathcal{B}$
rdfc3a	?C :intersectionOf (?C$_1$ … ?C$_n$) . ?x a ?C . ⇒ ?x a ?C$_1$, …, ?C$_n$.	$?C \in \mathcal{B}$
rdfc3b	?C :intersectionOf (**?C$_1$**) . ?x a ?C$_1$. ⇒ ?x a ?C . [b]	$?C \in \mathcal{B}$
	$\mathcal{R}2$: at least one terminological/multiple assertional patterns in antecedent	
rdfp1′	**?P** a :FunctionalProperty . ?x ?P ?y , ?z . ⇒ ?y :sameAs ?z .	
rdfp2	**?P** a :InverseFunctionalProperty . ?x ?P ?z . ?y ?P ?z . ⇒ ?x :sameAs ?y .	
rdfp4	**?P** a :TransitiveProperty . ?x ?P ?y . ?y ?P ?z . ⇒ ?x ?P ?z .	
rdfp15′	?C :someValuesFrom **?D** ; :onProperty **?P** . ?x ?P ?y . ?y a ?D . ⇒ ?x a ?C .	$?C \in \mathcal{B}$
rdfp16′	?C :allValuesFrom ?D ; :onProperty ?P . ?x a ?C ; ?P ?y . ⇒ ?y a ?D .	$?C \in \mathcal{B}$
rdfc3c	?C :intersectionOf (**?C$_1$ … ?C$_n$**) . ?x a ?C$_1$, …, ?C$_n$. ⇒ ?x a ?C .	$?C \in \mathcal{B}$
rdfc4a	?C :cardinality 1 ; :onProperty ?P . ?x a ?C ; ?P ?y , ?z . ⇒ ?y :sameAs ?z .	$?C \in \mathcal{B}$
rdfc4b	?C :maxCardinality 1 ; :onProperty ?P . ?x a ?C ; ?P ?y , ?z . ⇒ ?y :sameAs ?z .	$?C \in \mathcal{B}$
	$\mathcal{R}3$: only assertional patterns in antecedent	
rdfp6′	?x :sameAs ?y . ⇒ ?y :sameAs ?x .	
rdfp7	?x :sameAs ?y . ?y :sameas ?z . ⇒ ?x :sameAs ?z .	
rdfp11′	?x :sameAs ?_x ; ?P ?y .⇒ ?_x ?P ?y . [c]	
rdfp11″	?y :sameAs ?_y . ?x ?P ?y .⇒ ?x ?P ?_y . [c]	

[a] $?C_i \in \{?C_1, ..., ?C_n\}$
[b] **rdfs3b** is a special case of **rdfs3c** with one A-Box pattern and thus falls under $\mathcal{R}1$.
[c] Only where **?p** is not an RDFS/OWL property used in any of our rules (see $\mathcal{P}_{SAOR}$, Section 3.3)

terialise reflexive `:sameAs` membership assertions, which again hold for every URI and blank node in a graph. We see such statements as inflationary and orthogonal to our aim of reduced output.

pD* Amendments: :sameAs Inferencing

From the previous set of omissions, we do not infer reflexive `:sameAs` statements. However, such reflexive statements are required by pD* rule **rdfp11.** We thus fragment the rule into **rdfp11′** and **rdfp11″** which allows for the same inferencing without such reflexive statements.

In a related issue, we wittingly do not allow `:sameAs` inferencing to interfere with terminological data: for example, we do not allow `:sameAs` inferencing to affect properties in the predicate position of a triple or classes in the object position of an `rdf:type` triple. In (Hogan, Harth et al. 2007) we showed that `:sameAs` inferencing through `:InverseFunctionalProperty` reasoning caused fallacious equalities to be asserted due to noisy web data. This is the primary motivation for us also omitting rules **rdfp9, rdfp10** and the reason why we place the restriction on `?p` for our rule **rdfp11″;** we do not want noisy equality inferences to be reflected in the terminological segment of our knowledge-base, nor to affect the class and property positions of membership assertions.

pD* Omissions: Terminological Inferences

From pD*, we also omit rules which infer only terminological statements: namely **rdf1, rdfs5, rdfs6, rdfs8, rdfs10, rdfs11, rdfs12, rdfs13, rdfp9, rdfp10, rdfp12*** and **rdfp13*.** As such, our use-case is query-answering over assertional data; we therefore focus in this article on materialising assertional data.

We have already motivated omission of inference through `:sameAs` rules **rdfp9** and **rdfp10.** Rules **rdf1, rdfs8, rdfs12** and **rdfs13** infer memberships of, or subclass/subproperty relations to, RDF(S) classes and properties; we are not interested in these primarily syntactic statements which are not directly used in our inference rules. Rules **rdfs6** and **rdfs10** infer reflexive memberships of `rdfs:subPropertyOf` and `rdfs:subClassOf` meta-properties which are used in our inference rules; clearly however, these reflexive statements will not lead to unique assertional inferences through related rules **rdfs7′ or rdfs9** respectively. Rules **rdfs5 and rdfs1**1 infer transitive memberships again of `rdfs:subPropertyOf` and `rdfs:subClassOf`; again however, exhaustive application of rules **rdfs7′ or rdfs9** respectively ensures that all possible assertional inferences are materialised without the need for the transitive rules. Rules **rdfp12c and rdfp13c** infer additional `:equivalentClass`/`:equivalentProperty` statements from `rdfs:subClassOf`/ `rdfs:subPropertyOf` statements where assertional inferences can instead be conducted through two applications each of rules **rdfs9** and **rdfs7′** respectively.

pD* Amendments: Direct Assertional Inferences

The observant reader may have noticed that we did not dismiss inferencing for rules **rdfp12a,rdfp12b/ rdfp13a,rdfp13b** which translate `:equivalentClass`/`:equivalentProperty` to `rdfs:subClassOf`/ `rdfs:subPropertyOf`. In pD*, these rules are required to support indirect assertional inferences through rules **rdfs9** and **rdfs7** respectively; we instead support assertional inferences directly from the `:equivalentProperty`/`:equivalentClass` statements using symmetric rules **rdfp12a′,rdfp12b′/ rdfp13a′,rdfp13b′.**

pD* Omissions: Existential Variables in Consequent

We avoid rules with existential variables in the consequent; such rules would require adaptation of the T_r operator so as to “invent” new blank nodes for each rule application, with undesireable effects for forward-chaining reasoning regarding termination. For example, like pD*, we only support inferences in one direction for `:someValuesFrom` and avoid a rule such as:

```
>    ?C :someValuesFrom ?D ; :onProperty ?P .
     ?x a ?C ⇒ ?x ?P _:b . _:b a ?D .
```

Exhaustive application of the rule to, for example, the following data (more generally where `?D` is a subclass of `?C`):

```
>    ?ex:Person rdfs:subClassOf [:someValues-
     From ex:Person ; :onProperty ex:mother
     .]
>    ? _:Tim a ex:Person .
```

would infer infinite triples of the type:

```
>    ? _:Tim ex:mother _:b0 .
>    ? _:b0 a ex:Person ; ex:mother _:b1 .
>    ? _:b1 a ex:Person ; ex:mother _:b2 .
...
```

In fact, this rule is listed in (ter Horst 2005) as **rdf-svx** which forms an extension of pD* entailment called *pD*sv*. This rule is omitted from pD* and from SAOR due to obvious side-effects on termination and complexity.

Unlike pD*, we also avoid inventing so called "surrogate" blank nodes for the purposes of representing a literal in intermediary inferencing steps (Rules **lg, gl, rdf2-D, rdfs1** in **RDFS/D*** entailment). Thus, we also do not support datatype reasoning (Rule **rdf2-D)** which involves the creation of surrogate blank nodes. Although surrogate blank nodes are created according to a direct mapping from a finite set of literals (and thus, do not prevent termination), we view "surrogate statements" as inflationary.

pD* Amendments: Relaxing Literal Restrictions

Since we do not support surrogate blank nodes as representing literals, we instead relax restrictions placed on pD* rules. In pD*, blank nodes are allowed in the predicate position of triples; however, the restriction on literals in the subject and predicate position still applies: literals are restricted from travelling to the subject or predicate position of a consequent (see *where column*, Table 1). Thus, surrogate blank nodes are required in pD* to represent literals in positions where they would otherwise not be allowed.

We take a different approach whereby we allow literals directly in the subject and predicate position for intermediate inferences. Following from this, we remove pD* literal restrictions on rules **rdfs3, rdfs7, rdfp1, rdfp3, rdfp6, rdfp8*, rdfp14b, rdfp16** for intermediate inferences and omit any inferred **non-RDF** statements from being written to the final output.

Additions to pD*

In addition to pD*, we also include some "class based entailment" from OWL, which we call C-entailment. We name such rules using the **rdfc* stem,** following the convention from **P-entailment.** We provide limited support for enumerated classes (**rdfc0**), union class descriptions (**rdfc1**), intersection class descriptions (**rdfc3***),[6] as well as limited cardinality constraints (**rdfc2, rdfc4*).**

pD* Amendments: Enforcing OWL Abstract Syntax Restrictions

Finally, unlike pD*, we enforce blank nodes as mandated by the OWL Abstract Syntax (Patel-Schneider and Horrocks 2004), wherein certain abstract syntax constructs (most importantly in our case: *unionOf(description$_1$...*

description$_n$), intersectionOf(description$_1$... description$_n$), oneOf(iID$_1$...iID$_n$), restriction(ID allValuesFrom(range)), restriction(ID someValuesFrom(required)), restriction(ID value(value)), restriction(ID maxCardinality(max)), restriction(ID minCardinality(min)), restriction(ID cardinality(card)) and *SEQ item$_1$...item$_n$)* are strictly mapped to RDF triples with blank nodes enforced for certain positions: such mapping is necessitated by the idiosyncrasies of representing OWL in RDF. Although the use of URIs in such circumstances is allowed by RDF, we enforce the use of blank nodes for terminological patterns in our ruleset; to justify, let us look at the following problematic example of OWL triples taken from two sources:

```
>    SOURCE <ex:>
>    ex:Person :onProperty ex:parents ; :som-
     eValuesFrom ex:Person .
>    SOURCE <ex2:>
>    ex:Person :allValuesFrom ex2:Human .
```

According to the abstract syntax mapping, neither of the restrictions should be identified by a URI (if blank nodes were used instead of `ex:Person` as mandated by the abstract syntax, such a problem could not occur as each web-graph is given a unique set of blank nodes). If we consider the RDF-merge of the two graphs, we will be unable to distinguish which restriction the `:onProperty` value applies to. As above, allowing URIs in these positions would enable "syntactic interference" between data sources. Thus, in our ruleset, we always enforce blank-nodes as mandated by the OWL abstract syntax; this specifically applies to pD* rules **rdfp14*, rdfp15′** and **rdfp16′** and to all of our **C-entailment** rules **rdfc*.** We denote the restrictions in the where column of Table 2. Indeed, in our treatment of terminological collection statements, we enforced blank nodes in the subject position of `rdf:first`/`rdf:rest` membership assertions, as well as blank nodes in the object position of non-terminating `rdf:rest` statements; these are analogously part of the OWL abstract syntax restrictions.

Separation of T-Box from A-Box

Aside from the differences already introduced, our primary divergence from the pD* fragment and traditional rule-based approaches is that we separate terminological data from assertional data according to their use of the RDF(S) and OWL vocabulary; these are commonly known as the "T-Box" and "A-Box" respectively (loosely borrowing Description Logics terminology). In particular, we require a separation of T-Box data as part of a core optimisation of our approach; we wish to perform a once-off load of T-Box data from our input knowledge-base into main memory.

Let $\mathcal{P}_{SAOR}$ and $\mathcal{C}_{SAOR}$ and be, resp., the exact set of RDF(S)/OWL meta-properties and -classes used in our inference rules; viz. $\mathcal{P}_{SAOR}$ = { `rdfs:domain`, `rdfs:range`, `rdfs:subClassOf`, `rdfs:subPropertyOf`, `:allValuesFrom`, `:cardinality`, `:equivalentClass`, `:equivalentProperty`, `:hasValue`, `:intersectionOf`, `:inverseOf`, `:maxCardinality`, `:minCardinality`, `:oneOf`, `:onProperty`, `:sameAs`, `:someValuesFrom`, `:unionOf` } & $\mathcal{C}_{SAOR}$ = { `:FunctionalProperty`, `:InverseFunctionalProperty`, `:SymmetricProperty`, `:TransitiveProperty` }; our T-Box is a set of terminological triples restricted to only include membership assertions for $\mathcal{P}_{SAOR}$ and $\mathcal{C}_{SAOR}$ and the set of terminological collection statements. Table 2 identifies T-Box patterns by underlining. Statements from the input knowledge-base that match these patterns are all of the T-Box statements we consider in our reasoning process: inferred statements or statements that do not match one of these patterns are not considered being part of the T-Box, but are treated purely as assertional. We now define our T-Box:

Definition 1 (T-Box) *Let $\mathcal{T}_{\mathcal{G}}$ be the union of all graph pattern instances from a graph $\mathcal{G}$ for a* *terminological (underlined) graph pattern in Table 2; i.e., $\mathcal{T}_{\mathcal{G}}$ is itself a graph. We call $\mathcal{T}_{\mathcal{G}}$ the T-Box of $\mathcal{G}$.*

Also, let $\mathcal{P}^{domP}_{SAOR}$ = {`rdfs:domain`, `rdfs:range`, `rdfs:subPropertyOf`, `:equivalentProperty`, `:inverseOf` } and $\mathcal{P}^{ranP}_{SAOR}$ = { `rdfs:subPropertyOf`, `:equivalentProperty`, `:inverseOf`, `:onProperty`}, We call ϕ a property in T-Box $\mathcal{T}$ if there exists a triple $t \in \mathcal{T}$ where

- $s = \phi$ and $p \in \mathcal{P}^{domP}_{SAOR}$
- $p \in \mathcal{P}^{ranP}_{SAOR}$ and $o = \phi$
- $s = \phi$, $p =$ `rdf:type` and $o \in \mathcal{C}_{SAOR}$

Similarly, let $\mathcal{P}^{domC}_{SAOR}$ = {`rdfs:subClassOf`, `:allValuesFrom`, `:cardinality`, `:equivalentClass`, `:hasValue`, `:intersectionOf`, `:maxCardinality`, `:minCardinality`, `:oneOf`, `:onProperty`, `:someValuesFrom`, `:unionOf`} and $\mathcal{P}^{ranC}_{SAOR}$ = {`rdfs:domain`, `rdfs:range`, `rdfs:subClassOf`, `rdf:first`, `:allValuesFrom`, `:equivalentClass`, `:someValuesFrom`}. We call $\mathcal{X}$ a class in T-Box $\mathcal{T}$ if there exists a triple $t \in \mathcal{T}$ where

- $p \in \mathcal{P}^{domC}_{SAOR}$ *and* $s = \mathcal{X}$
- $p \in \mathcal{P}^{ranC}_{SAOR}$ *and* $o = \mathcal{X}$

We define the signature of a T-Box $\mathcal{T}$ to be the set of all properties and classes in $\mathcal{T}$ as above, which we denote by $sig(\mathcal{T})$.

For our knowledge-base $\mathbb{KB}$, we define our T-Box $\mathbb{T}$ as the set of all pairs $(\mathcal{T}_{\mathcal{W}'}, c)$ where $(\mathcal{W}', c) \in \mathbb{KB}$, and $\mathcal{T}_{\mathcal{W}'} \neq \emptyset$. Again, we may use the intuitive notation $\mathcal{T}_{\mathcal{W}'} \in \mathbb{T}$. We define our A-Box $\mathbb{A}$ as containing all of the statements in $\mathbb{KB}$, including $\mathbb{T}$ and the set of class and property membership assertions possibly using identifiers in $\mathcal{P}_{SAOR} \cup \mathcal{C}_{SAOR}$; i.e., unlike description logics, our $\mathbb{A}$ is synonymous with our $\mathbb{KB}$. We use the term A-Box to distinguish data that are stored on-disk (which includes T-Box data also stored in memory).

We now define our notion of a $\mathcal{T}$-split inference rule, whereby part of the antecedent is a basic graph pattern strictly instantiated by a static T-Box $\mathcal{T}$.

Definition 2 ($\mathcal{T}$-split inference rule) *We define a $\mathcal{T}$-split inference rule r as the triple $(\mathcal{A}nte_{\mathcal{T}}, \mathcal{A}nte_{\mathcal{G}}, \mathcal{C}on)$, where $\mathcal{A}nte_{\mathcal{T}}$ is a basic graph pattern matched by a static T-Box $\mathcal{T}$ and $\mathcal{A}nte_{\mathcal{G}}$ is matched by data in the graph $\mathcal{G}$, $\mathcal{C}on$ does not contain blank nodes, $\mathcal{V}(\mathcal{C}on) \neq \emptyset$, $\mathcal{V}(\mathcal{C}on) \subseteq \mathcal{V}(\mathcal{A}nte_{\mathcal{T}}) \cup \mathcal{V}(\mathcal{A}nte_{\mathcal{G}})$; also, if both $\mathcal{A}nte_{\mathcal{T}}$ and $\mathcal{A}nte_{\mathcal{G}}$ are non-empty, then $\mathcal{V}(\mathcal{A}nte_{\mathcal{T}}) \cap \mathcal{V}(\mathcal{A}nte_{\mathcal{G}}) \neq \emptyset$.*

We generally write $(\mathcal{A}nte_{\mathcal{T}}, \mathcal{A}nte_{\mathcal{G}}, \mathcal{C}on)$ as $\underline{\mathcal{A}nte_{\mathcal{T}}}\mathcal{A}nte_{\mathcal{G}} \Rightarrow \mathcal{C}on$. We call $\mathcal{A}nte_{\mathcal{T}}$ the terminological or T-Box antecedent pattern and $\mathcal{A}nte_{\mathcal{G}}$ the assertional or A-Box pattern.

Definition 3 (Rule-sets $\mathcal{R}_{\mathcal{T}}$, $\mathcal{R}_{\mathcal{TG}}$, $\mathcal{R}_{\mathcal{G}}$) *We define $\mathcal{R}_{\mathcal{T}}$ as the set of $\mathcal{T}$-split rules for which $\mathcal{A}nte_{\mathcal{T}} \neq \emptyset$ and $\mathcal{A}nte_{\mathcal{G}} = \emptyset$. We define $\mathcal{R}_{\mathcal{TG}}$ as the set of $\mathcal{T}$-split rules for which $\mathcal{A}nte_{\mathcal{T}} \neq \emptyset$ and $\mathcal{A}nte_{\mathcal{G}} \neq \emptyset$. We define $\mathcal{R}_{\mathcal{G}}$ as the set of $\mathcal{T}$-split rules for which $\mathcal{A}nte_{\mathcal{T}} \neq \emptyset$ and $\mathcal{A}nte_{\mathcal{G}} \neq \emptyset$.*

In Table 2, we categorise the $\mathcal{T}$-split rules into four rulesets: $\mathcal{R}0 \subset \mathcal{R}_{\mathcal{T}}$; $\mathcal{R}1 \subset \mathcal{R}_{\mathcal{TG}}$ where $|\mathcal{A}nte_{\mathcal{G}}| = 1$; $\mathcal{R}2 \subset \mathcal{R}_{\mathcal{TG}}$ where $|\mathcal{A}nte_{\mathcal{G}}| > 1$ and $\mathcal{R}0 \subset \mathcal{R}_{\mathcal{G}}$.

We now introduce the notion of a *$\mathcal{T}$-split inference* rule application for a graph $\mathcal{G}$ w.r.t. a *T-Box $\mathcal{T}$:*

Definition 4 ($\mathcal{T}$-split inference rule application) *We define a $\mathcal{T}$-split rule application to be $\mathcal{T}_r(\mathcal{T}, \mathcal{G})$ for $r = (\mathcal{A}nte_{\mathcal{T}}, \mathcal{A}nte_{\mathcal{G}}, \mathcal{C}on)$ as follows:*

$T_r(\mathcal{T}, \mathcal{G}) = \{\mu(\mathcal{C}on) | \exists\mu$ such that $\mu(\mathcal{A}nte_{\mathcal{T}}) \subseteq T$ and $\mu(\mathcal{A}nte_{\mathcal{G}}) \subseteq G\}$

Again, $\mathcal{T}_{\mathcal{R}}(\mathcal{T}, \mathcal{G}) = \bigcup_{r \in \mathcal{R}} \mathcal{T}_r(\mathcal{T}, \mathcal{G})$; also, given $\mathcal{T}$ as static, the exhaustive application of the $\mathcal{T}_{\mathcal{R}}(\mathcal{T}, \mathcal{G})$ up to the least fixpoint is called the *$\mathcal{T}$-split* closure of $\mathcal{G}$, denoted as $\mathcal{C}l_{\mathcal{R}}(\mathcal{T}, \mathcal{G})$. Again we use abbreviations such as $\mathcal{T}_{\mathcal{R}}(\mathbb{T}, \mathbb{KB})$ and $\mathcal{Cl}_{\mathcal{R}}(\mathbb{T}, \mathbb{KB})$, where $\mathbb{KB}$ should be interpreted as $\bigcup_{\mathcal{W}' \in \mathbb{KB}} \mathcal{W}'$ and $\mathbb{T}$ as $\bigcup_{\mathcal{T}_{\mathcal{W}'} \in \mathbb{T}} \mathcal{T}_{\mathcal{W}'}$.

Please note that since we enforce blank nodes in all positions mandated by the OWL abstract syntax for our rules, each instance of a given graph pattern $\mathcal{A}nte_{\mathcal{T}}$ can only contain triples from one web-graph $\mathcal{W}'$ where $\mathcal{T}_{\mathcal{W}'} \in \mathbb{T}$. Let $\mathcal{V}_{\mathcal{B}}(\mathcal{GP})$ be the set of all variables in a graph pattern $\mathcal{GP}$ which we restrict to only be instantiated by a blank node according to the abstract syntax. For all $\mathcal{A}nte_{\mathcal{T}}$ in our rules where $|\mathcal{A}nte_{\mathcal{T}}| > 1$ let $\mathcal{A}nte_{\overline{\mathcal{T}}}$ be any proper non-empty subset of $\mathcal{A}nte_{\mathcal{T}}$; we can then say that $\mathcal{V}_{\mathcal{B}}(\mathcal{A}nte_{\overline{\mathcal{T}}}) \cap \mathcal{V}_{\mathcal{B}}(\mathcal{A}nte_{\mathcal{T}} \setminus \mathcal{A}nte_{\overline{\mathcal{T}}}) \neq \emptyset$. In other words, since for every rule either (i) $\mathcal{A}nte_{\mathcal{T}} = \emptyset$; or (ii) $\mathcal{A}nte_{\mathcal{T}}$ consists of a single triple pattern; or (iii) no sub-pattern of any $\mathcal{A}nte_{\mathcal{T}}$ in our rules contains a unique set of blank-node enforced variables; then a given instance of $\mathcal{A}nte_{\mathcal{T}}$ can only contain triples from one web-graph with unique blank nodes as is enforced by our knowledge-base. For our ruleset, we can then say that $T_R(\mathbb{T}, \mathbb{KB}) = T_R(\bigcup_{\mathcal{T}_{\mathcal{W}'} \in \mathbb{T}} \mathcal{T}_{\mathcal{W}'}, \mathbb{KB}) = \bigcup_{\mathcal{T}_{\mathcal{W}'} \in \mathbb{T}} T_R(\mathcal{T}_{\mathcal{W}'}, \mathbb{KB})$. In other words, one web-graph cannot re-use structural statements in another web-graph to instantiate a T-Box pattern in our rule; this has bearing on our notion of authoritative reasoning which will be highlighted at the end of Section 3.4.

Further, a separate static T-Box within which inferences are not reflected has implications upon the completeness of reasoning w.r.t. the presented ruleset. Although, as presented in Section 3.2, we do not infer terminological statements and thus can support most inferences directly from our static T-Box, SAOR still does not fully support meta-modelling (Motik 2007): by separating the T-Box segment of the knowledge-base, we do not support all possible entailments from the simultaneous description of both a class (or property) and an individual. In other words, we do not fully support inferencing for meta-classes or meta-properties defined outside of the RDF(S)/OWL specification.

However, we do provide limited reasoning support for meta-modelling in the spirit of "punning" by conceptually separating the individual-, class- or property-meanings of a resource; c.f. (Grau, Horrocks et al. 2006). More precisely, during reasoning we not only store the T-Box data in memory, but also store the data on-disk in the A-Box. Thus, we perform punning in one direction: viewing class and property descriptions which form our T-Box also as individuals in our A-Box. Interestingly, although we do not support terminological reasoning directly, we can through our limited punning perform reasoning for terminological data based on the RDFS descriptions provided for the RDFS and OWL specifications. For example, we would infer the following by storing the three input statements in both the T-Box and the A-Box:

```
> rdfs:subClassOf rdfs:domain rdfs:Class;
  rdfs:range rdfs:Class .
> ex:Class1 rdfs:subClassOf ex:Class2 . ⇒
> ex:Class1 a rdfs :Class .
  ex:Class2 a rdfs:Class .
```

However, again our support for meta-modelling is limited; SAOR does not fully support so-called "non-standard usage" of RDF(S) and OWL: the use of properties and classes which make up the RDF(S) and OWL vocabularies in locations where they have not been intended, cf. (de Bruijn and Heymans 2007; Muñoz, Pérez et al. 2007). We adapt and refine the definition of non-standard vocabulary use for our purposes according to the parts of the RDF(S) and OWL vocabularies relevant for our inference ruleset:

Definition 5 (Non-standard Vocabulary Usage) *An RDF triple t has non-standard vocabulary usage for our ruleset if one of the following conditions holds:*

- a property in $\mathcal{P}_{SAOR}$ appears in a position different from the predicate position; or

- a class in $\mathcal{C}_{SAOR}$ appears in a position different from the object position of an `rdf:type` triple.

Continuing, we now introduce the following example wherein the first input statement is a case of non-standard usage with `rdfs:subClassOf` $\in \mathcal{P}_{SAOR}$ in the object position:[7]

> `ex:subClassOf rdfs:subPropertyOf rdfs:subClassOf .`
> `ex:Class1 ex:subClassOf ex:Class2 .` ⇒
> `ex:Class1 rdfs:subClassOf ex:Class2 .`

We can see that SAOR provides inference through `rdfs:subPropertyOf` as per usual; however, the inferred triple will not be reflected in the T-Box, thus we are incomplete and will not translate members of `ex:Class1` into `ex:Class2`. As such, non-standard usage may result in T-Box statements being produced which, according to our limited form of punning, will not be reflected in the T-Box and will lead to incomplete inference.

Indeed, there may be good reason for not fully supporting non-standard usage of the ontology vocabulary: non-standard use could have unpredictable results even under our simple rule-based entailment if we were to fully support meta-modelling. One may consider a finite combination of only four non-standard triples that, upon naive reasoning, would explode all web resources R by inferring $|R|^3$ triples, namely:

> `rdfs:subClassOf rdfs:subPropertyOf rdfs:Resource .`
> `rdfs:subClassOf rdfs:subPropertyOf rdfs:subPropertyOf .`
> `rdf:type rdfs:subPropertyOf rdfs:subClassOf .`
> `rdfs:subClassOf rdf:type :SymmetricProperty .`

The exhaustive application of standard RDFS inference rules plus inference rules for property symmetry together with the inference for class membership in `rdfs:Resource` for all collected resources in typical rulesets such as pD* lead to inference of any possible triple (r_1, r_2, r_3) for arbitrary $r_1, r_2, r_3 \in R$. Thus, although by maintaining a separate static T-Box we are incomplete w.r.t non-standard usage, we show that complete support of such usage of the RDFS/OWL vocabularies is undesirable for the Web.[8]

Authoritative Reasoning against Ontology Hijacking

During initial evaluation of a system which implements reasoning upon the above ruleset, we encountered a behaviour which we term "ontology hijacking", symptomised by a perplexing explosion of materialised statements. For example, we noticed that for a single `foaf:Person` membership assertion, SAOR inferred in the order of hundreds of materialised statements as opposed to the expected six. Such an explosion of statements is orthogonal to the aim of reduced materialised statements we have outlined for SAOR; thus, SAOR is designed to annul the diagnosed problem of ontology hijacking through anaylsis of the authority of web sources for T-Box data. Before formally defining ontology hijacking and our proposed solution, let us give some preliminary definitions:

Definition 6 (Authoritative Source) *A web-graph $\mathcal{W}$ from source (context) c speaks authoritatively about an RDF term n iff:*

- $n \in \mathcal{B}$; or
- $n \in \mathcal{U}$ and c coincides with, or is redirected to by, the namespace[9] of n.

Firstly, all graphs are authoritative for blank nodes defined in that graph (remember that according to the definition of our knowledge-base, all blank nodes are unique to a given graph). Secondly, we support namespace redirects so as to conform to best practices as currently adopted by web ontology publishers.[10]

For example, as taken from the Web:

- Source `http://usefulinc.com/ns/doap` is authoritative for all classes and properties which are within the `http://usefulinc.com/ns/doap` namespace; e.g., `http://usefulinc.com/ns/doap#Project`.
- Source `http://xmlns.com/foaf/spec/` is authoritative for all classes and properties which are within the `http://xmlns.com/foaf/0.1/` namespace; e.g., `http://xmlns.com/foaf/0.1/knows`; since the property `http://xmlns.com/foaf/0.1/knows` redirects to `http://xmlns.com/foaf/spec/`.

We consider the authority of sources speaking about classes and properties in our T-Box to counter-act ontology hijacking; ontology hijacking is the assertion of a set of non-authoritative T-Box statements such that could satisfy the terminological pattern of a rule in $\mathcal{R}_{\mathcal{TG}}$ (i.e., those rules with at least one terminological and at least one assertional triple pattern in the antecedent). Such third-party sources can then cause arbitrary inferences over membership assertions of classes or properties (contained in the A-Box) for which they speak non-authoritatively. We can say that only rules in $\mathcal{R}_{\mathcal{TG}}$ are relevant to ontology hijacking since: (i) inferencing on $\mathcal{R}_{\mathcal{G}}$, which does not contain any T-Box patterns, cannot be affected by non-authoritative T-Box statements; and (ii) the $\mathcal{R}_{\mathcal{T}}$ ruleset does not contain any assertional antecedent patterns and therefore, cannot hijack assertional data (i.e., in our scenario, the `:oneOf` construct can be viewed as directly asserting memberships, and is unable, according to our limited support, to redefine sets of individuals). We now define ontology hijacking:

Definition 7 (Ontology Hijacking) *Let $\mathcal{T}_{\mathcal{W}}$ be the T-Box extracted from a web-graph $\mathcal{W}$ and let $\widehat{sig}(\mathcal{W})$ be the set of classes and properties for which $\mathcal{W}$ speaks authoritatively; then if $\mathcal{C}l_{\mathcal{R}_{\mathcal{TG}}}(\mathcal{T}_{\mathcal{W}}, \mathcal{G}) \neq \mathcal{G}$ for any $\mathcal{G}$ not mentioning any element of $\widehat{sig}(\mathcal{W})$, we say that web-graph $\mathcal{W}$ is performing ontology hijacking.*

In other words, ontology hijacking is the contribution of statements about classes or properties in a non-authoritative source such that reasoning on members of those classes or properties is affected. One particular method of ontology hijacking is defining new super-classes or properties of third-party classes or properties. As a concrete example, if one were to publish today a description of a property in an ontology (in a location non-authoritative for `foaf:` but authoritative for `my:`), `my:name`, within which the following was stated: `foaf:name rdfs:subPropertyOf my:name .`, that person would be hijacking the `foaf:name` property and effecting the translation of all `foaf:name` statements in the web knowledge-base into `my:name` statements as well. However, if the statement were instead `my:name rdfs:subPropertyOf foaf:name .`, this would not constitute a case of ontology hijacking but would be a valid example of translating from a local authoritative property into an external non-authoritative property.

Ontology hijacking is problematic in that it vastly increases the amount of statements that are materialised and can potentially harm inferencing on data contributed by other parties. With respect to materialisation, the former issue becomes prominent: members of classes/properties from popular/core ontologies get translated into a plethora of conceptual models described in obscure ontologies; we quantify the problem in Section 5. However, taking precautions against harmful ontology hijacking is growing more and more important as the Semantic Web features more and more attention; motivation for spamming and other malicious activity propagates amongst certain parties with ontology hijacking being a prospective avenue. With this in mind, we assign sole responsibility for classes and properties and reasoning upon their members to those who maintain the authoritative specification.

Related to the idea of ontology hijacking is the idea of "non-conservative extension" described in the Description Logics literature: cf. (Ghilardi, Lutz et al. 2006; Lutz, Walther et al. 2007; Jiménez-Ruiz, Grau et al. 2008). However, the notion of a "conservative extension" was defined with a slightly different objective in mind: according to the notion of deductively conservative extensions, a graph $\mathcal{G}_a$ is only considered malicious towards $\mathcal{G}_b$ if it causes additional inferences with respect to the intersection of the signature of the original $\mathcal{G}_b$ with the newly inferred statements. Returning to the former `my:name` example from above, defining a super-property of `foaf:name` would still constitute a conservative extension: the closure without the non-authoritative `foaf:name rdfs:subPropertyOf my:name .` statement is the same as the closure with the statement after all of the `my:name` membership assertions are removed. However, further stating that `my:name a :InverseFunctionalProperty.` would not satisfy a model conservative extension since members of `my:name` might then cause equalities in other remote ontologies as side-effects, independent from the newly defined signature. Summarising, we can state that every non-conservative extension (with respect to our notion of deductive closure) constitutes a case of ontology hijacking, but not vice versa; non-conservative extension can be considered "harmful" hijacking whereas the remainder of ontology hijacking cases can be considered "inflationary".

To negate ontology hijacking, we only allow inferences through *authoritative rule applications*, which we now define:

Definition 8 (Authoritative Rule Application) *Again let $\widehat{sig}(\mathcal{W})$ be the set of classes and properties for which $\mathcal{W}$ speaks authoritatively and let $\mathcal{T}_{\mathcal{W}}$ be the T-Box of $\mathcal{W}$. We define an authoritative rule application for a graph $\mathcal{G}$ w.r.t. the T-Box $\mathcal{T}_{\mathcal{W}}$ to be a $\mathcal{T}$-split rule application $\mathcal{T}_r(\mathcal{T}_{\mathcal{W}}, \mathcal{G})$ where additionally, if both $\mathcal{A}nte_{\mathcal{T}}$ and $\mathcal{A}nte_{\mathcal{G}}$ are non-empty ($r \in \mathcal{R}_{\mathcal{TG}}$), then for the mapping μ of $\mathcal{T}_r(\mathcal{T}_{\mathcal{W}}, \mathcal{G})$ there must exist a variable $v \in (\mathcal{V}(\mathcal{A}nte_{\mathcal{T}}) \cap \mathcal{V}(\mathcal{A}nte_{\mathcal{G}}))$ such that $\mu(v) \in \widehat{sig}(\mathcal{W})$. We denote an authoritative rule application by $\mathcal{T}_{\hat{r}}(\mathcal{T}_{\mathcal{W}}, \mathcal{G})$.*

In other words, an authoritative rule application will only occur if the rule consists of only assertional patterns ($\mathcal{R}_{\mathcal{G}}$); or the rules consists of only terminological patterns ($\mathcal{R}_{\mathcal{T}}$); or if in application of the rule, the terminological pattern instance is from a web-graph authoritative for at least one class or property in the assertional pattern instance. The $T_{\hat{R}}$ operator follows naturally as before for a set of authoritative rules $\hat{\mathcal{R}}$, as does the notion of authoritative closure which we denote by $Cl_{\hat{R}}(\mathcal{T}, \mathcal{W})$. We may also refer to, e.g., $T_{\hat{R}}(\mathbb{T}, \mathbb{KB})$ and $Cl_{\hat{R}}(\mathbb{T}, \mathbb{KB})$ as before for a $\mathcal{T}$-split rule application.

Table 2 identifies the authoritative restrictions we place on our rules wherein the underlined T-Box pattern is matched by a set of triples from a web-graph $\mathcal{W}$ iff $\mathcal{W}$ speaks authoritatively for at least one element matching a boldface variable in Table 2; i.e., again, for each rule, at least one of the classes or properties matched by the A-Box pattern of the antecedent must be authoritatively spoken for by an instance of the T-Box pattern. These restrictions only apply to $\mathcal{R}1$ and $\mathcal{R}2$ (which are both a subset of $\mathcal{R}_{\mathcal{TG}}$). Please note that, for example in rule **rdfp14b′** where there are no boldface variables, the variables enforced to be instantied by blank nodes will always be authoritatively spoken for: a web-graph is always authoritative for its blank nodes.

We now make the following proposition relating to the prevention of ontology-hijacking through authoritative rule application:

Proposition 1 *Given a T-Box $\mathcal{T}_{\mathcal{W}}$ extracted from a web-graph $\mathcal{W}$ and any graph $\mathcal{G}$ not mentioning any element of $\widehat{sig}(\mathcal{W})$, then $Cl_{\widehat{\mathcal{R}_{\mathcal{TG}}}}(\mathcal{T}_{\mathcal{W}}, \mathcal{G}) = \mathcal{G}$.*

Proof: Informally, our proposition is that the authoritative closure of a graph $\mathcal{G}$ w.r.t. some T-Box $\mathcal{T}_{\mathcal{W}}$ will not contain any inferences which constitute ontology hijacking, defined in terms of ruleset $\mathcal{R}_{\mathcal{TG}}$.

Firstly, from Definition 3, for each rule $r \in \mathcal{R}_{\mathcal{TG}}$, $\mathcal{A}nte_{\mathcal{T}} \neq \emptyset$ and $\mathcal{A}nte_{\mathcal{G}} \neq \emptyset$. Therefore, from Definitions 4 & 8, for an authoritative rule application to occur for any such r, there must exist both (i) a mapping μ such that $\mu(\mathcal{A}nte_{\mathcal{T}}) \subseteq \mathcal{T}_{\mathcal{W}}$ and $\mu(\mathcal{A}nte_{\mathcal{G}}) \subseteq \mathcal{G}$; and (ii) a variable $v \in ((\mathcal{A}nte_{\mathcal{T}})(\mathcal{A}nte_{\mathcal{G}}))$ such that $\mu(v) \in \widehat{sig}(\mathcal{W})$. However, since $\mathcal{G}$ does not mention any element of $\widehat{sig}(\mathcal{W})$, then there is no such mapping μ where $\mu(v) \in \widehat{sig}(\mathcal{W})$ for $v \in \mathcal{V}(\mathcal{A}nte_{\mathcal{G}})$, and $\mu(\mathcal{A}nte_{\mathcal{G}}) \subseteq \mathcal{G}$. Hence, for $r \in \mathcal{R}_{\mathcal{TG}}$, no such application $\mathcal{T}_{\hat{r}}(\mathcal{T}_{\mathcal{W}}, \mathcal{G})$ will occur; it then follows that $\mathcal{T}_{\widehat{\mathcal{R}_{\mathcal{TG}}}}(\mathcal{T}_{\mathcal{W}}, \mathcal{G}) = \emptyset$ and $Cl_{\widehat{\mathcal{R}_{\mathcal{TG}}}}(\mathcal{T}_{\mathcal{W}}, \mathcal{G}) = \mathcal{G}$. □

The above proposition and proof holds for a given web-graph $\mathcal{W}$; however, given a set of web-graphs where an instance of $\mathcal{A}nte_{\mathcal{T}}$ can consist of triples from more that one graph, it is possible for ontology hijacking to occur whereby some triples in the instance come from a non-authoritative graph and some from an authoritative graph. To illustrate we refer to the following example, wherein (and without enforcing abstract syntax blank nodes) the second source could cause ontology hijacking by interfering with the authoritative definition of the class restriction in the first source as follows:

```
>   RULE (adapted so that ?C need not be a
    blank node)
>   ?C :allValuesFrom ?D ; :onProperty ?P .
    ?x a ?C ; ?P ?y . ⇒ ?y a ?D .
>   SOURCE <ex:>
>   ex:Person :onProperty ex:parent.
>   SOURCE <ex2:>
>   ex:Person :allValuesFrom ex2:Human
>   ASSERTIONAL
>   _:Jim a ex:Person ; ex:parent _:Jill .
>   ⇒
>   _:Jill a ex2:Human .
```

Here, the above inference is authoritative according to our definition since the instance of $\mathcal{A}nte_{\mathcal{T}}$ (specifically the first statement from `source <ex:>`) speaks authoritatively for a class/property in the assertional data; however, the statement from the `<ex2:>` source is causing inferences on assertional data not containing a class or property for which `<ex2:>` is authoritative.

As previously discussed, for our ruleset, we enforce the OWL abstract syntax and thus we enforce that $\mu(\mathcal{A}nte_{\mathcal{T}}) \subseteq \mathcal{T}_{\mathcal{W}'}$ where $\mathcal{T}_{\mathcal{W}'} \in \mathbb{T}$. However, where this condition does not hold (i.e., an instance of $\mathcal{A}nte_{\mathcal{T}}$ can comprise of data from more than one graph), then an authoritative rule application should only occur if each web-graph contributing to an instance of $\mathcal{A}nte_{\mathcal{T}}$ speaks authoritatively for at least one class/property in the $\mathcal{A}nte_{\mathcal{G}}$ instance.

REASONING ALGORITHM

In the following we first present observations on web data that influenced the design of the SAOR algorithm, then give an overview of the algorithm, and next discuss details of how we handle T-Box information, perform statement-wise reasoning, and deal with equality for individuals.

Characteristics of Web Data

Our algorithm is intended to operate over a web knowledge-base as retrieved by means of a web crawl; therefore, the design of our algorithm is motivated by observations on our web dataset:

1. Reasoning accesses a large slice of data in the index: we found that approximately 61% of statements in the 147m dataset and 90% in the 1.1b dataset produced inferred statements through authoritative reasoning.
2. Relative to assertional data, the volume of terminological data on the Web is small: < 0.9% of the statements in the 1.1b dataset and <1.7% of statements in the 147m dataset were classifiable as SAOR T-Box statements.[11]
3. The T-Box is the most frequently accessed segment of the knowledge-base for reason-

ing: although relatively small, all but the rules in $\mathcal{R}3$ require access to T-Box information.

Following from the first observation, we employ a file-scan batch-processing approach so as to enable sequential access over the data and avoid disk-lookups and dynamic data structures which would not perform well given high disk latency; also we avoid probing the same statements repeatedly for different rules at the low cost of scanning a given percentage of statements not useful for reasoning.

Following from the second and third observations, we optimise by placing T-Box data in a separate data structure accessible by the reasoning engine. Currently, we hold all T-Box data in-memory, but the algorithm can be generalised to provide for a caching on-disk structure or a distributed in-memory structure as needs require.[12]

To be able to scale, we try to minimise the amount of main memory needed, given that main memory is relatively expensive and that disk-based algorithms are thus more economical (Kunkle and Cooperman 2008). Given high disk latency, we avoid using random-access on-disk data structures. In our previous work a disk-based updateable random-access data structure (a B+-Tree) proved to be the bottleneck for the reasoning due to a high volume of inserts, leading to frequent index reorganisations and hence inadequate performance. As a result, our algorithms are now build upon two disk-based primitives known to scale: file scanning and sorting.

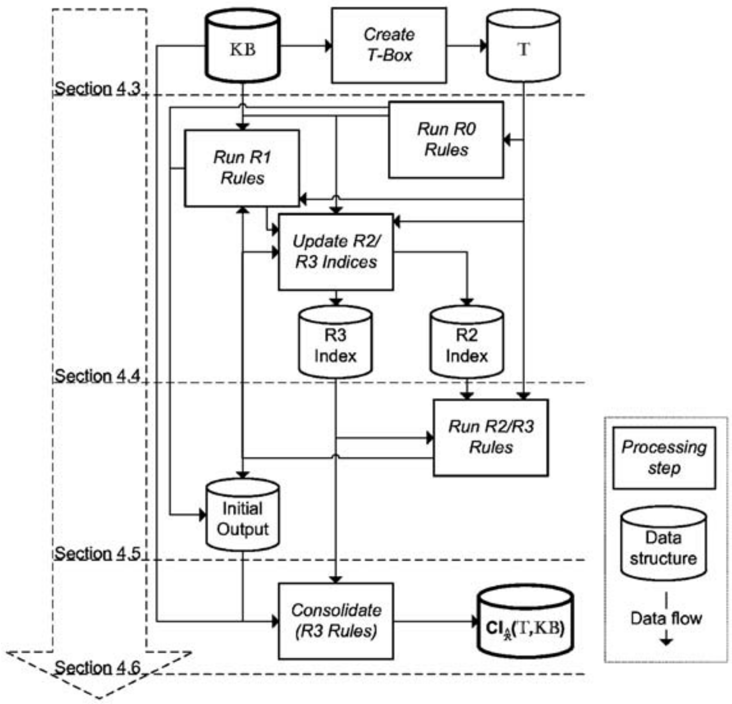
Algorithm Overview

The SAOR algorithm performs a fixpoint computation by iteratively applying the rules in Table 2. Figure 1 outlines the architecture. The reasoning process can be roughly divided into the following steps:

1. Separate $\mathbb{T}$ from $\mathbb{KB}$, build in-memory representation $\mathbb{T}$, and apply ruleset $\mathcal{R}0$ (Section 4.3).
2. Perform reasoning over $\mathbb{KB}$ in a statement-wise manner (Section 4.4):
   - 2.1 Execute rules with only a single A-Box triple pattern in the antecedent ($\mathcal{R}1$): join A-Box pattern with in-memory T-Box; recursively execute steps over inferred statements; write inferred RDF statements to output file.
   - 2.2 Write on-disk files for computation of rules with multiple A-Box triple patterns in the antecedent ($\mathcal{R}2$); when a statement matches one of the A-Box triple patterns for these rules and the necessary T-Box join exists, the statement is written to the on-disk file for later rule computation.
   - 2.3 Write on-disk equality file for rules which involve equality reasoning ($\mathcal{R}3$); `:sameAs` statements found during the scan are written to an on-disk file for later computation.
3. Execute ruleset $\mathcal{R}3 \cup \mathcal{R}3$: on-disk files containing partial A-Box instances for rules in $\mathcal{R}2$ and $\mathcal{R}3$ are sequentially analysed producing further inferred statements. Newly inferred statements are again subject to step 2 above; fresh statements can still be written to on-disk files and so the process is iterative until no new statements are found (Section 4.5).
4. Finally, consolidate source data along with inferred statements according to `:sameAs` computation ($\mathcal{R}3$) and write to final output (Section 4.6).

In the following sections, we discuss the individual components and processes in the architecture as highlighted, whereafter, in Section 4.7 we show how these elements are combined to achieve closure.

Figure 1. High-level architecture



Handling Terminological Data

In the following, we describe how to separate the T-Box data and how to create the data structures for representing the T-Box.

T-Box data from RDFS and OWL specifications can be acquired either from conventional crawling techniques, or by accessing the locations pointed to by the dereferenced URIs of classes and properties in the data. We assume for brevity that all the pertinent terminological data have already been collected and exist within the input data. If T-Box data are sourced separately via different means we can build an in-memory representation directly, without requiring the first scan of all input data.

We apply the following algorithm to create the T-Box in-memory representation, which we will analyse in the following sections:

- *FULL SCAN 1:* separate T-Box information as described in Definition 1.
- *TBOX SCAN 1 & 2:* reduce irrelevant RDF collection statements.
- *TBOX SCAN 3:* perform authoritative analysis of the T-Box data and load in-memory representation.

Separating and Reducing T-Box Data

Firstly, we wish to separate all possible T-Box statements from the main bulk of data. $\mathcal{P}_{SAOR}$ and $\mathcal{C}_{SAOR}$ are stored in memory and then the data dump is scanned. Quadruples with property $\in \mathcal{P}_{SAOR} \cup \{$ `rdf:first`, `rdf:rest` $\}$ or `rdf:type` statements with object $\in \mathcal{C}_{SOAR}$ (which, where applicable, abide by the OWL abstract syntax) are buffered to a T-Box data file.

However, the T-Box data file still contains a large amount of RDF collection statements (property ∈ { `rdf:first`, `rdf:rest` }) which are not related to reasoning. SAOR is only interested in such statements wherein they form part of a `:unionOf`, `:intersectionOf` or `:oneOf` class description. Later when the T-Box is being loaded, these collection fragments are reconstructed in-memory and irrelevant collection fragments are discarded; to reduce the amount of memory required we can quickly discard irrelevant collection statements through two T-Box scans:

- scan the T-Box data and store contexts of statements where the property ∈ { `:unionOf`, `:intersectionOf`, `:oneOf` }.
- scan the T-Box data again and remove statements for which both hold:
  - property ∈ { `rdf:first`, `rdf:rest` }
  - the context does not appear in those stored from the previous scan.

These scans quickly remove irrelevant collection fragments where a `:unionOf`, `:intersectionOf`, `:oneOf` statement does not appear in the same source as the fragment (i.e., collections which cannot contribute to the T-Box pattern of one of our rules).

Authoritative Analysis

We next apply authoritative analysis to the T-Box and load the results into our in-memory representation; in other words, we build an *authoritative T-Box* which pre-computes authority of T-Box data. We denote our authoritative T-Box by $\widehat{\mathbb{T}}$, whereby $\mathcal{C}l_{\widehat{\mathcal{R}}}(\mathbb{T}, \mathbb{KB}) = \mathcal{C}l_{\mathcal{R}}(\widehat{\mathbb{T}}, \mathbb{KB})$; for each rule, $\widehat{\mathbb{T}}$ only contains T-Box pattern instances for $\mathcal{A}nte_{\mathcal{T}}$ which can lead to an authoritative rule application.

Each statement read is initially matched without authoritative analysis against the patterns

Table 3. T-Box patterns and associated links kept between classes and properties in the in-memory T-Box

	$\mathcal{R}0$	
rdfc0	?C :oneOf ($?x_1$ ... $?x_n$) .	($?x_1$ ... $?x_n$) $\xrightarrow{\text{rdfc0}}$?C
	$\mathcal{R}1$	
rdfs2	**?P** rdfs:domain ?C .	**?P** $\xrightarrow{\text{rdfs2}}$?C
rdfs3′	**?P** rdfs:range ?C .	**?P** $\xrightarrow{\text{rdfs3}'}$?C
rdfs7′	**?P** rdfs:subPropertyOf ?Q .	**?P** $\xrightarrow{\text{rdfs7}'}$?Q
rdfs9	**?C** rdfs:subClassOf ?D .	**?C** $\xrightarrow{\text{rdfs9}}$?D
rdfp3′	**?P** a :SymmetricProperty .	**?P** $\xrightarrow{\text{rdfp3}'}$ TRUE
rdfp8a′	**?P** :inverseOf ?Q .	**?P** $\xrightarrow{\text{rdfp8a}'}$?Q
rdfp8b′	?P :inverseOf **?Q** .	**?Q** $\xrightarrow{\text{rdfp8b}'}$?P
rdfp12a′	**?C** :equivalentClass ?D .	**?C** $\xrightarrow{\text{rdfp12a}'}$?D
rdfp12b′	?C :equivalentClass **?D** .	**?D** $\xrightarrow{\text{rdfp12b}'}$?C
rdfp13a′	**?P** :equivalentProperty ?Q .	**?P** $\xrightarrow{\text{rdfp13a}'}$?Q
rdfs13b′	?P :equivalentProperty **?Q** .	**?Q** $\xrightarrow{\text{rdfs13b}'}$?P
rdfp14a′	?C :hasValue ?y ; :onProperty **?P** .	**?P** $\xrightarrow{\text{rdfp14a}'}$ { ?C, ?y }
rdfp14b′	?C :hasValue ?y ; :onProperty ?P .	?C $\xrightarrow{\text{rdfp14b}'}$ { ?P, ?y }
rdfc1	?C :unionOf ($?C_1$...$\mathbf{?C_i}$...$?C_n$) .	$\mathbf{?C_i}$ $\xrightarrow{\text{rdfc1}}$?C
rdfc2	?C :minCardinality 1 ; :onProperty **?P** .	**?P** $\xrightarrow{\text{rdfc2}}$?C
rdfc3a	?C :intersectionOf ($?C_1$ ... $?C_n$) .	?C $\xrightarrow{\text{rdfc3a}}$ { $?C_1$, ..., $?C_n$ }
rdfc3b	?C :intersectionOf ($\mathbf{?C_1}$) .	$\mathbf{?C_1}$ $\xrightarrow{\text{rdfc3b}}$?C
	$\mathcal{R}2$	
rdfp1′	**?P** a :FunctionalProperty .	**?P** $\xrightarrow{\text{rdfp1}'}$ TRUE
rdfp2	**?P** a :InverseFunctionalProperty .	**?P** $\xrightarrow{\text{rdfp2}}$ TRUE
rdfp4	**?P** a :TransitiveProperty .	**?P** $\xrightarrow{\text{rdfp4}}$ TRUE
rdfp15′	?C :someValuesFrom **?D** ; :onProperty **?P** .	**?P** $\xleftrightarrow{\text{rdfp15}'}$ **?D** $\xrightarrow{\text{rdfp15}'}$?C
rdfp16′	?C :allValuesFrom ?D ; :onProperty ?P .	?P $\xleftrightarrow{\text{rdfp16}'}$?C $\xrightarrow{\text{rdfp16}'}$?D
rdfc3c	?C :intersectionOf ($\mathbf{?C_1}$ ... $\mathbf{?C_n}$) .	{ $\mathbf{?C_1}$, ..., $\mathbf{?C_n}$ } $\xrightarrow{\text{rdfc3c}}$?C
rdfc4a	?C :cardinality 1 ; :onProperty ?P .	?C $\xleftrightarrow{\text{rdfc4a}}$?P
rdfc4b	?C :maxCardinality 1 ; :onProperty ?P .	?C $\xleftrightarrow{\text{rdfc4b}}$?P

enumerated in Table 3. If a pattern is initially matched, the positions required to be authoritative, as identified in boldface, are checked. If one such authoritative check is satified, the pattern is loaded into the T-Box. Indeed the same statement may be matched by more than one T-Box pattern for different rules with different authoritative restrictions; for example the statement `foaf:name :equivalentProperty my:name .` retrieved from `my:` namespace matches the T-Box pattern of rules **rdfp13a′** & **rdfp13b′**, but only conforms to the authoritative restriction for rule **rdfp13b′**. Therefore, we only store the statement in such a fashion as to apply to rule **rdfp13b′**; that is, the authoritative T-Box stores T-Box pattern instances separately for each rule, according to the authoritative restrictions for that rule.

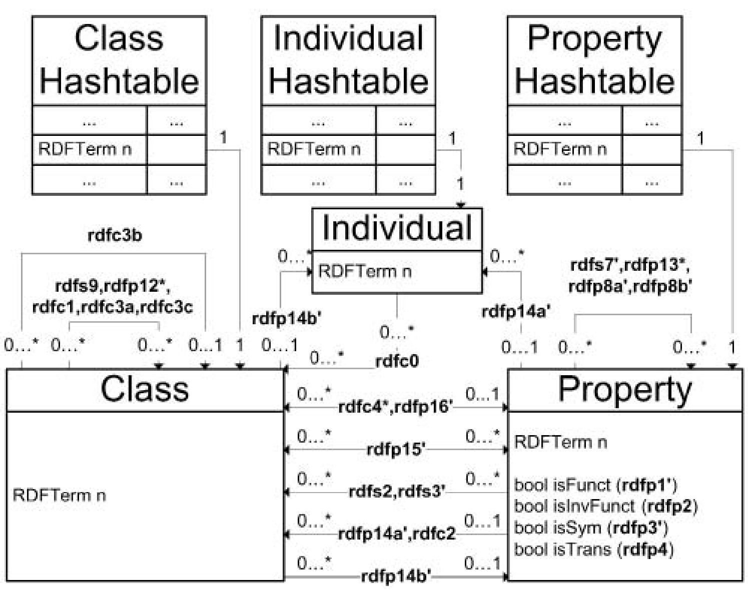
Checking the authority of a source for a given namespace URI, as presented in Definition 6, may require a HTTP connection to the namespace URI so as to determine whether a redirect exists to the authoritative document (HTTP Response Code 303). Results of accessing URIs are cached once in-memory so as to avoid establishing repetitive connections. If the pattern is authoritatively matched, the statement is reflected in the in-memory T-Box. Alternatively, where available, a crawler can provide a set of redirect pairs which can be loaded into the system to avoid duplicating HTTP lookups; we presume for generality that such information is not provided.

In-Memory T-Box

Before we proceed, we quickly discuss the storage of `:oneOf` constructs in the T-Box for rule **rdfc0**. Individuals ($?x_1 \ldots ?x_n$) are stored with pointers to the one-of class ?C. Before input data are read, these individuals are asserted to be of the `rdf:type` of their encompassing one-of class.

Besides the one-of support, for the in-memory T-Box we employ two separate hashtables, one for classes and another for properties, with RDF terms as key and a Java representation of the class or property as value. The representative Java objects contain labelled links to related objects as defined in Table 3 and Figure 2. The property and class objects are designed to contain all of the infor-

Figure 2. In-memory T-Box structure



mation required for reasoning on a membership assertion of that property or class: that is, classes/properties satisfying the A-Box antecedent pattern of a rule are linked to the classes/properties appearing in the consequent of that rule, with the link labelled according to that rule. During reasoning, the class/property identifier used in the membership assertion is sent to the corresponding hashtable and the returned internal object used for reasoning on that assertion. The objects contain the following:

- Property objects contain the property URI and references to objects representing domain classes (**rdfs2)**, range classes **(rdfs3′)**, super properties **(rdfs7')**, inverse properties **(rdfs8*)** and equivalent properties **(rdfp13*)**. References are kept to restrictions where the property in question is the object of an `:onProperty` statement (**rdfp14a′, rdfp16′, rdfc2, rdfc4*)**. Where applicable, if the property is part of a some-values-from restriction, a pointer is kept to the some-values-from class **(rdfp15′)**. Boolean values are stored to indicate whether the property is functional **(rdfp1′)**, inverse-functional **(rdfp2)**, symmetric **(rdfp3′)** and/or transitive **(rdfp4)**.
- Class objects contain the class URI and references to objects representing super classes **(rdfs9)**, equivalent classes **(rdfp12*)** and classes for which this class is a component of a union **(rdfc1)** or intersection **(rdfc3b/c)**. On top of these core elements, different references are maintained for different types of class description:
  - intersection classes store references to their constituent class objects **(rdfc3a)**
  - restriction classes store a reference to the property the restriction applies to **(rdfp14b′, rdfp15′, rdfc2, rdfc4*)** and also, if applicable to the type of restriction:
    - the values which the restriction property must have **(rdfp14b′)**
    - the class for which this class is a some-values-from restriction value **(rdfp15′)**

The algorithm must also performs in-memory joining of collection segments according to `rdf:first` and `rdf:rest` statements found during the scan for the purposes of building union, intersection and enumeration class descriptions. Again, any remaining collections not relevant to the T-Box segment of the knowledge-base (i.e., not terminological collection statements) are discarded at the end of loading the input data; we also discard cyclic and branching lists as well as any lists not found to end with the `rdf:nil` construct.

We have now loaded the final T-Box for reasoning into memory; this T-Box will remain fixed throughout the whole reasoning process.

Initial Input Scan

Having loaded the terminological data, SAOR is now prepared for reasoning by statement-wise scan of the assertional data. We provide the high-level flow for reasoning over an input statement in Function 1. The reasoning scan process can be described as recursive depth-first reasoning whereby each unique statement produced is again input immediately for reasoning. Statements produced thus far for the original input statement are kept in a set to provide uniqueness testing and avoid cycles; a uniquing function is maintained for a common subject group in the data, ensuring that statements are only produced once for that statement group. Once all of the statements produced by a rule have been themselves recursively analysed, the reasoner moves on to analysing the proceeding rule and loops until no unique statements are inferred.

```
Function 1:ReasonStatement(s)
  Input: statement s
  Global: T, index, output
  foreach rule r ∈ R1 do
      fire rule r w.r.t s and T;
      foreach inferred statement: i do
          if i is unique then
              if i is valid RDF then
                  write to output;
              ReasonStatement (i) ;
  foreach rule r ∈ R2 ∪ R3 do
      if s relevant for r w.r.t. T then
          write s to index, ;
```

There are three disjoint categories of statements which require different handling: namely (i) `rdf:type` statements, (ii) `:sameAs` statements, (iii) all other statements. We assume disjointness between the statement categories: we do not allow any external extension of the core `rdf:type`/ `:sameAs` semantics (non-standard use / non-authoritative extension). Further, the assertions about `rdf:type` in the RDFS specification define the `rdfs:domain` and `rdfs:range` of `rdf:type` as being `rdfs:Resource` and `rdfs:Class`; since we are not interested in inferring membership of such RDFS classes we do not subject `rdf:type` statements to property-based entailments. The only assertions about `:sameAs` from the OWL specification define domain and range as `:Thing` which we ignore by the same justification.

The `rdf:type` statements are subject to class-based entailment reasoning and require joins with class descriptions in the T-Box. The `:sameAs` statements are handled by ruleset $\mathcal{R}3$, which we discuss in Section 4.6. All other statements are subject to property-based entailments and thus requires joins with T-Box property descriptions.

Ruleset $\mathcal{R}2 \bigcup \mathcal{R}3$ cannot be computed solely on a statement-wise basis. Instead, for each rule, we assign an on-disk file (blocked and compressed to save disk space). Each file contains statements which may contribute to satisfying the antecedent of its pertinent rule. During the scan, if an A-Box statement satisfies the necessary T-Box join for a rule, it is written to the index for that rule. For example, when the statement

> `ex:me foaf:isPrimaryTopicOf ex:myHomepage .`

is processed, the property object for `foaf:isPrimaryTopicOf` is retrieved from the T-Box property hashtable. The object states that this property is of type `:InverseFunctionalProperty`. The rule cannot yet be fired as this statement alone does not satisfy the A-Box segment of the antecedent of **rdfp2** and the method is privy to only one A-Box statement at a time. When, later, the statement:

> `ex:me2 foaf:isPrimaryTopicOf ex:myHomepage .`

is found, it also is written to the same file – the file now contains sufficient data to (although it cannot yet) fire the rule and infer:

> `ex:me :sameAs ex:me2 .`

During the initial scan and inferencing, all files for ruleset $\mathcal{R}2 \bigcup \mathcal{R}3$ are filled with pertinent statements analogously to the example above. After the initial input statements have been exhausted, these files are analysed to infer, for example, the `:sameAs` statement above.

On-Disk A-Box Join Analysis

In this section, we discuss handling of the on-disk files containing A-Box statements for ruleset $\mathcal{R}2 \bigcup \mathcal{R}3$. We firstly give a general overview of the execution for each rule using an on-disk file and then look at the execution of each rule.

Table 4 presents the joins to be executed via the on-disk files for each rule: the key join variables, used for computing the join, are shown in boldface. In this table we refer to *SPOC and OPSC sorting order:* these can be intuitively interpreted as quads sorted according to subject, predicate,

Table 4. Table enumerating the A-Box joins to be computed using the on-disk files with key join variables in boldface font and sorting order required for statements to compute join

	$\mathcal{R}2$	
rdfp1′	**?x** ?P ?y , ?z .	SPOC
rdfp2	?x ?P **?z** . ?y ?P **?z** .	OPSC
rdfp4	?x ?P **?y** . **?y** ?P ?z .	SPOC & OPSC
rdfp15′	?x ?P **?y** . **?y** a ?D .	SPOC / OPSC
rdfp16′	**?x** a ?C ; ?P ?y .	SPOC
rdfc3c	**?x** a ?C_1, ..., ?C_n .	SPOC
rdfc4a	**?x** a ?C ; ?P ?y , ?z .	SPOC
rdfc4b	**?x** a ?C ; ?P ?y , ?z .	SPOC
	$\mathcal{R}3$	
rdfp7	?x :sameAs **?y** . **?y** :sameas ?z .	SPOC & OPSC
rdfp11′	**?x** :sameAs ?_x ; ?P ?y .	SPOC
rdfp11″	**?y** :sameAs ?_y . ?x ?P **?y** .	SPOC / OPSC

object, context (natural sorting order) and object, predicate, subject, context (inverse sorting order) respectively. For the internal index files, we use context to encode the sorting order of a statement and the iteration in which it was added; only joins with at least one new statement from the last iteration will infer novel output.

Again, an on-disk file is dedicated for each rule/join required. The joins to be computed are a simple "star shaped" join pattern or "one-hop" join pattern (which we reduce to a simple star shaped join computation by inverting one one or more patterns to inverse order). The statements in each file are initially sorted according to the key join variable. Thus, common bindings for the key join variable are grouped together and joins can be executed by means of sequential scan for common key join variable binding groups.

We now continue with a more detailed description of the process for each rule beginning with the more straightforward rules.

Functional Property Reasoning - Rule *rdfp1′*

From the initial input scan, we have a file containing only statements with functional properties in the predicate position (as described in Section 4.4). As can be seen from Table 4, the key join variable is in the subject position for all A-Box statements in the pattern. Thus, we can sort the file according to SPOC (natural) order. The result is a file where all statements are grouped according to a common subject, then predicate, then object. We can now scan this file, storing objects with a common subject-predicate. We can then fire the rule stating equivalence between these objects.

Inverse Functional Reasoning - ***Rule rdfp2***

Reasoning on statements containing inverse functional properties is conducted analogously to functional property reasoning. However, the key join variable is now in the object position for all A-Box statements in the pattern. Thus, we instead sort the file according to OPSC (inverse) order and scan the file inferring equivalence between the subjects for a common object-predicate group.

Intersection Class Reasoning - Rule *rdfc3c*

The key join variable for rule rdfc3c is in the subject position for all A-Box triple patterns. Thus we can sort the file for the rule (filled with memberships assertions for classes which are part of some intersection) according to SPOC order. We can scan common subject-predicate (in any case, the predicates all have value `rdf:type`) groups storing the objects (all types for the subject resource which are part of an intersection). The containing intersection for each type can then be retrieved and the intersection checked to see if all of it's constituent types have been satisfied. If so, membership of the intersection is inferred.

All-Values-From Reasoning - Rule *rdfp16′*

Again, the key join variable for rule **rdfp16'** is in the subject position for all A-Box triple patterns and again we can sort the file according to SPOC order. For a common subject group, we store `rdf:type` values and also all predicate/object edges for the given subject. For every membership assertion of an all-values-from restriction class (as is given by all of the `rdf:type` statements in the file), we wish to infer that objects of the `:onProperty` value (as is given by all the non-`rdf:type` statements) are of the all-values-from class. Therefore, for each restriction membership assertion, the objects of the corresponding `:onProperty`-value membership-assertions are inferred to be members of the all-values-from object class (`?D`).

Some-Values-From Reasoning - Rule *rdfp15′*

For some-values-from reasoning, the key join variable is in the subject position for `rdf:type` statements (all membership assertions of a some-values-from object class) but in the object position for the `:onProperty` membership assertions. Thus, we order class membership assertions in the file according to natural SPOC order and property membership assertions according to inverse OPSC order. In doing so, we can scan common `?y` binding groups in the file, storing `rdf:type` values and also all predicate/subject edges. For every member of a some-values-from object class (as is given by all of the `rdf:type` statements in the file according to the join with the T-Box on the `?D` position), we infer that subjects of the `:onProperty`-value statements (as is given by all the non-`rdf:type` statements according to the T-Box join with `?P`) are members of the restriction class (`?C`).

Transitive Reasoning (Non-Symmetric) - *Rule rdfp4*

Transitive reasoning is perhaps the most challenging to compute: the output of rule **rdfp4** can again recursively act as input to the rule. For closure, recursive application of the rule must be conducted in order to traverse arbitrarily long transitive paths in the data.

Firstly, we will examine sorting order. The key join variable is in the subject position for one pattern and in the object position for the second pattern. However, both patterns are identical: a statement which matches one pattern will obviously match the second. Thus, every statement in the transitive reasoning file is duplicated with one version sorted in natural SPOC order, and another in inverse OPSC.

Take for example the following triples where `ex:comesBefore` is asserted as being of type `:TransitiveProperty` in the T-Box:

```
>    INPUT:
>    ex:a ex:comesBefore ex:b .
>    ex:b ex:comesBefore ex:c .
>    ex:c ex:comesBefore ex:d .
```

In order to compute the join, we must write the statements in both orders, using the context to mark which triples are in inverse order, and sort them accordingly (for this internal index, we temporarily relax the requirement that context is a URI).

```
>    SORTED FILE - ITERATION 1:[13]
>    ex:a ex:comesBefore ex:b _:spoc1 .
>    ex:b ex:comesBefore ex:a _:opsc1 .
```

```
>    ex:b ex:comesBefore ex:c _:spoc1 .
>    ex:c ex:comesBefore ex:b _:opsc1 .
>    ex:c ex:comesBefore ex:d _:spoc1 .
>    ex:d ex:comesBefore ex:c _:opsc1 .
```

The data, as above, can then be scanned and for each common join-binding/predicate group (e.g., `ex:b ex:comesBefore`), the subjects of statements in inverse order (e.g., `ex:a`) can be linked to the object of naturally ordered statements (e.g., `ex:c`) by the transitive property. However, such a scan will only compute a single one-hop join. From above, we only produce:

```
>    OUTPUT - ITERATION 1 / INPUT - ITERATION
     2:
>    ex:a ex:comesBefore ex:c .
>    ex:b ex:comesBefore ex:d .
```

We still not have not computed the valid statement `ex:a ex:comesBefore ex:d .` which requires a two hop join. Thus we must iteratively feedback the results from one scan as input for the next scan. The output from the first iteration, as above, is also reordered and sorted as before and merge-sorted into the main

```
>    SORTED FILE - ITERATION 2:
>    ex:a ex:comesBefore ex:b _:spoc1 .
>    ex:a ex:comesBefore ex:c _:spoc2 .
>    ex:b ex:comesBefore ex:a _:opsc1 .
>    ex:b ex:comesBefore ex:c _:spoc1 .
>    ex:b ex:comesBefore ex:d _:spoc2 .
>    ex:c ex:comesBefore ex:a _:opsc2 .
>    ex:c ex:comesBefore ex:b _:opsc1 .
>    ex:c ex:comesBefore ex:d _:spoc1 .
>    ex:d ex:comesBefore ex:b _:opsc2 .
>    ex:d ex:comesBefore ex:c _:opsc1 .
```

The observant reader may already have noticed from above that we also mark the context with the iteration for which the statement was added. In every iteration, we only compute inferences which involve the delta from the last iteration; thus the process is comparable to semi-naïve evaluation. Only joins containing at least one newly added statement are used to infer new statements for output. Thus, from above, we avoid repeat inferences from ***ITERATION*** 1 and instead infer:

```
>    OUTPUT - ITERATION 2:
>    ex:a ex:comesBefore ex:d .
```

A fixpoint is reached when no new statements are inferred. Thus we would require another iteration for the above example to ensure that no new statements are inferable. The number of iterations required is in $\bigcirc(\log n)$ according to the longest unclosed transitive path in the input data. Since the algorithm requires scanning of not only the delta but all data, performance using on-disk file scans alone would be sub-optimal. For example, if one considers that most of the statements constitute paths of, say ≤8 vertices, one path containing 128 vertices would require four more scans after the bulk of the paths have been closed.

With this in mind, we accelerate transitive closure by means of an in-memory transitivity index. For each transitive property found, we store sets of linked lists which represent the graph extracted for that property. From the example ***INPUT*** from above, we would store.

```
>    ex:comesBefore | ex:a -> ex:b -> ex:c ->
     ex:d
```

From this in-memory linked list, we would then collapse all paths of length ≥2 (all paths of length 1 are input statements) and infer closure at once:

```
>    OUTPUT - ITERATION 1 / INPUT - ITERATION
     2:
>    ex:a ex:comesBefore ex:c .
>    ex:a ex:comesBefore ex:d .
>    ex:b ex:comesBefore ex:d .
```

Obviously, for scalability requirements, we do not expect the entire transitive body of statements to fit in-memory. Thus, before each iteration we calculate the in-memory capacity and only store a pre-determined number of properties and vertices. Once the in-memory transitive index is full, we infer the appropriate statements and continue by file-scan. The in-memory index is only used to store the delta for a given iteration (everything for the first iteration). Thus, we avoid excess iterations to compute closure of a small percentage of statements which form a long chain and greatly accelerate the fixpoint calculation.

Transitive Reasoning (Symmetric) - Rules *rdfp3′/rdfp4*

We use a separate on-disk file for membership assertions of properties which are both transitive and symmetric. A graph of symmetric properties is direction-less, thus the notion of direction as evident above though use of inverted ordered statements is unnecessary. Instead, all statements and their inverses (computed from symmetric rule **rdfp3′**) are written in natural SPOC order and direct paths are inferred between all objects in a common subject/predicate group. The in-memory index is again similar to above; however, we instead use a direction-less doubly-linked list.

Equality Reasoning - Ruleset $\mathcal{R}3$

Thus far, we have not considered `:sameAs` entailment, which is supported in SAOR through rules in $\mathcal{R}3$. Prior to executing rules **rdfp11′** & **rdfp11″**, we must first perform symmetric transitive closure on the list of all `:sameAs` statements (rules **rdfp6′** & **rdfp7**). Thus, we use an on-disk file analogous to that described in Section 4.5.7.

However, for rules **rdfp6′** & **rdfp7**, we do not wish to experience an explosion of inferencing through long equivalence chains (lists of equivalent individuals where there exists a `:sameAs` path from each individual to every other individual). The closure of a symmetric transitive chain of n vertices results in $n\ (n-1)$ edges or statements (ignoring reflexive statements). For example, in (Hogan, Harth et al. 2007) we found a chain of 85,803 equivalent individuals inferable from a Web dataset.[14] Naïvely applying symmetric transitive reasoning as discussed in Section 4.5.7 would result in a closure of 7.362b `:sameAs` statements for this chain alone.

Similarly, `:sameAs` entailment, as according to rules **rdfp11′** & **rdfp11″**, duplicates data for all equivalent individuals which could result in a massive amount of duplicate data (particularly when considering uniqueness on a quad level: i.e., including duplicate triples from different sources). For example, if each of the 85,803 equivalent individuals had attached an average of 8 unique statements, then this could equate to 8*85,803*85,803 = 59b inferred statements.

Obviously, we must avoid the above scenarios, so we break from complete inference with respect to the rules in $\mathcal{R}3$. Instead, for each set of equivalent individuals, we chose a pivot identifier to use in rewriting the data. The pivot identifier is used to keep a consistent identifier for the set of equivalent individuals: the alphabetically highest pivot is chosen for convenience of computation. For alternative choices of pivot identifiers on web data see (Hogan, Harth et al. 2007). We use the pivot identifier to consolidate data by rewriting all occurrences of equivalent identifiers to the pivot identifier effectively merging the equivalent set into one individual.

Thus, we do not derive the entire closure of `:sameAs` statements as indicated in rules **rdfp6′** & **rdfp7** but instead only derive an equivalence list which points from equivalent identifiers to their pivots. As highlighted, use of a pivot identifier is necessary to reduce the amount of output statements, effectively compressing equivalent resource descriptions: we hint here that a fully expanded view of the descriptions could instead be supported through backward-chaining over the semi-materialised data.

To achieve the pivot compressed inferences we use an on-disk file containing :sameAs statements. Take for example the following statements:

```
>    INPUT:
>    ex:a :sameAs ex:b .
>    ex:b :sameAs ex:c .
>    ex:c :sameAs ex:d .
```

We only wish to infer the following output for the pivot identifier ex:a:

```
>    OUTPUT PIVOT EQUIVALENCES:
>    ex:b :sameAs ex:a .
>    ex:c :sameAs ex:a .
>    ex:d :sameAs ex:a .
```

The process is the same as that for symmetric transitive reasoning as described before: however, we only close transitive paths to nodes with the highest alphabetical order. So, for example, if we have already materialised a path from ex:d to ex:a we ignore inferring a path from ex:d to ex:b as ex:b>ex:a.

To execute rules **rdfp11′** & **rdfp11″** and perform "consolidation" (rewriting of equivalent identifiers to their pivotal form), we perform a zig-zag join: we sequentially scan the :sameAs inference output as above and an appropriately sorted file of data, rewriting the latter data according to the :sameAs statements. For example, take the following statements to be consolidated:

```
>    UNCONSOLIDATED DATA:
>    ex:a foaf:mbox <mail@example.org> .
>    ...
>    ex:b foaf:mbox <mail@example.org> .
>    ex:b foaf:name "Joe Bloggs" .
>    ...
>    ex:d :sameAs ex:b .
>    ...
>    ex:e foaf:knows ex:d .
```

The above statements are scanned sequentially with the closed :sameAs pivot output from above. For example, when the statement ex:b foaf:mbox <mailto:mail@example.org> . is first read from the unconsolidated data, the :sameAs index is scanned until ex:b :sameAs ex:a . is found (if ex:b is not found in the :sameAs file, the scan is paused when an element above the sorting order of ex:b is found). Then, ex:b is rewritten to ex:a.

```
>    PARTIALLY CONSOLIDATED DATA:
>    ex:a foaf:mbox <mail@example.org> .
>    ...
>    ex:a foaf:mbox <mail@example.org> .
>    ex:a foaf:name "Joe Bloggs" .
>    ...
>    ex:a :sameAs ex:b .
>    ...
>    ex:e foaf:knows ex:d .
```

We have now executed rule **rdfp11′** and have the data partially consolidated as shown. However, the observant reader will notice that we have not consolidated the object of the last two statements. We must sort the data again according to inverse OPSC order and again sequentially scan both the partially consolidated data and the :sameAs pivot equivalences, this time rewriting ex:b and ex:d in the object position to ex:a and producing the final consolidated data. This equates to executing rule **rdfp11″.**

For the purposes of the on-disk files for computing rules requiring A-Box joins, we must consolidate the key join variable bindings according to the :sameAs statements found during reasoning. For example consider the following statements in the functional reasoning file:

```
>    ex:a ex:mother ex:m1 .
>    ex:b ex:mother ex:m2 .
```

Evidently, rewriting the key join position according to our example pivot file will lead to inference of:

```
>    ex:m1 :sameAs ex:m2
```

which we would otherwise miss. Thus, whenever the index of `:sameAs` statements is changed, for the purposes of closure it is necessary to attempt to rewrite all join index files according to the new `:sameAs` statements. Since we are, for the moment, only concerned with consolidating on the join position we need only apply one consolidation scan.

The final step in the SAOR reasoning process is to finalise consolidation of the initial input data and the newly inferred output statements produced by all rules from scanning and on-disk file analysis. Although we have provided exhaustive application of all inferencing rules, and we have the complete set of `:sameAs` statements, elements in the input and output files may not be in their equivalent pivotal form. Therefore, in order to ensure proper consolidation of all of the data according to the final set of `:sameAs` statements, we must firstly sort both input and inferred sets of data in SPOC order, consolidate subjects according to the pivot file as above; sort according to OPSC order and consolidate objects.

However, one may notice that `:sameAs` statements in the data become consolidated into reflexive statements: i.e., from the above example `ex:a :sameAs ex:a` . Thus, for the final output, we remove any `:sameAs` statements in the data and instead merge the statements contained in our final pivot `:sameAs` equivalence index, and their inverses, with the consolidated data. These statements retain the list of all possible identifiers for a consolidated entity in the final output.

Achieving Closure

We conclude this section by summarising the approach, detailing the overall fixpoint calculations (as such, putting the jigsaw together) and detailing how closure is achieved using the individual components. Along these lines, in Algorithm 2, we provide a summary of the steps seen so far and, in particular, show the fixpoint calculations involved for exhaustive application of ruleset $\mathcal{R}2 \cup \mathcal{R}3$; we compute one main fixpoint over all of the operations required, within which we also compute two local fixpoints.

Firstly, since all rules in $\mathcal{R}2$ are dependant on `:sameAs` equality, we perform `:sameAs` inferences first. Thus, we begin closure on $\mathcal{R}2 \cup \mathcal{R}3$ with a local equality fixpoint which (i) executes all rules which produce `:sameAs` inferences (**rdfp1′,rdfp2,rdfc4***); (ii) performs symmetric-transitive closure using pivots on all `:sameAs` inferences; (iii) rewrites **rdfp1′, rdfp2** and **rdfc4*** indexes according to `:sameAs` pivot equivalences and (iv) repeats until no new `:sameAs` statements are produced.

Next, we have a local transitive fixpoint for recursively computing transitive property reasoning: (i) the transitive index is rewritten according to the equivalences found through the above local fixpoint; (ii) a transitive closure iteration is run, output inferences are recursively fed back as input; (iii) ruleset $\mathcal{R}1$ is also recursively applied over output from previous step whereby the output from ruleset $\mathcal{R}1$ may also write new statements to any $\mathcal{R}2$ index. The local fixpoint is reached when no new transitive inferences are computed.

Finally, we conclude the main fixpoint by running the remaining rules: **rdfp15′, rdfp16′ and rdfc3c.** For each rule, we rewrite the corresponding index according to the equivalences found from the first local fixpoint, run the inferencing over the index and send output for reasoning through ruleset $\mathcal{R}1$. Statements inferred directly from the rule index, or through subsequent application of ruleset $\mathcal{R}1$, may write new statements for $\mathcal{R}2$ indexes. This concludes one iteration of the main fixpoint, which is run until no new statements are inferred.

For each ruleset $\mathcal{R}0$ - $\mathcal{R}3$, we now justify our algorithm in terms of our definition of closure with respect to our static T-Box. Firstly, closure is achieved immediately upon ruleset $\mathcal{R}0$, which requires only T-Box knowledge, from our static

Algorithm 2.

```
Algorithm 2: SAOR reasoning algorithm
Input: KB
Output: Cl_R̂(T,KB)
for scan KB (Section 4.3.1) do
    obtain candidate statements for T;
reduce T, derive T̂ (Sect. 4.3.2) ;
load T̂ in-memory (Sect. 4.3.3) ;
run R0 rules ;
for inferred statement i do
    if i is valid RDF then
        write to output;
    ReasonStatement (i) (Funct. 1) ;
for s ∈ KB (Sect. 4.4) do
    ReasonStatement (s) (Funct. 1) ;
// output contains R0 ∪ R1 inferences for initial input
// index contains initial statements relevant for R2
// sameas contains initial statements relevant for R3
repeat
    repeat
        for rule r ∈ {rdfp1',rdfp2,rdfc4*} (Sect. 4.5) do
            if new_r is set then
                rewrite index_r w.r.t. sameas/ unset new_r;
            if index_r has changed or was rewritten then
                run r on index_r;
                write to sameas;
        if sameas has changed then
            run rules rdfp6' and rdfp7 on sameas (Sect. 4.6) ;
            write sameas/ set all new;
    until fixpoint reached (no changes in previous iteration) ;
    if new_rdfp4 is set then
        rewrite index_rdfp4 w.r.t. sameas/ unset new_rdfp4;
    repeat
        run rule rdfp4 on index_rdfp4 (Sect. 4.5.6 and 4.5.7) ;
        for inferred statement i do
            write to index_rdfp4 ;
            if i is RDF then
                write to output;
            ReasonStatement (i) (Funct. 1) ;
    until fixpoint reached (no changes in previous iteration) ;
    for rule r ∈ {rdfp15',rdfp16',rdfc3c} (Sect. 4.5) do
        if new_r is set then
            rewrite index_r w.r.t. sameas/ unset new_r;
        if index_r has changed or was rewritten then
            run r on index_r;
            for inferred statement i do
                if i is RDF then
                    write to output;
                ReasonStatement (i) (Funct. 1) ;
until fixpoint reached (no changes in previous iteration) ;
// output contains all inferences in non-pivotal form
for subject and object (Sect. 4.6) do
    for scan KB and output do
        rewrite according to sameas;
        write to Cl_R̂(T,KB);
    write sameas and sameas⁻ to Cl_R̂(T,KB);
```

T-Box. Secondly, with respect to the given T-Box, every input statement is subject to reasoning according to ruleset $\mathcal{R}1$, as is every statement inferred from ruleset $\mathcal{R}0$, those recursively inferred from ruleset $\mathcal{R}1$ itself, and those recursively inferred from on-disk analysis for ruleset $\mathcal{R}1 \bigcup \mathcal{R}2$. Next, every input statement is subject to reasoning according to ruleset $\mathcal{R}2$ with respect to our T-Box; these again include all inferences from $\mathcal{R}0$, all statements inferred through $\mathcal{R}1$ alone, and all inferences from recursive application of ruleset $\mathcal{R}1 \bigcup \mathcal{R}2$.

Therefore, we can see that our algorithm applies exhaustive application of ruleset $\mathcal{R}0 \bigcup \mathcal{R}1 \bigcup \mathcal{R}2$ with respect to our T-Box, leaving only consideration of equality reasoning in ruleset $\mathcal{R}3$. Indeed, our algorithm is not complete with respect to ruleset $\mathcal{R}3$ since we choose pivot identifiers for representing equivalent individuals as justified in Section 4.6. However, we still provide a form of "pivotal closure" whereby backward-chaining support of rules **rdfp11′** and **rdfp11″** over the output of our algorithm would provide a view of closure as defined; i.e., our output contains all of the possible inferences according to our notion of closure, but with equivalent individuals compressed in pivotal form.

Firstly, for rules **rdfp6′** and **rdfp7**, all statements where p = `:sameAs` from the original input or as produced by $\mathcal{R}0 \bigcup \mathcal{R}1 \bigcup \mathcal{R}2$ undergo on-disk symmetric-transitive closure in pivotal form. Since both rules only produce more `:sameAs` statements, and according to the standard usage restriction of our closure, they are not applicable to reasoning under $\mathcal{R}0 \bigcup \mathcal{R}1 \bigcup \mathcal{R}2$. Secondly, we loosely apply rules **rdfp11′ and rdfp11″** such as to provide closure with respect to joins in ruleset $\mathcal{R}2$; i.e., all possible joins are computed with respect to the given `:sameAs` statements. Equivalence is clearly not important to $\mathcal{R}0$ since we strictly do not allow `:sameAs` statements to affect our T-Box; $\mathcal{R}1$ inferences do not require joins and, although the statements produced will not be in pivotal form, they will be output and rewritten later; inferences from $\mathcal{R}2$ will be produced as discussed, also possibly in non-pivotal form. In the final consolidation step, we then rewrite all statements to their pivotal form and provide incoming and outgoing `:sameAs` relations between pivot identifiers and their non-pivot equivalent

Table 5. Comparison of authoritative and non-authoritative reasoning for the number of unique inferred RDF statements produced (w.r.t. ruleset $\mathcal{R}1$ over) the five most frequently occurring classes and properties in both input datasets. '' indicates a datatype property where the object of $m(P)$ is a literal. The amount of statements produced for authoritative reasoning for a single membership assertion of the class or property is denoted by $|\mathcal{C}l_{\mathcal{R}1}(\widehat{\mathbb{T}}, \{m(C)\})|$ and $|\mathcal{C}l_{\mathcal{R}1}(\widehat{\mathbb{T}}, \{m(P)\})|$ respectively. Non-authoritative counts are given by $|\mathcal{C}l_{\mathcal{R}1}(\mathbb{T}, \{m(C)\})|$ and $|\mathcal{C}l_{\mathcal{R}1}(\mathbb{T}, \{m(P)\})|$. n is the number of membership assertions for the class C or property P in the given dataset.*

147m Dataset

C	$\lvert Cl_{\mathcal{R}1}(\widehat{\mathbb{T}},\{m(C)\})\rvert$	$\lvert Cl_{\mathcal{R}1}(\mathbb{T},\{m(C)\})\rvert$	n	$n\lvert Cl_{\mathcal{R}1}(\widehat{\mathbb{T}},\{m(C)\})\rvert$	$n\lvert Cl_{\mathcal{R}1}(\mathbb{T},\{m(C)\})\rvert$
`rss:item`	0	356	3,558,055	0	1,266,667,580
`foaf:Person`	6	388	3,252,404	19,514,424	1,261,932,752
`rdf:Seq`	2	243	1,934,852	3,869,704	470,169,036
`foaf:Document`	1	354	1,750,365	1,750,365	619,629,210
`wordnet:Person`	0	236	1,475,378	0	348,189,208
TOTAL	9	1,577	11,971,054	25,134,493	3,966,587,786

P	$\lvert Cl_{\mathcal{R}1}(\{\widehat{\mathbb{T}},m(P)\})\rvert$	$\lvert Cl_{\mathcal{R}1}(\mathbb{T},\{m(P)\})\rvert$	n	$n\lvert Cl_{\mathcal{R}1}(\{\widehat{\mathbb{T}},m(P)\})\rvert$	$n\lvert Cl_{\mathcal{R}1}(\mathbb{T},\{m(P)\})\rvert$
`dc:title*`	0	14	5,503,170	0	77,044,380
`dc:date*`	0	377	5,172,458	0	1,950,016,666
`foaf:name*`	3	418	4,631,614	13,894,842	1,936,014,652
`foaf:nick*`	0	390	4,416,760	0	1,722,536,400
`rss:link*`	1	377	4,073,739	4,073,739	1,535,799,603
TOTAL	4	1,576	23,797,741	17,968,581	7,221,411,701

1.1b Dataset

C	$\lvert Cl_{\mathcal{R}1}(\widehat{\mathbb{T}},\{m(C)\})\rvert$	$\lvert Cl_{\mathcal{R}1}(\mathbb{T},\{m(C)\})\rvert$	n	$n\lvert Cl_{\mathcal{R}1}(\widehat{\mathbb{T}},\{m(C)\})\rvert$	$n\lvert Cl_{\mathcal{R}1}(\mathbb{T},\{m(C)\})\rvert$
`foaf:Person`	6	4,631	63,271,689	379,630,134	293,011,191,759
`foaf:Document`	1	4,523	6,092,322	6,092,322	27,555,572,406
`rss:item`	0	4,528	5,745,216	0	26,014,338,048
`oboInOwl:DbXref`	0	0	2,911,976	0	0
`rdf:Seq`	2	4,285	2,781,994	5,563,988	11,920,844,290
TOTAL	9	17,967	80,803,197	391,286,444	358,501,946,503

P	$\lvert Cl_{\mathcal{R}1}(\widehat{\mathbb{T}},\{m(P)\})\rvert$	$\lvert Cl_{\mathcal{R}1}(\mathbb{T},\{m(P)\})\rvert$	n	$n\lvert Cl_{\mathcal{R}1}(\widehat{\mathbb{T}},\{m(P)\})\rvert$	$n\lvert Cl_{\mathcal{R}1}(\mathbb{T},\{m(P)\})\rvert$
`rdfs:seeAlso`	2	8,647	113,760,738	227,521,476	983,689,101,486
`foaf:knows`	14	9,269	77,335,237	1,082,693,318	716,820,311,753
`dc:title*`	0	4,621	71,321,437	0	329,576,360,377
`foaf:nick*`	0	4,635	65,855,264	0	305,239,148,640
`foaf:weblog`	7	9,286	55,079,875	385,559,125	511,471,719,250
TOTAL	23	36,458	383,352,551	1,695,773,919	2,846,796,641,506

identifiers. This constitutes our output, which we call *pivotal authoritative closure*.

EVALUATION AND DISCUSSION

We now provide evaluation of the SAOR methodology firstly with quantitative analysis of the importance of authoritative reasoning, and secondly we provide performance measurements and discussion along with insights into the fecundity of each rule w.r.t. reasoning over web data. All experiments are run on one machine with a single Opteron 2.2 GHz CPU and 4 GB of main memory. We provide evaluation on two datasets: we provide complete evaluation for a dataset of 147m statements collected from 665k sources and scale-up experiments running scan-reasoning (rules in $\mathcal{R}0 \cup \mathcal{R}1$) on a dataset of 1.1b statements collected from 6.5m sources; both datasets are from web-crawls using MultiCrawler (Harth, Umbrich et al. 2006). We create a unique set of blank nodes for each graph $\mathcal{G}' \in M(\mathcal{S}_{\mathcal{W}})$ using a function on c and the original blank node label which ensures a one-to-one mapping from the original blank node labels and uniqueness of the blank nodes for a given context c.

To show the effects of ontology hijacking we constructed two T-Boxes with and without authoritative analysis for each dataset. We then ran reasoning on single membership assertions for the top five classes and properties found natively in each dataset. Table 5 summarises the results. Taking `foaf:Person` as an example, with an authoritative T-Box, six statements are output for every input `rdf:type foaf:Person` statement in both datasets. With the non-authoritative T-Box, 388 and 4,631 statements are output for every such input statement for the smaller and larger datasets respectively. Considering that there are 3.25m and 63.33m such statements in the respective datasets, overall output for `rdf:type foaf:Person` input statements alone approach 1.26b and 293b statements for non-authoritative reasoning respectively. With authoritative reasoning we only produce 19.5m and 379.6m statements, a respective saving of 65x and 772x on output statement size.[15]

It should be noted that reasoning on a membership assertion of the top level class (`:Thing`/ `rdfs:Resource`) is very large for both the 147m (234 inferences) and the 1.1b dataset (4251 inferences). For example, in both datasets, there are many `:unionOf` class descriptions with `:Thing` as a member;[16] for the 1.1b dataset, many inferences on the top level class stem from, for example, the OWL W3C Test Repository[17]. Of course we do not see such documents as being malicious in any way, but clearly they would cause inflationary inferences when naïvely considered as part of web knowledge-base.

Next, we present some metrics regarding the first step of reasoning: the separation and in-memory construction of the T-Box. For the 1.1b dataset, the initial scan of all data found 9,683,009 T-Box statements (0.9%). Reducing the T-Box by removing collection statements as described in Section 4.3.1 dropped a further 1,091,698 (11% of total) collection statements leaving 733,734 such statements in the T-Box (67% collection statements dropped) and 8,591,311 (89%) total. Table 6 shows, for membership assertions of each class and property in $\mathcal{C}_{SAOR}$ and $\mathcal{P}_{SAOR}$, the result of applying authoritative analysis. Of the 33,157 unique namespaces probed, 769 (2.3%) had a redirect, 4068 (12.3%) connected but had no redirect and 28,320 (85.4%) did not connect at all. In total, 14,227,116 authority checks were performed. Of these, 6,690,704 (47%) were

Table 6. Authoritative analysis of T-Box statements in 1.1b dataset for each primitive where dropped statements are highlighted in bold

Property	AuthSub	AuthObj	AuthBoth	AuthNone	Total	Drop
`rdfs:subClassOf`	25,076	**583,399**	1,595,850	**1,762,414**	3,966,739	2,345,813
`:onProperty`	1,041,873	-	97,921	-	1,139,843	-
`:someValuesFrom`	681,968	-	217,478	-	899,446	-
`rdf:first`	273,805	-	392,707	-	666,512	-
`rdf:rest`	249,541	-	416,946	-	666,487	-
`:equivalentClass`	574	189,912	162,886	**3,198**	356,570	3,198
`:intersectionOf`	-	-	216,035	-	216,035	-
`rdfs:domain`	5,693	**7,788**	66,338	**79,748**	159,567	87,536
`rdfs:range`	32,338	**4,340**	37,529	**75,338**	149,545	79,678
`:hasValue`	9,903	0	82,853	0	92,756	-
`:allValuesFrom`	51,988	-	22,145	-	74,133	-
`rdfs:subPropertyOf`	3,365	**147**	22,481	**26,742**	52,734	26,888
`:maxCardinality`	26,963	-	-	-	26,963	-
`:inverseOf`	75	52	6,397	**18,363**	24,887	18,363
`:cardinality`	20,006	-	-	-	20,006	-
`:unionOf`	-	-	21,671	-	21,671	-
`:minCardinality`	15,187	-	-	-	15,187	-
`:oneOf`	-	-	6,171	-	6,171	-
`:equivalentProperty`	105	24	187	**696**	1,012	696
Class						
`:FunctionalProperty`	9,616	-	-	**18,111**	27,727	18,111
`:InverseFunctionalProperty`	872	-	-	**3,080**	3,952	3,080
`:TransitiveProperty`	807	-	-	**1,994**	2,801	1,994
`:SymmetricProperty`	265	-	-	**351**	616	351
OVERALL	2,450,020	785,661	3,365,595	1,990,035	8,591,311	2,585,708

negative and 7,536,412 (53%) were positive. Of the positive, 4,236,393 (56%) were blank-nodes, 2,327,945 (31%) were a direct match between namespace and source and 972,074 (13%) had a redirect from the namespace to the source. In total, 2,585,708 (30%) statements were dropped as they could not contribute to a valid authoritative inference. The entire process of separating, analysing and loading the T-Box into memory took 6.47 hours: the most costly operation here is the large amount of HTTP lookups required for authoritative analysis, with many connections unsuccessful after our five second timeout. The process required ~3.5G of Java heap-space and ~10M of stack space.

For the 147m dataset, 2,649,532 (1.7%) T-Box statements were separated from the data, which was reduced to 1,609,958 (61%) after reducing the amount of irrelevant collection statements; a further 536,564 (33%) statements were dropped as they could not contribute to a valid authoritative inference leaving 1,073,394 T-Box statements (41% of original). Loading the T-Box into memory for the 147m dataset took 1.04 hours.

We proceed by evaluating the application of all rules on the 147m dataset. Figure 3 shows performance for reaching an overall fixpoint for application of all rules. Clearly, the performance plateaus after 79 mins. At this point the input statements have been exhausted, with rules in $\mathcal{R}0$ and $\mathcal{R}1$ having been applied to the input data and statements written to the on-disk files for $\mathcal{R}2$ and $\mathcal{R}3$. SAOR now switches over to calculating a fixpoint over the on-disk computed $\mathcal{R}2$ and $\mathcal{R}3$ rules, the results of which become the new input for $\mathcal{R}0$ and $\mathcal{R}1$ and further recursive input to the $\mathcal{R}2$ and $\mathcal{R}3$ files.

Figure 4 shows performance specifically for achieving closure on the on-disk $\mathcal{R}2$ and $\mathcal{R}3$ rules. There are three pronounced steps in the output of statements. The first one shown at (a) is due to inferencing of `:sameAs` statements from rule **rdfp2 (** `:InverseFunctionalProperty` - 2.1m inferences). Also part of the first step are `:sameAs` inferences from rules **rdfp1′** (`:FunctionalProperty` - 31k inferences) and rules **rdfc4*** **(**`:cardinality`/`:maxCardinality` - 449 inferences).

Figure 3. Performance of applying entire ruleset on the 147m statements dataset (without final consolidation step)

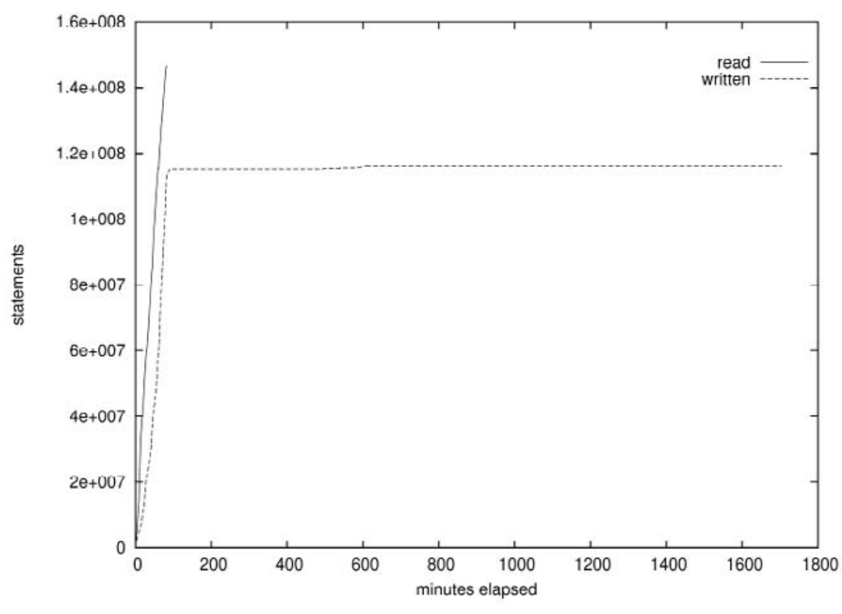
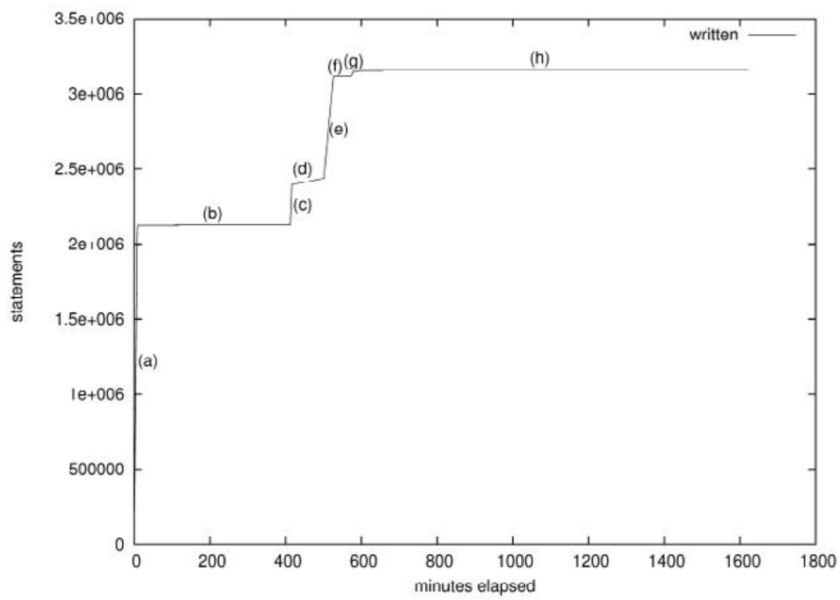


Figure 4. Performance of inferencing over $\mathcal{R}2$ and $\mathcal{R}3$ on-disk indexes for the 147m statements dataset (without final consolidation)



For the first plateau shown at (b), the `:sameAs` equality file is closed for the first time and a local fixpoint is being calculated to derive the initial `:sameAs` statements for future rules; also during the plateau at (b), the second iteration for the `:sameAs` fixpoint (which, for the first time, consolidates the key join variables in files for rules **rdfp2, rdfp1′, rdfc4a, rdfc4b according to all** `:sameAs` statements produced thus far) produces 1,018 new such statements, with subsequent iterations producing 145, 2, and 0 new statements respectively.

The second pronounced step at (c) is attributable to 265k transitive inferences, followed by 1.7k symmetric-transitive inferences. The proceeding slope at (d) is caused by inferences on **rdfc3c** (`:intersectionOf` - 265 inferences) and **rdfp15′** (`:someValuesFrom` - 36k inferences) with rule **rdfp16′** (`:allValuesFrom` - 678k inferences) producing the final significant step at (e). The first complete iteration of the overall fixpoint calculation is now complete.

Since the first local `:sameAs` fixpoint, 22k mostly `rdf:type` statements have been written back to the cardinality rule files, 4 statements to the `:InverseFunctionalProperty` file and 14 to the `:FunctionalProperty` file. Thus, the `:sameAs` fixpoint is re-executed at (f), with no new statements found. The final, minor, staggered step at (g) occurs after the second `:sameAs` fixpoint when, most notably, rule **rdfp4** (`:Transitive-Property`) produces 24k inferences, rule **rdfc3c** (`:intersectionOf`) produces 6.7k inferences, and rule **rdfp16′** (`:allValuesFrom`) produces 7.3k new statements.

The final, extended plateau at (h) is caused by rules which produce/consume `rdf:type` statements. In particular, the fixpoint encounters `:allValuesFrom` inferencing producing a minor contribution of statements (≤2) which lead to an update and re-execution of `:allValuesFrom` inferencing and `:intersectionOf` reasoning. In particular, `:allValuesFrom` required 66 recursive iterations to reach a fixpoint. We identified the problematic data as follows:

```
> @prefix veml: <http://www.icsi.berkeley.edu/snarayan/VEML.owl#>
> @prefix verl: <http://www.icsi.berkeley.edu/snarayan/VERL.owl#>
> @prefix data: <http://www.icsi.berkeley.edu/snarayan/meeting01.owl#>
> ...
> FROM veml: (T-BOX):
> veml:sceneEvents rdfs:range veml:EventList .
> veml:EventList rdfs:subClassOf _:r1 ; rdfs:subClassOf _:r2 .
> _:r1 :allValuesFrom verl:Event ; :onProperty rdf:first .
> _:r2 :allValuesFrom veml:EventList ; :onProperty rdf:rest .
> FROM data: (A-BOX):
> data:scene veml:sceneEvents ( data:1 , ..., data:65 ) .
> EXAMPLE COLLECTION SNIPPET:
> _:cN rdf:first data:N ; rdf:rest _:cN+1 .
```

From the above data, each iteration of :allValuesFrom reasoning and subsequent subclass reasoning produced:

```
> IN ALL-VALUES-FROM, ITER 0:
> FROM INPUT:
> ( _:c1 ... _:c65) rdf:first (data:1 ... data:65) .
> FROM RANGE:
> _:c1 a veml:EventList .
> OUTPUT ALL-VALUES-FROM, ITER N:
> _:dataN a verl:Event .
> _:cN+1 a veml:EventList .
> FROM SUBCLASS ON ABOVE
> ADDED TO ALL-VALUES-FROM, ITER N+1:
> _:cN+1 rdf:type _:r1 ; rdf:type _:r2.
```

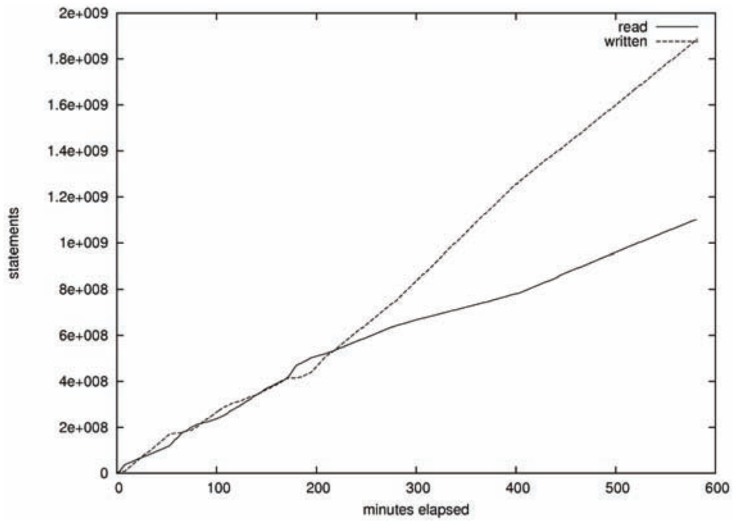
In particular, a small contribution of input statements requires a merge-sort and re-scan of the file in question. This could indeed be solved by implementing binary-search lookup functionality over the sorted files for small input from a previous round; however, this would break with our initial aim of performing reasoning using only the primitives of file-scanning and multi-way merge-sort.

Finally in the reasoning process, we must perform consolidation of the input data and the output inferred statements according to the `:sameAs` index produced in the previous step. The first step involves sorting the input and inferred data according to natural SPOC order; the process took 6.4 hours and rewrote 35.4m statements into pivotal form. The second step involves subsequent sorting of the data according to inverse OPSC order; the process took 8.2 hours and rewrote 8.5m statements. The expense of these steps is primarily attributable to applying multi-way merge-sorting over all data in both sorting orders.

Although the degradation of performance related to the on-disk fixpoint computation of ruleset $\mathcal{R}2 \cup \mathcal{R}3$ is significant, if one is prepared to trade completeness (as we define it) for computational efficiency, the fixpoint calculation can be restrained to only perform a small, known amount of iterations (e.g., inferencing of the majority of statements in Figure 4 takes place over approx. 3 hours). Only minute amounts of inferred statements are produced in latter iterations of the fixpoint.

Further still, most inferences are produced after the initial scan which takes approx. 79 minutes. Thus, even after application of only $\mathcal{R}0$ and $\mathcal{R}1$ rules, the majority of inferencing has been conducted. This simpler more practical reasoning subset exhibits linear scale, as is visible for the first stage of Figure 3 prior to the on-disk computations. Along these lines, we present in Figure 5 the performance of applying rules $\mathcal{R}0$ and $\mathcal{R}1$ to the 1.1b statement dataset, in one scan, with respect to the T-Box derived from that dataset as described above. In particular, we refer to the linear trend present; upon inspection, one can see that minor slow-down in the rate of statements read is attributable to an increased

Figure 5. Performance of applying ruleset $\mathcal{R}0 \cup \mathcal{R}1$ on the 1.1b dataset



throughput in terms of output statements (disk write operations).

Finally, Table 7 lists the number of times each rule was fired for reasoning on the 1.1b dataset, reasoning using only $\mathcal{R}0 \cup \mathcal{R}1$ on the 147m dataset and also of applying all rules to the 147m dataset. Again, from both Figure 3 and Table 7 we can deduce that the bulk of current web reasoning is covered by those rules ($\mathcal{R}0 \cup \mathcal{R}1$) which exhibit linear scale.

RELATED WORK

OWL reasoning, specifically query answering over OWL Full, is not tackled by typical DL Reasoners; such as FaCT++ (Haarslev and Möller 2003; Tsarkov and Horrocks 2006), RACER (Haarslev and Möller 2003) or Pellet (Sirin, Parsia et al. 2007); which focus on complex reasoning tasks such as subsumption checking and provable completeness of reasoning. Likewise, KAON2 (Motik 2006), which reports better results on query answering, is limited to OWL-DL expressivity due to completeness requirements. Despite being able to deal with complex ontologies in a complete manner, these systems are not tailored for the particular challenges of processing large amounts of RDF data and particularly large A-Boxes.

Systems such as TRIPLE (Sintek and Decker 2002), JESS[18], or Jena[19] support rule representable RDFS or OWL fragments as we do, but only work in-memory whereas our framework is focused on conducting scalable reasoning using persistent storage.

The OWLIM (Kiryakov, Ognyanov et al. 2005) family of systems allows reasoning over a version of pD* using the TRREE: Triple Reasoning and Rule Entailment Engine. Besides the in-memory version SwiftOWLIM, which uses TRREE, there is also a version offering query-processing over a persistent image of the repository, BigOWLIM, which comes closest technically to our approach. In evaluation on 2 x Dual-Core 2GHz machines with 16GB of RAM, BigOWLIM is claimed to index over 1 bn triples from the LUBM benchmark (Guo, Pan et al. 2005) in just under 70 hours ("BigOWLIM Sys. Doc.", 2006); however, this figure includes indexing of the data for query-answering, and is not directly comparable with our results, and in any case, our reasoning approach strictly focuses on sensible reasoning for web data.

Table 7. Count of number of statements inferred for applying the given ruleset on the given dataset.

Rule	1.1b - $\mathcal{R}0-1$	147M - $\mathcal{R}0-1$	147M - $\mathcal{R}0-3$
		$\mathcal{R}0$	
rdfc0	35,157	6,084	6,084
		$\mathcal{R}1$	
rdfs2	591,304,476	30,203,111	30,462,570
rdfs3′	596,661,696	31,789,905	32,048,477
rdfs7′	156,744,587	27,723,256	27,882,492
rdfs9	1,164,619,890	64,869,593	65,455,001
rdfp3′	562,426	483,204	483,204
rdfp8a′	231,661,554	9,404,319	9,556,544
rdfp8b′	231,658,162	9,404,111	9,556,336
rdfp12a′	8,153,304	23,869	38,060
rdfp12b′	57,116	17,769	25,362
rdfp13a′	5,667,464	11,478	11,478
rdfp13b′	6,642	4,350	4,350
rdfp14a′	98,601	39,422	39,902
rdfp14b′	104,780	43,886	44,390
rdfc1	15,198,615	1,492,395	1,595,293
rdfc2	584,913	337,141	337,279
rdfc3a	115,416	3,075	17,224
rdfc3b	54	8	8
		$\mathcal{R}2$	
rdfp1′	-	-	31,174
rdfp2	-	-	2,097,007
rdfp4	-	-	291,048
rdfp15′	-	-	42,098
rdfp16′	-	-	685,738
rdfc3c	-	-	6,976
rdfc4a	-	-	211
rdfc4b	-	-	246

Some existing systems already implement a separation of T-Box and A-Box for scalable reasoning, where in particular, assertional statements are stored in some RDBMS; e.g. DLDB (Pan and Heflin 2003), Minerva (Zhou, Ma et al. 2006) and OntoDB (Hondjack, Pierra et al. 2007). Similar to our approach of reasoning over web data, (Pan, Qasem et al. 2007) demonstrates reasoning over 166m triples using the DLDB system. Also like us, (and as we had previously introduced in (Hogan, Harth et al. 2007)) they internally choose pivot identifiers to represent equivalent sets of individuals. However, they use the notion of perspectives to support inferencing based on T-Box data; in their experiment they manually selected nine T-Box perspectives, unlike our approach that deals with arbitrary T-Box data from the Web. Their evaluation was performed on a workstation with dual 64-bit CPUs and 10GB main memory on which they loaded 760k documents / 166m triples (14% larger than our 147m statement dataset) in about 350 hrs; however, unlike our evaluation, the total time taken includes indexing for query-answering.

In a similar approach to our authoritative analysis, (Cheng, Ge et al. 2008) introduced restrictions for accepting sub-class and equivalent-class statements from third-party sources; they follow similar arguments to that made in this paper. However, their notion of what we call authoritativeness is based on hostnames and does not consider redirects; we argue that in both cases, e.g., use of PURL services[20] is not properly supported: (i) all documents using the same service (and having the same namespace hostname) would be 'authoritative' for each other, (ii) the document cannot be served directly by the namespace location, but only through a redirect. Indeed, further work presented in (Cheng and Qu 2008) introduced the notion of an *authoritative description* which is very similar to ours. In any case, we provide much more extensive treatment of the issue, supporting a much more varied range of RDF(S)/OWL constructs.

One promising alternative to authoritative reasoning for the Web is the notion of "context-dependant" or "quarantined reasoning" introduced in (Delbru, Polleres et al. 2008), whereby inference results are only considered valid within the given context of a document. As opposed to our approach whereby we construct one authoritative model for all web data, their approach uses a unique model for each document, based on implicit and explicit imports of the document; thus, they would infer statements within the local context which we would consider to be non-authoritative. However, they would miss inferences which can only be conducted by considering a merge of documents, such as transitive closure or equality inferences based on inverse-functional properties over multiple documents. Their evaluation was completed on three machines with quad-core 2.33GHz and 8GB main memory; they claimed to be able to load, on average, 40 documents per second.

CONCLUSION AND FUTURE WORK

We have presented SAOR: a system for performing reasoning over web data based on primitives known to scale: file-scan and sorting. We maintain a separate optimised T-Box index for our reasoning procedure. To keep the resulting knowledge-base manageable, both in size and quality, we made the following modifications to traditional reasoning procedures:

- only consider a positive fragment of OWL reasoning;
- analyse the authority of sources to counter ontology hijacking;
- use pivot identifiers instead of full materialisation of equality.

We show in our evaluation that naïve inferencing over web data leads to an explosion of materialised statements and show how to prevent this explosion through analysis the authority of data sources. We also present metrics relating to the most productive rules with regards inferencing on the Web.

Although SAOR is currently not optimised for reaching full closure, we show that our system is suitable for optimised computation of the approximate closure of a web knowledge-base w.r.t. the most commonly used RDF(S) and OWL constructs. In our evaluation, we showed that the bulk of inferencing on web data can be completed with two scans of an unsorted web-crawl.

Future work includes investigating possible distribution methods: indeed, by limiting our tool-box to file scans and sorts, our system can be implemented on multiple machines, as-is, according to known distribution methods for our foundational operations.

ACKNOWLEDGMENT

A preliminary version of this article has been accepted at ASWC 2008 (Hogan, Harth et. al. 2008). SAOR: Authoritative Reasoning for the Web. Proceedings of the 3rd Asian Semantic Web Conference (ASWC 2008). Bankok, Thailand.

Compared to that version, we have added significant material. The added contributions in this version include (i) a better formalisation of authoritative reasoning, (ii) improvements in the algorithms, and (iii) respectively updated experimental results with additional metrics on a larger dataset.

We thank the anonymous reviewers of this and related articles for their valuable feedback.

This work has been supported by Science Foundation Ireland project Lion (SFI/02/CE1/I131), European FP6 project inContext (IST-034718), COST Action "Agreement Technologies'' (IC0801) and an IRCSET Postgraduate Research Scholarship.

REFERENCES

(2006, October). *BigOWLIM: System Documentation*.

Bechhofer, S., F. van Harmelen, et al. (2004, February). OWL Web Ontology Language Reference.

Bechhofer, S., & Volz, R. (2004). Patching Syntax in OWL Ontologies. *International Semantic Web Conference*, 3298, 668-682, Hiroshima, Japan. Springer.

Brickley, D., & Guha, R. V. (2004, February). *RDF Vocabulary Description Language 1.0: RDF Schema*.

Brickley, D., &Miller, L. (2007, November). FOAF Vocabulary Specification 0.91.

Cheng, G., Ge, W., et al. (2008). Searching Semantic Web Objects Based on Class Hierarchies. Proceedings of Linked Data on the Web Workshop.

Cheng, G. & Qu, Y. (2008). *Term Dependence on the Semantic Web. International Semantic Web Conference (pp. 665-680),* Karlsruhe, Germany.

de Bruijn, J. (2008). *Semantic Web Language Layering with Ontologies, Rules, and Meta-Modeling.* University of Innsbruck.

de Bruijn, J., & Heymans, S. (2007). *Logical Foundations of (e)RDF(S): Complexity and Reasoning.* 6th International Semantic Web Conference (pp. 86-99), Busan, Korea.

Delbru, R., A. Polleres, et al. (2008). Context dependent reasoning for semantic documents in Sindice. *Proceedings of the 4th International Workshop on Scalable Semantic Web Knowledge Base Systems* (SSWS 2008), Karlsruhe, Germany.

Fensel, D., & van Harmelen, F. (2007). Unifying Reasoning and Search to Web Scale. *IEEE Internet Computing*, *11*(2). 96, 94-95.

Ghilardi, S., Lutz, C., et al. (2006). Did I Damage My Ontology? A Case for Conservative Extensions in Description Logics. *Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning* (pp. 187-197), Lake District of the UK..

Grau, B. C., Horrocks, I., et al. (2006). *Next Steps for OWL*. OWL: Experiences and Directions Workshop, Athens, Georgia, USA.

Grosof, B., Horrocks, I., et al. (2004). Description Logic Programs: Combining Logic Programs with Description Logic. *13th International Conference on World Wide Web.*

Guha, R. V., McCool, R., et al. (2004). Contexts for the Semantic Web. *3rd International Semantic Web Conference*, Hiroshima.

Guo, Y., Pan, Z., et al. (2005). LUBM: A benchmark for OWL knowledge base systems. *Journal of Web Semantics, 3*(2-3), 158-182.

Gutiérrez, C., Hurtado, C., et al. (2004). Foundations of Semantic Web Databases. *23rd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Paris.

Haarslev, V., & Möller, R. (2003). Racer: A Core Inference Engine for the Semantic Web. *International Workshop on Evaluation of Ontology-based Tools.*

Harth, A., & Decker, S. (2005). Optimized Index Structures for Querying RDF from the Web. *3rd Latin American Web Congress* (pp. 71-80). Buenos Aires, Argentina: IEEE Press.

Harth, A., Umbrich, J., et al. (2006). MultiCrawler: A Pipelined Architecture for Crawling and Indexing Semantic Web Data. *International Semantic Web Conference* (pp. 258-271).

Hayes, P. (2004, February). *RDF Semantics.*

Hogan, A., Harth, A., et al. (2007). Performing Object Consolidation on the Semantic Web Data Graph. *1st I3 Workshop: Identity, Identifiers,* Identification Workshop.

Hogan, A., Harth, A., et al. (2008). SAOR: Authoritative Reasoning for the Web. *Proceedings of the 3rd Asian Semantic Web Conference* (ASWC 2008), Bankok, Thailand.

Hondjack, D., Pierra, G., et al. (2007). OntoDB: An Ontology-Based Database for Data Intensive Applications. *Proceedings of the 12th International Conference on Database Systems for Advanced Applications* (pp. 497-508), Bangkok, Thailand.

Horrocks, I., & Patel-Schneider, P. F. (2004). Reducing OWL entailment to description logic satisfiability. *Journal of Web Semamtics, 1*(4), 345-357.

Jiménez-Ruiz, E., Grau, B. C., et al. (2008). Safe and Economic Re-Use of Ontologies: A Logic-Based Methodology and Tool Support. *Proceedings of the 21st International Workshop on Description Logics* (DL2008), Dresden, Germany.

Kiryakov, A., Ognyanov, D., et al. (2005). OWLIM - A Pragmatic Semantic Repository for OWL. *Web Information Systems Engineering Workshops* (pp. 182-192), New York, USA.

Kunkle, D., & Cooperman, G. (2008). Solving Rubik's Cube: disk is the new RAM. *Communications of the ACM, 51*(4), 31-33.

Lloyd, J. W. (1987). *Foundations of Logic Programming* (2nd ed.). Springer-Verlag.

Lutz, C., Walther, D., et al. (2007). Conservative Extensions in Expressive Description Logics. IJCAI 2007, *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. Hyderabad, India: 453-458.

Motik, B. (2006). *Reasoning in Description Logics using Resolution and Deductive Databases.* Karlsruhe, Germany, Forschungszentrum Informatik.

Motik, B. (2007). On the Properties of Metamodeling in OWL. *Journal of Logic and Computation, 17*(4), 617-637.

Muñoz, S., Pérez, J., et al. (2007). *Minimal Deductive Systems for RDF. ESWC* (pp. 53-67).

Pan, Z., & Heflin, J. (2003). DLDB: Extending Relational Databases to Support Semantic Web Queries. PSSS1 - Practical and Scalable Semantic Systems, *Proceedings of the First International Workshop on Practical and Scalable Semantic Systems.* Sanibel Island, Florida, USA.

Pan, Z., Qasem, A., et al. (2007). Hawkeye: A Practical Large Scale Demonstration of Semantic Web Integration. *Proceedings of On the Move to Meaningful Internet Systems Workshops* (OTM-II) (pp. 1115-1124).

Patel-Schneider, P. F., & Horrocks, I. (2004, February). *OWL Web Ontology Language Semantics and Abstract Syntax*

Prud'hommeaux, E., & Seaborne, A. (2008, January). *SPARQL Query Language for RDF.*

Sintek, M., & Decker, S. (2002). TRIPLE - A Query, Inference, and Transformation Language for the Semantic Web. *1st International Semantic Web Conference* (pp. 364-378).

Sirin, E., Parsia, B., et al. (2007). Pellet: A practical OWL-DL reasoner. *Journal of Web Semantics, 5*(2), 51-53.

Smith, M. K., Welty, C., et al. (2004, February). *OWL Web Ontology Language Guide.*

ter Horst, H. J. (2005). Combining RDF and Part of OWL with Rules: Semantics, Decidability, Complexity. *4th International Semantic Web Conference* (pp. 668-684).

ter Horst, H. J. (2005). Completeness, decidability and complexity of entailment for RDF Schema ans a semantic extension involving the OWL vocabulary. *Journal of Web Semantics,* 3, 79-115.

Theoharis, Y., Christophides, V., et al. (2005). Benchmarking Database Representations of RDF/S Stores. *Proceedings of the Fourth International Semantic Web Conference* (pp. 685-701). Galway, Ireland.

Tsarkov, D., & Horrocks, I. (2006). FaCT++ Description Logic Reasoner: System Description. *International Joint Conf. on Automated Reasoning* (pp. 292-297).

Wang, T. D., Parsia, B., et al. (2006). A Survey of the Web Ontology Landscape. *Proceedings of the 5th International Semantic Web Conference* (ISWC 2006) (pp. 682-694), Athens, GA, USA.

Wu, Z., Eadon, G., et al. (2008). Implementing an Inference Engine for RDFS/OWL Constructs and User-Defined Rules in Oracle. *24th International Conference on Data Engineering. IEEE*.

Zhou, J., Ma, L., et al. (2006). Minerva: A Scalable OWL Ontology Storage and Inference System. *Proceedings of The First Asian Semantic Web Conference* (ASWC), (pp. 429-443), Beijing, China.

ENDNOTES

1. Throughout this article, we assume that http://www.w3.org/2002/07/owl# is the default namespace with prefix ":", i.e. we write e.g. just "`:Class`", "`:disjointWith`", etc. instead of using the commonly used `owl:` prefix. Other prefixes such as `rdf:`, `rdfs:`, `foaf:` are used as in other common documents. Moreover, we often use the common abbreviation 'a' as a convenient shortcut for `rdf:type`.
2. http://www.dajobe.org/2004/01/turtle/
3. Unlike some other rule systems for RDF, the most prominent of which being CONSTRUCT statements in SPARQL, we forbid blank nodes; i.e., we forbid existential variables in rule consequents which would require the "invention" of blank nodes.
4. http://xmlns.com/foaf/spec/index.rdf
5. Tim (now the same entity as the W3C) is asserted to be a member of the two disjoint classes: `foaf:Person` and `foaf:Organization`.
6. In (ter Horst 2005), rules using RDF collection constructs were not included (such as our rules **rdfc0,rdfc1,rdfc3*)** as they have variable antecedent-body length and, thus, can affect complexity considerations. It was informally stated that `:intersectionOf` and `:unionOf` could be supported under pD* through reduction into subclass relations; however no rules were explicitly defined and our rule **rdfc3b** could not be supported in this fashion. We support such rules here since we are not so concerned for the moment with theoretical worst-case complexity, but are more concerned with the practicalities of web-reasoning.
7. A similar example from the Web can be found at http://thesauri.cs.vu.nl/wordnet/rdfs/wordnet2b.owl.
8. In any case, as we will see in Section 3.4, our application of authoritative analysis would not allow such arbitrary third-party re-definition of core RDF(S)/OWL constructs.
9. Here, slightly abusing XML terminology by "namespace" of a URI we mean the prefix of the URI obtained from stripping off the final NCname
10. See Appendix A&B of http://www.w3.org/TR/swbp-vocab-pub/
11. Includes some RDF collection fragments which may not be part of a class description

[12] We expect that a caching on-disk index would work well considering the distribution of membership assertions for classes and properties in web data; there would be a high hit-rate for the cache.

[13] In N-Quads format: c.f. http://sw.deri.org/2008/07/n-quads/

[14] This is from incorrect use of the FOAF ontology by prominent exporters. We refer the interested reader to (Hogan, Harth et. al. 2007)

[15] For example, the document retrievable from http://pike.kw.nl/files/documents/pietzwart/RDF/PietZwart200602.owl defines super-classes/-properties for all of the FOAF vocabulary.

[16] Thirty-four such `:unionOf` class descriptions can be found in http://colab.cim3.net/file/work/SICoP/ontac/reference/ProtegeOntologies/COSMO-Versions/TopLevel06.owl; fifty-five can be found in http://lsdis.cs.uga.edu/ oldham/ontology/wsag/wsag.owl

[17] http://www.w3.org/2002/03owlt/

[18] http://herzberg.ca.sandia.gov/

[19] http://jena.sourceforge.net/

[20] http://purl.org/

This work was previously published in the International Journal on Semantic Web & Information Systems, Vol. 5, Issue 2, edited by A. Sheth, pp. 49-90, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 7.20
A Framework for Integrating the Social Web Environment in Pattern Engineering

Pankaj Kamthan
Concordia University, Canada

ABSTRACT

In the last decade, patterns have emerged as a notable problem-solving approach in various disciplines. This paper aims to address the communication requirements of the elements of pattern engineering (namely, actors, activities, and artifacts) in general and the pattern realization process in particular. To that regard, a theoretical framework using the Social Web as the medium is proposed and its implications are explored. The prospects of using the Social Web are analyzed by means of practical scenarios and concrete examples. The concerns of using the Social Web related to cost to actors, decentralization and distribution of control, and semiotic quality of representations of patterns are highlighted. The directions for future research including the use of patterns for Social Web applications, and the potential of the confluence of the Social Web and the Semantic Web for communicating the elements of pattern engineering, are briefly explored.

INTRODUCTION

The reliance on the knowledge garnered from past experience and expertise is important for any creative endeavor. A pattern is one such type of conceptually reusable knowledge (Buschmann, Henney, & Schmidt, 2007b). From their origins in urban planning and architecture in the 1970s (Alexander, Ishikawa, & Silverstein, 1977; Alexander, 1979), followed by object-oriented software design in the late 1980s and the early 1990s (Gamma et al., 1995), patterns have found applications in various domains of interest (Rising, 2000; Henninger & Corrêa, 2007). For novices, patterns are means of guidance; for experts, they are means of reference. The use of patterns has, for example, enabled the construction of high-quality distributed software architectures (Buschmann, Henney, & Schmidt, 2007a), electronic commerce applications (Kamthan & Pai, 2008), mobile interaction design (Ballard, 2007), secure systems software (Schumacher et al., 2006), use case models (Kamthan, 2009), and Web Applications (Kamthan, 2008), to name a few.

The human-centric nature of patterns has been known for some time (Coplien, 1996; Schumacher et al., 2006). For its broad acceptance and use, the knowledge in form of patterns needs to be explicably communicated to its actors.

The Social Web, or as it is more commonly referred to by the pseudonym Web 2.0 (O'Reilly, 2005; White, 2007), is the perceived evolution of the Web in a direction that is driven by 'collective intelligence' (Engelbart, 1995), realized by information technology, and characterized by user participation, openness, and network effects. The purpose of this paper is to assess the viability of the Social Web environment in serving as an ecosystem for many-to-many asynchronous and synchronous communication of the elements of pattern engineering in general and the pattern realization process in particular. For the sake of this paper, the Social Web *environment* includes Social Web technologies, applications based on those technologies, and tools for managing both.

The rest of the paper is organized as follows. The background and related work necessary for the discussion that follows is first outlined. This is followed by introduction of a theoretical framework for communicating the elements of pattern engineering via the Social Web (namely, SW4PE) that includes identifying and classifying actors of patterns, a model for the pattern realization process, and communication requirements for pattern engineering. Then, a detailed analysis of the prospects and concerns of using the Social Web for communicating the elements of pattern engineering along different dimensions is carried out. In particular, the role of 'collective intelligence' and of the technologies/applications underlying the Social Web including blogs, folksonomy, mashups, microformats, podcasting, social bookmarking, social networking, and Wikis, is highlighted. Next, challenges and directions for future research are outlined. Finally, concluding remarks are given.

BACKGROUND AND RELATED WORK

This section presents a synopsis of terminology specific to patterns and a perspective of related work. In particular, limitations of the current media towards communicating the elements of pattern engineering are highlighted.

A Terminological Overview of the Pattern Space

There is currently no standard or a reference model for terminology related to patterns. Therefore, for the definition of the members in the *pattern space*, this section relies on selected publications (Appleton, 1997; Meszaros & Doble, 1998; Buschmann, Henney, & Schmidt, 2007b) that can be considered as authoritative.

A *pattern* is defined as an empirically proven solution to a recurring problem that occurs in a particular context. There are several possible views of a pattern. From a structural viewpoint, a pattern is typically described using an ordered list of elements that are labeled as (pattern) name, author, context, problem, forces, solution, examples, and related patterns. At times, the labels may vary across community, and other (optional) elements, such as those related to *metadata*, may be included to enrich the description.

The name element of a pattern is an evocative, often a noun-phrase, metaphor reflecting the nature of the solution; the author element gives the identity of the pattern author(s); the context element provides the situation or pre-conditions within which the problem occurs; the forces element provides the constraints that are resolved to arrive at a solution; the solution element provides an abstract, general, and reusable solution to the problem and is shown to work in practice via an examples element; and the related patterns element outlines any other pattern(s) to which a pattern is related to in some way. It is this structure that makes patterns more practical in their applicability compared to other expert bodies of knowledge such as principles, guidelines (Wesson & Cowley, 2003), and heuristics.

A pattern is usually referred to by its name. In this paper, the name of a pattern is listed in uppercase in order to distinguish it from the surrounding text.

There are other members in the pattern space closely related to a pattern. An *anti-pattern* is a pattern that suggests a 'negative' solution to a given problem, and occurs when the context of the problem is not understood or the underlying forces are not optimally balanced. A *patlet* is a 'simplified' description of a pattern, providing only a short statement of the problem and solution, and does not include other elements. A *pattern thumbnail* is similar to a patlet except that it is usually accompanied with a picture of the solution. The purpose of both patlet and a pattern thumbnail is to briefly introduce a pattern (without engaging in details) so that a pattern reader can make an informed decision whether or not to read any further. This can be particularly significant when there is a large collection of patterns to select from.

It is rarely the case that a pattern exists in isolation. Indeed, a pattern is often intimately related to other patterns in many different ways. A *pattern language* is a network of patterns that are intimately related to each other by a common goal and collectively solves a larger problem than that possible by any individual pattern. The collection of patterns in a pattern language when taken together forms a 'vocabulary' that can be used by the actors for communication.

A sequential reading through the text of a lengthy pattern language may not be sufficient to gain an overall picture that is necessary for its understanding and subsequent use. A compact graphical representation can be useful in such a case. For a given pattern language, a *pattern language map* is a visual presentation of patterns and their relationships.

A *pattern management system* (PMS) is an interactive software system with responsibilities that include archiving a selected collection of patterns that could evolve (added, deleted, or modified), facilitating the discovery of those patterns via navigation or searching, and rendering those patterns on a user agent. For example, a PMS could be based on a client-server environment of the Web (Kamthan, 2008).

A *pattern realization process* (PRP) is a collection of activities and their interrelationships for specifying a pattern. The activities themselves can either be individual or social.

Finally, *pattern engineering* (PE) is a systematic and disciplined approach to the definition, subsequent use and maintenance, and interface to humans, machines, and other entities of knowledge of a member of the pattern space within the given constraints of available resources. A PRP is a part of PE.

Medium for Communicating the Elements of Pattern Engineering and the Human/Social Factors

Every means of communication requires a medium. From 1970s to about mid-1990s, patterns were essentially restricted to print medium like commercial-only books or event proceedings. According to an empirical study (Henninger & Corrêa, 2007), the print medium continues to dominate as a major channel for publishing patterns.

However, a print medium provides limited opportunities for communication. In particular, it enables only a one-to-many communication paradigm; there is no interaction; information in modalities like animation, audio, video, or three-dimensional graphics can not be communicated; there is notion of presentation, not of representation; a reuse can only be realized through citation or duplication; entities (such as books) are essentially isolated from each other; it is not possible to provide multiple views of the same information (for example, change the level of magnification) or, on-demand, present information at different levels of abstraction; and there is no support for hypertext.

In the past decade or so, the electronic (digital) medium, particularly the distributed environment of the Internet and the Web, has proved to be a useful vehicle for communicating elements of PE in different sensory modalities. The use of electronic mail (e-mail) and Internet Relay Chat (IRC) services, both of which predate the Web, has conventionally been made for communicating asynchronously and synchronously, respectively. As indicated by surveys (Deng, Kemp, & Todd, 2005; Henninger & Corrêa, 2007), mailing lists and newsgroups dedicated to patterns have spawned and various domain-specific portals/repositories for patterns, usually equipped with navigation and search mechanisms, have been established.

However, mailing lists and newsgroups provide limited capabilities for organizing patterns (Manolescu et al., 2007). Furthermore, a conventional repository is limited by one or more of the following issues: it tends to be prescriptive; it still only enables a one-to-many communication paradigm; it usually only provides an author-view of patterns where the role of a reader is that of a mere observer, not a contributor; and, in general, any human or social relationships in PE are not always made explicit. This is prohibitive to the advancement of patterns and could potentially undermine their significance.

To alleviate some of these issues, human and social aspects of the PRP and subsequent deployment, in addition to technical considerations, is necessary. There have been some partial efforts in that direction such as the use of Wikis (Weiss & Birukou, 2007) but an in-depth analysis has not been carried out. This lends one of the motivations for this paper.

A FRAMEWORK FOR COMMUNICATING THE ELEMENTS OF PATTERN ENGINEERING USING THE SOCIAL WEB ENVIRONMENT

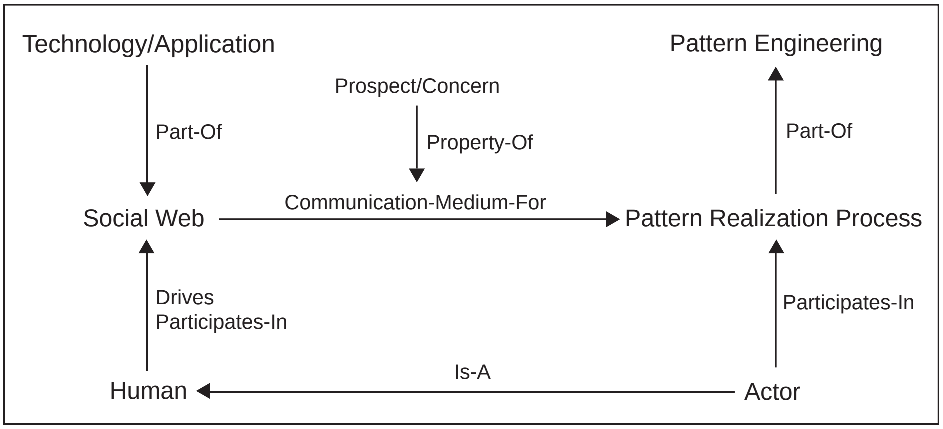
This section posits SW4PE, a framework for communicating the elements of PE via the Social Web illustrated in Figure 1. In the forthcoming sections, the actors of patterns are identified and classified, the details of the PRP are specified, the actor-specific requirements for communicating the elements of PE are outlined and, based on these, the prospects/concerns of integrating the Social Web technologies/applications in PE in general and PRP in particular are discussed in detail.

A Model for Actors of Patterns

According to the CLEAR TARGET AUDIENCE pattern (Meszaros & Doble, 1998) and the CONSISTENT "WHO" pattern (Harrison, 2003), the external entities to which a pattern is being communicated need to be recognized. An *actor* is a person who has interest in a pattern for some purpose. Based upon their *roles*, the possible actors of patterns can be identified and classified as follows:

- **Producer:** *Pattern Author* (responsible for authoring a pattern), *Pattern Shepherd* (responsible for inspection and feedback on a pattern), *Pattern Writers' Workshop Participant* (responsible for inspection and feedback on a pattern), *Pattern Engineer* (responsible for providing means for representation and presentation of a pattern), and *Pattern Administrator* (responsible for maintenance and management of patterns).
- **Consumer:** *Pattern Reader* (target for perceiving a pattern) and *Pattern User* (target for using a pattern).

Figure 1. A high-level view of SW4PE



The actor classification scheme has a few properties. In it, the actors are not necessarily mutually exclusive. For example, there is a generalization-specialization relationship between a pattern reader and a pattern shepherd and the same between a pattern reader and a pattern user; however, the converse in both cases is not necessarily the case. The same person can also take upon different roles, and the same role can be taken upon by different persons. For example, a person casually reading a pattern plays the role of a pattern reader but given the write permission can (at least in part) play the role of a pattern administrator. The Social Web has made the boundaries among actors increasingly fuzzy as a professional pattern consumer can become a voluntary pattern co-producer, or a pattern 'prosumer' (Shuen, 2008).

Remarks

The actor classification scheme can be granularized further if needed. For example, a pattern reader and a pattern user could both be categorized further into *novice* and *expert*, or into *putative*, *potential*, or *future*. Also, pattern user could be labeled as a secondary actor, while all other actors as primary actors.

The actor classification scheme can also be extended. For example, a pattern reader could be generalized to a pattern percipient. A means of creating a taxonomy of actors for a general software system is available (Alexander, 2005), the discussion of which is beyond the scope of this paper.

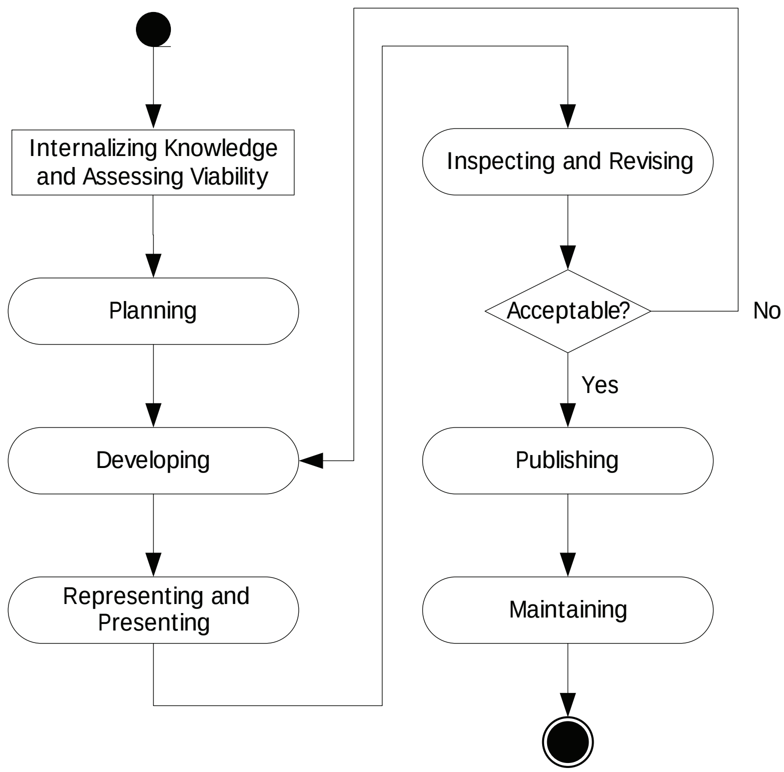
A Model for a Human-Centric and Evolutionary Pattern Realization Process

The PRP is an actor-centric, iterative, and incremental process, the resulting product of which is a pattern. A *workflow* of PRP is a high level organization unit that consists of one or more activities. As shown in Figure 2, there are a number of workflows in PRP, including (1) *planning*, (2) *developing*, (3) *representing and presenting*, (4) *inspecting and revising*, (5) *publishing*, and (6) *maintaining*. These are prefixed by (0) *internalizing knowledge and assessing viability*, which is a prerequisite to the workflows that follow.

[0]. Internalizing Knowledge and Assessing Viability

The two recommended approaches for acquiring internal knowledge are *individual* and

Figure 2. The sequence of workflows in a pattern realization process



sociological. In an individual approach, there is a single pattern author who relies on personal experiences *and* extrospections (observations) based on others' experiences; in a sociological approach, there are multiple pattern authors relying on each other's experiences.

For the sake of discussion, consider the individual approach. It is assumed in this paper that practice involves action (doing). The practice over a period of time by a person of repeatedly solving a problem in some domain leads to experiences. Among other factors, the ensemble of the experiences themselves and of the retrospective after each experience (and perhaps extrospections based on others' experiences), lead to insight. The experiences also lead to acquirement of skill. It is then insight *and* skill together that form expertise. The a posteriori, situated, and experiential knowledge internalized by the person is either implicit or tacit. The person may voluntarily decide to share the implicit knowledge with others. For that, the implicit knowledge needs to be externalized (articulated) to explicit knowledge. At that point, the person takes upon the role of a (potential) pattern author.

Based on the expertise and research of existing pattern base, the pattern author determines the viability of proposing a 'new' pattern (also known as the *proto-pattern*) to the pattern community at-large (including target pattern readers and potential pattern users). As part of this workflow, the pattern author also checks for the existence of patterns that may be similar or variants of the one being proposed. This workflow concludes with a favorable decision to proceed with the definition of a proto-pattern.

[1]. Planning

In order for the PRP to be productive, the definition of the proto-pattern requires appropriate planning by the pattern author(s). The planning needs to include an assessment of the availability of resources including time, effort, expert body of knowledge, and tools. In case of multiple pattern authors, schedules for meetings also need to be decided upon.

There is a cost factor associated with the PRP. In particular, if the pattern author seeks a 'formal' analysis of the proto-pattern, there is cost involved in traveling to the Pattern Languages of Programming (PLoP) 'family' of conferences. There can perhaps also be ancillary cost related to publishing and administering the proto-pattern. The exceptions to this are the voluntary time and effort of the pattern authors and pattern shepherds.

There are a few non-commercial resources at the pattern author's disposal to assist in the PRP. These include expert bodies of knowledge like guidelines (Buschmann, Henney, & Schmidt, 2007b) and patterns (Meszaros & Doble, 1998; Harrison, 2003) for describing patterns, and patterns for shepherding (Harrison, 2000) and Writers' Workshops (Coplien, 2000). There are also tools for representing and presenting the patterns, the discussion of which is delegated to later sections.

[2]. Developing

For the purpose of referencing, the pattern author assigns an evocative name to the proto-pattern. From an analysis of the given information, the pattern author then abstracts the problem and, guided by previous personal experiences and extrospections based on others' experiences from [0], explicitly places the problem in a specific context that reflects the scope of the problem.

Next, a general and (conceptually) reusable solution for the problem needs to be devised. This is achieved via abstraction of instances of the solution from personal experiences and extrospections based on others' experiences from [1]. The solution describes both the process *and* the thing (created by the process). The purpose of the process aspect of the solution is pedagogy. It is likely that the problem has more than one solution, and that each solution has its own advantages and disadvantages. The 'best' solution is chosen based on an optimal balance (or equilibrium) of forces (constraints), which are usually the desirable quality attributes of the solution (Lea, 1994). From an examination of previous work, it appears that the means of achieving this balance are not given.

Since even the best solution is not absolute, the pattern author examines the implications (consequences) of applying the solution. The consequences could include forces that are not entirely resolved as well as new forces that may arise. This may lead to the need for other pattern(s), which is the inception of a pattern language.

The solution proposed by the proto-pattern must be generative (Lea, 1994), that is, it must be demonstrably proven to work. Therefore, based on the 'rule of three' (Meszaros & Doble, 1998), the pattern author elicits three solution instances or examples that best demonstrate the feasibility of the proposed solution. The examples could possibly be from earlier personal experiences and extrospections based on others' experiences from [0]. However, since the proto-pattern is based on empirical knowledge, it is prone to subjectivity. Therefore, to lend some degree of objectivity, these examples should not exclusively be internal, that is, they should not be all from the pattern author's personal experiences. In other words, there must be at least one external example.

Finally, the proto-pattern is placed in its social context. To do that, related patterns (if any) along with their relationships to the proto-pattern are listed.

[3]. Representing and Presenting

In order to become explicit, the information in [2] needs a suitable means of representation. A representation can subsequently be presented in one or more ways, in one or more sensory modalities, to make it perceptible to an actor.

In this workflow, the pattern author selects one of the available means for representing and presenting the proto-pattern (that are made possible by a pattern engineer), keeping the needs of the readership (Meszaros & Doble, 1998; Harrison, 2003) into consideration. The possible means for representing and presenting a proto-pattern can vary across the spectrum of formality (informal, semi-formal, formal), modes (text, graphics), open/closed technology, and so on.

For example, proto-patterns (and even an entire proto-pattern language) may be represented (Kamthan & Pai, 2006a) in the Extensible Markup Language (XML) and, depending on the target device, subsequently presented in one of the profiles of the Extensible HyperText Markup Language (XHTML) that is targeted for the Web or in the Portable Document Format (PDF) that is targeted for printing. This marks the end of the first iteration of the PRP.

[4]. Inspecting and Revising

The proto-pattern may go through an informal inspection (a non-anonymous, highly recommended but optional, review process) to evaluate the characteristics of the proto-pattern.

The prime example of inspection includes submission of the proto-pattern to one of the members of the PLoP 'family' of conferences, which leads to shepherding (which is one-on-one mentoring of the author by another person, namely the pattern shepherd, who is familiar with the underlying domain and is experienced in describing patterns) followed by participation in a Writers' Workshop (which is a face-to-face structured peer review process involving domain experts). The inspection may lead to a few iterations of the proto-pattern and thereby a re-visitation of [2] and [3]. At the end of the inspection, the proto-pattern may reach the candidacy of a pattern.

The pattern author, individually or otherwise as a result of the inspection, may associate a rating reflecting the confidence or maturity level of the pattern. Before publication, the pattern author may also optionally include metadata information related to configuration management, copyright/licensing, and so on, in the description of the pattern.

[5]. Publishing

Up until now, the pattern is limited to internal consumption. In order for the pattern to reach a broader community (beyond the pattern author(s), pattern shepherd, and participants of the Writers' Workshop), it needs to be published in a public environment.

The pattern is published in some (usually print and/or electronic) medium that is deemed reachable to the patterns community. The Web in general and the Social Web in particular is one candidate medium for publication of patterns. For example, a Web Application for patterns could be developed in a systematic manner (Kamthan, 2008) that archives and serves desirable patterns.

[6]. Maintaining

If needed, a pattern administrator carries out corrective and/or adaptive maintenance of pattern(s) on a timely basis. Furthermore, these pattern(s) may also be integrated (into a larger collection) and organized (classified and indexed) in some way. This concludes the conventional PRP from the viewpoint of the pattern producers.

A pattern, once published, does not get 'retired' or 'terminated' (in the sense of hardware or software systems). However, for various reasons, it may lose support and go out of use.

Remarks

A few remarks concerning PRP are in order. From the description of PRP, it is evident that PRP is non-linear and its workflows are not necessarily mutually exclusive. PRP also relies on human creativity and can not be completely automated. In some ways, PRP is similar to agile software development methodologies and open source software development processes but it is not as rigorous or formal. There has traditionally been modest involvement (if any) of the pattern consumers in PRP. This changes significantly by the introduction of the Social Web as a medium in the PRP.

A multidisciplinary, interdisciplinary, and participatory methodology called Identification-Development-Refinement (IDR) for realizing interaction design patterns has been proposed previously (Winters & Mor, 2008). However, IDR does not precisely identify actors or provide details of their involvement, rejects shepherding, and its steps are subsumed by that of PRP.

Actor-Specific Requirements for Communicating the Elements of Pattern Engineering

This section lists computing environment-, domain-, and technology-independent requirements for communicating the elements of PE driven by the needs of the actors. These informally collated requirements are identified by the prefix [PE-CR-n], n = 1, 2, 3, 4, and stated as follows:

- **[PE-CR-1]** It should be possible for a pattern author to readily describe a proto-pattern; it should also be straightforward for a pattern shepherd to readily inspect and provide feedback on it (in proximity as well as remotely). In general, a pattern author and a pattern shepherd should be able to collaborate and share a proto-pattern.
- **[PE-CR-2]** It should be possible for a pattern engineer to provide a means to represent and present proto-patterns.
- **[PE-CR-3]** It should be possible for a pattern administrator to manage (including store, retrieve, process (manipulate, transform), modify, and delete) a collection of patterns with minimal effort.
- **[PE-CR-4]** It should be relatively easy for a pattern reader (and a potential pattern user) to be able to locate, read, and understand a pattern and, with appropriate permissions, be able to distribute the pattern. It should be possible for a pattern reader to contact the pattern author and the pattern administrator.

These requirements serve as a guide for the rest of the paper. The use of qualitative terms in the statements [PE-CR-1] – [PE-CR-4] is intentional: their quantification is possible but is beyond the scope of this paper. In the following, symbols [+PE-CR-n] and [–PE-CR-n], n = 1, 2, 3, 4, are used to respectively denote strong and weak conformance to the corresponding requirement.

Communicating Patterns using the Social Web Environment

The Social Web has recently emerged as a perceived extension of the current Web that fosters ‘collective intelligence’ (Engelbart, 1995) and further decentralization. The notion of the apparent ‘humanization’ and ‘socialization’ of the Web is not new and dates back to the early days of the Web. Indeed, amazon.com and eBay are classical exemplars of consumer participation that introduced product review/recommendation and feedback, respectively. The notion of decentralization also has its predecessor in file sharing via Peer-to-Peer (P2P) computing.

However, it appears that there are three primary factors that have brought the vision of the Social Web to a mainstream realization: (1) it enables a many-to-many communication paradigm; (2) the maturation of the underlying technological infrastructure and the availability of its implementations as open source, and (3) the awareness, followed by immense interest and large-scale participation, by the public in general.

This paper advocates retaining the advantages that the Web offers towards communicating the elements of PE and assessing the viability of the Social Web in extending those advantages. The Social Web provides the medium in which human-to-machine-to-human communication takes place to realize human-to-human communication.

Scope of SW4PE: Open Problems in Patterns—Beyond Information Technology and the Web

There are certain impediments that the actors currently face in their dealings with patterns. In the future, it is theoretically possible that some of the issues, like the existence of a meta-index of repositories (Manolescu et al., 2007) or means for more precise search/retrieval might be addressed and even get resolved by technological means within the realm of the Web and its extensions.

Still, there are other pressing issues, like existence of a coordinating body for patterns, standardization of terminology related to pattern engineering body of knowledge (PEBOK), standardization of representations of patterns, or making all patterns as open/freely available content, that are obstacles to pervasiveness of patterns. These issues naturally constrain SW4PE, and are likely to remain beyond the scope of the Web and its foreseeable extensions.

AN ASSESSMENT OF THE PROSPECTS AND CONCERNS FOR COMMUNICATING THE ELEMENTS OF PATTERN ENGINEERING USING THE SOCIAL WEB ENVIRONMENT

In this section, based on the background set forth in the previous sections, the prospects as well as the concerns in deploying the Social Web are each assessed along certain dimensions.

Prospects for Communicating the Elements of Pattern Engineering via the Social Web

There are certain aspects of patterns that make them a natural fit within the environment of Social Web. In this section, the potential of the Social Web for communicating the elements of PE is explored along the lines of [PE-CR-1] – [PE-CR-4] and, to that regard, specific examples are provided.

Collaborating and Sharing

There is need for collaboration and sharing during the different workflows of the PRP. A pattern reflects *shared* understanding of a domain, and can be viewed as a shared resource or 'commons' (Hess & Ostrom, 2007). In the PRP, the transition of a proto-pattern to the status of a pattern inevitably involves collaboration during shepherding and during Writers' Workshops. Also, a pattern author needs to be cognizant and be sensitive to pattern readers' concerns. The Social Web lends various opportunities for collaboration and sharing, which we consider next.

Collaborative Researching

The Web has become an indispensable source for researching for information, and the same holds for elicitation of domain knowledge (such as elicitation of pattern instances) during the PRP. There are Social Web applications like Google Notebook and Microsoft OneNote that allow one to attach notes to and clip text, graphics, and links during researching. These 'notebooks' can be saved, and can subsequently be used for collaboration and sharing with others. Furthermore, the 'notebooks' in Google Notebook can be exported to Google Docs.

Social Scheduling

A face-to-face meeting, whether it is for sociological approach to elicitation of domain knowledge or for Writers' Workshop, requires scheduling. A schedule that is agreeable to all, particularly as the number of persons involved increases, can become difficult to manage. The use of Social Web applications that facilitate calendar sharing (such as the Google Calendar) can reduce some of the tedium involved in scheduling a meeting agenda.

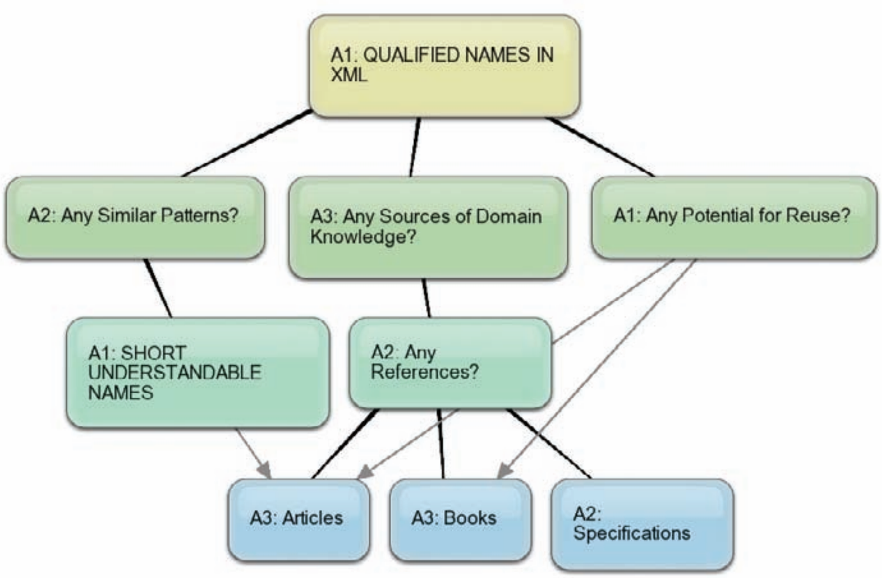
Brainstorming

Brainstorming is implicit to various activities in the PRP. For example, in a sociological approach to elicitation of domain knowledge, the authors often engage in brainstorming (for collectively organizing their thoughts and recall, for collaborative decision making, and so on). The same idea applies to the dynamics of shepherding.

One way to brainstorm is through visualization, and mind mapping is a graphically-oriented approach to realize it. A mind map is a graph where nodes represent sub-ideas (or sub-goals or sub-tasks) at different levels of granularity and vertices represent 'semantic' relationships, all of which are arranged radially around a central idea (or goal or task, respectively). The pattern authors can share these mind maps over the Web and, depending on the permissions, read and/or edit others' maps.

Figure 3 illustrates a snapshot in time (work in progress) of a mind map using the bubbl.us tool

Figure 3. An example of a partial mind map reflecting a brainstorming session on the viability of the 'new' QUALIFIED NAMES IN XML pattern



(http://www.bubbl.us/). In it, three authors, namely A1, A2, and A3 are in a brainstorming session on the viability of a proposed pattern. The 'bubbles' reflect respective inputs by pattern authors.

Collaborative Authoring

The Social Web presents a suitable environment for collaborative authoring of patterns using various means including Google Docs and Wiki. The concept of Wiki (Leuf & Cunningham, 2001) was invented in the mid-1990s as a group communication utility. It allowed open editing of information (like patterns) as well as the organization of the contributions and, with various enhancements, continues to serve well in that vein (Weiss & Birukou, 2007). A properly administered Wiki assists pattern authors, pattern shepherds, and pattern readers. Indeed, barring certain reservations, a Wiki environment enables a person to play the dual role of a pattern reader and a pattern administrator.

There are several, opens source flavors of Wiki available today addressing different target groups and organizational needs. Most flavors of Wiki, including MediaWiki and TinyWiki, can be easily acquired, installed, and administered under commonly-deployed computing platforms (Ebersbach, Glaser, & Heigl, 2006). For example, Asynchronous JavaScript and XML (AJAX) Patterns (http://ajaxpatterns.org/) and Perl Design Patterns (http://perldesignpatterns.com/) are collections of patterns based on MediaWiki and TinyWiki, respectively.

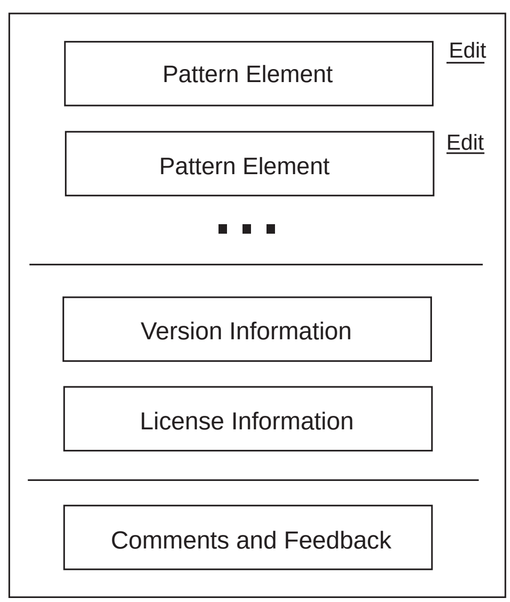
Figure 4 presents a simplified view of the description of a pattern within the Wiki environment. The mandatory elements of a pattern can be presented where they can be progressively disclosed (Lieberman, 2007) and edited; the details of history of the document can be automatically highlighted through version information; licensing terms can be made explicit; and pattern reader's feedback can be solicited and included. The figure could, for example, be extended by the addition of a block for the table of contents (in case the document is lengthy).

It is not absolutely necessary (although it may be relatively easier) to deploy Social Web technologies/applications for realizing collaboration. The conventional technologies/applications that have found success on the Web such as the Personal Hypertext Preprocessor (PHP) can be used to achieve similar effect as a Wiki. For example, a collection of patterns for 'living spaces' (http://architypes.net/patterns.php) is built from collaboration between architects, interior designers, and photographers.

Syndication

Every so often pattern consumers need to keep track of several resources on patterns, including announcements of new patterns or modifications of the existing ones, availability of relevant books or podcasts, and so on. However, individually and

Figure 4. A glimpse into an abstract, partial view of the description of a pattern within the Wiki environment



somewhat arbitrarily visiting each Web Site of interest can be inconvenient and time consuming. The subscription to periodically refreshable news feeds helps ameliorate this issue.

Syndication is a type of metadata implemented in form of channels that point to relevant resources on a topic that the pattern readers can subscribe to. There are a variety of syndication technologies of which Really Simply Syndication (RSS) and Atom are beginning to find broad support in conventional user agents and news feed readers.

For example, the following RSS markup fragment represents news for a specific day from a single channel:

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0">
  <channel>
     <title>Pedagogical Patterns Channel</
     title>
    <link>http://www.pattern.ca/</link>
    <description>
     This is a channel for news on patterns
     related to teaching and learning
       in the classroom.
    </description>
    <item>
        <title>News for January 15, 2008</
     title>
   <link>http://www.pattern.ca/2008/01/15/</
     link>
       <description>
        The interview of the author of the
     recently published book titled
        Patterns in the Classroom is avail-
     able as a podcast ...
       </description>
    </item>
  </channel>
</rss>
```

It could, for instance, be stored in a file named `PedagogicalPatterns.rss` and linked from a place that pattern readers could readily discover.

Social Bookmarking

Bookmarking has traditionally been one of the most common ways of remembering the resources of interest visited while browsing on the Web. However, these bookmarks reside on the user's computer and are not accessible by other devices (and therefore are not shareable).

Social bookmarking enables management (for example, storage, organization, search, and sharing) of booksmarks residing remotely at third-party services. By unifying their knowledge base, social bookmarking can help pattern authors and pattern shepherds communicate more effectively during the PRP. By expending modest effort, it can also help pattern consumers share links to resources (including desirable patterns) amid themselves. The notion of social bookmarking was pioneered by itlist.com in the mid-1990s and brought into mainstream around 2003 by del.icio.us. Since then other social bookmarking services like Google Bookmarks have spawned.

Organizing and Social Networking

It was claimed more than a decade ago that "finding patterns is much easier than describing them" (Gamma et al., 1995). On the Web, one could locate a pattern one of the following three ways: (1) directly using the (known) address where a pattern resides, (2) by navigating to the address, or (3) by searching through a collection. However, the rapid growth (Henninger & Corrêa, 2007) of the number of patterns and pattern languages has made the task of locating desirable patterns increasingly challenging for the pattern readers. Still, by participating in the Social Web, the pattern consumers can help one another (and indirectly help the pattern producers) in somewhat easing the task of locating desirable patterns, and folksonomy is one way to do that.

Folksonomy

A suitable organization of patterns is critical for locating desirable patterns. However, the search for a suitable organizing scheme for locating desirable patterns continues to remain a persistent and elusive problem (Hafiz, Adamczyk, & Johnson, 2007).

A crucial aspect of organization is classification. There is no universal scheme for classifying patterns: a pattern placed in one category by its author(s) can reappear as belonging to a different category in a different pattern language by another set of author(s). For example, the MODEL-VIEW-CONTROLLER (MVC) pattern (Buschmann, Henney, & Schmidt, 2007a) can be classified in multiple different ways, including categories that are not envisioned by its original author(s) but are considered relevant by its pattern readers and pattern users. This 'post-publication' faceted classification of a pattern is possible by social annotation, specifically via the notion of folksonomy or social tagging (Smith, 2008).

Folksonomy enables pattern readers to associate with a resource words or phrases that they deem meaningful, relevant, and significant in describing the resource. By doing so, there is an implicit assumption that other (new) pattern readers will share and benefit from this understanding of the resource.

Folksonomy can be realized in several different ways. For instance, the semantics of the XHTML documents can be extended using certain mechanisms within the div, span, and class elements, and id and rel attributes. Microformats (Allsopp, 2007) are an effort to standardize the conventions for using these extension mechanisms. In XHTML, the attribute-value pair rel="tag" can be used to indicate that the resource that has been linked-to acts as a tag for the current context. For example, to tag an XHTML document describing the MVC pattern with 'Distributed Computing,' markup such as Distributed Computing could be used.

A collection of tags can lead to the formation of a tag cloud. (There are some resemblances between a tag cloud and the classical Web concepts of image map and site map.) A tag cloud is set of related tags with associated weights that represent frequency of use of each tag. The tags within a tag cloud are usually ordered lexicographically and the frequency of use of each tag is illustrated by visual cues such as distinct font color and size. It is preferable to use a style sheet language such as the Cascading Style Sheets (CSS) for associating presentation semantics with tags.

The 'human element' of the Social Web—as personified by mutual collaboration among the actors in locating desirable patterns through navigation—can be realized in the following manner: by proper organization of tags and representation of weights in a tag cloud, pattern administrators and pattern engineers can help the pattern readers, and by a careful selection of tags, pattern readers can help each other.

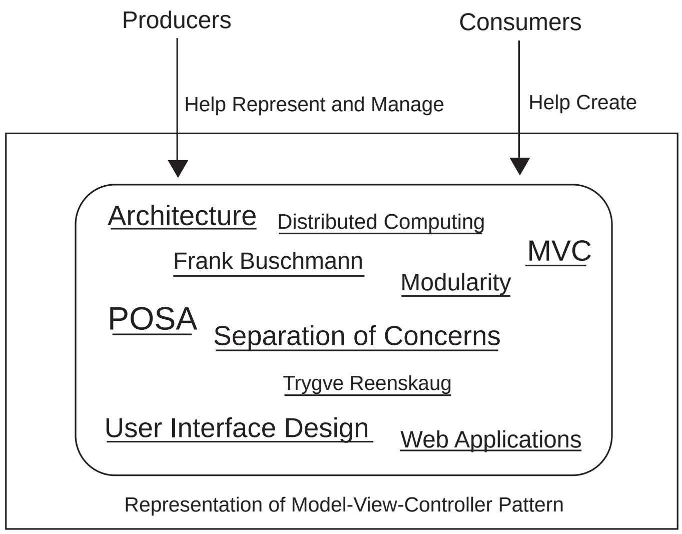
Figure 5 shows a tag cloud for the MVC pattern. It can be noted that the tags are not all of the same type: while some are about things, others are about people. For example, Modularity is a software engineering principle; POSA is the acronym for Patterns for Software Architecture; and Trygve Reenskaug is the person who is ascribed for first introducing MVC, and the tag could, for instance, link to a resource that acknowledges this.

The above idea can be extended to tagging a pattern language map that is dynamically-generated and expressed in a vector graphical language such as the Scalable Vector Graphics (SVG).

Social Networking

There are plenty of opportunities for social networking (Freeman, 2004) during and after the PRP. A variety of different types of social relationships can exist among the actors including pattern author-to-pattern author, pattern author-

Figure 5. A tag cloud embedded in the abstract representation of the MODEL-VIEW-CONTROLLER pattern



to-pattern shepherd, and pattern reader-to-pattern reader that can be made explicit.

The XHTML Friends Network (XFN) is a specification for explicitly indicating social networking relationships using `rel` attribute values in XHTML on blogroll links. For example, XFN values can be added to the `rel` attribute by a pattern author to indicate that John Smith is the shepherd, is a colleague, and is someone the author has met, using the following markup:

```
<ul>
  <li>
    <a href="http://john.smith.ca/"
        rel="colleague met shepherd">John
     Smith</a>
  </li>
</ul>
```

The social networking relationships expressed in XFN could be exploited by programs like bloggers, search engines, or spiders.

Publishing

The authors of patterns and pattern languages publish their work at the end of the process. In doing so, there are several issues involved. The representation and presentation of a published pattern can often be heterogeneous in nature, involving the use of different modes of information (text, graphics, source code, and so on). For instance, the solution of a pattern, particularly that is structurally-oriented, is usually accompanied with a picture illustrating the solution in an abstract manner or of its instances (examples). Any publishing also needs to take into account that some information can be reused and repeated at multiple places, and it needs to evolve independently for an effective maintenance. For example, the picture in a solution may also be a part of a pattern thumbnail. It is also possible that the author may not have complete technical control or legal rights to physically include some of the desirable information but nevertheless still needs to point to it. From a Social Web viewpoint, these considerations, if appropriately carried out, are suitable for a mashup.

Mashups

A mashup aggregates and reuses information from multiple sources. This is accomplished via some means such as a transclusion (Nelson, 1982), which is an on-demand inclusion of one resource into another via hyperlinking and/or a programmatic mechanism like an Enterprise Mashup Service (EMS), which is a Web Service based on the Service-Oriented Architecture (SOA). It then presents information in a federated manner that is transparent to a pattern reader.

Figure 6 illustrates an abstract construction of a pattern mashup in which author information, picture of an example, and licensing terms are 'transcluded' from external sources.

Both pattern authors and pattern readers (using, say, iGoogle) can benefit from mashups. For example, the description of a user interface pattern could reside on one server and, after receiving appropriate permissions, 'transclude' a picture depicting a solution instance from another server (like that of Flickr). As another example, the author can indicate that the pattern is licensed under the Creative Commons 3.0 Attribution Required License by including the markup `<a href="http://creativecommons.org/licenses/by/3.0/" rel="license">CC by 3.0</a>` in an XHTML document.

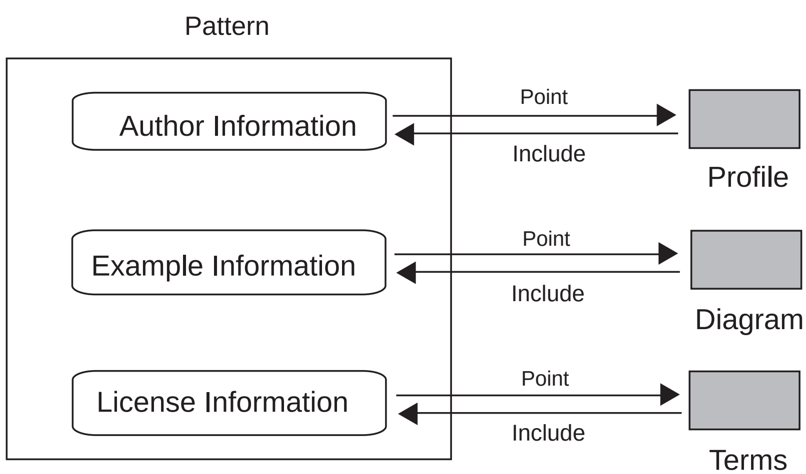
Deliberating, Educating, and Outreaching

It is the responsibility of the pattern authors to make a considerate effort to reach their audience of pattern consumers and, in doing so, use any means available to them. Being able to publish is only one aspect of communicating the elements of PE: it is also indirect, passive, and static, where the human side of patterns is less apparent. The Social Web presents an array of possibilities for direct, active, and dynamic means for communicating the elements of PE through a judicious use of media outlets. In particular, blogging and podcasting are asynchronous and synchronous means of communicating, respectively.

Blogging

A blog is an outlet for anybody to publicly express opinion, in writing, on just about any subject. In

Figure 6. A pattern mashup that points to and includes three different external sources



side bar of a personal blog, a blogger can provide a list of other blogs (or a blogroll).

There are a few benefits of blogging for the actors of patterns. Blogging gives pattern authors to respond to concerns of pattern consumers in an informal environment, inform others of their scholarly activities related to patterns (like relevant presentations at events), or state their current position about the (mis)use of patterns. For example, for the past few years, Grady Booch has been blogging about issues related to the Handbook for Software Architecture (http://www.booch.com/architecture/), which is one of the largest collections of patterns on software architecture. Blogging also gives pattern readers an opportunity to ask questions, and make their agreements and disagreements known in a public forum. For example, the Software Patterns Blog (http://pattern.ijop.org/) has been used by Mohamed Fayad to express concerns regarding current obstacles to learning and selecting patterns.

Podcasting

Podcasting provides an avenue to the pattern community for education and outreach by means of audio and video. For example, YouTube (http://www.youtube.com/results?search_query=Patterns) provides a platform for authors to post their presentations from events or other forms of media supporting patterns for a general audience without having to provide the service themselves. The pattern readers can provide comments on the videos, rate the videos, and add the videos to their list of social bookmarks.

As another example, the Software Engineering Radio (http://www.se-radio.net/) offers podcasts for interviews on software engineering related topics in general (Rech, 2007) and software patterns in particular. These interviews provide an opportunity for the authors to *converse* about patterns: demonstrate the patterns process by answering questions related to their experience in the domain, including successes and failures; draw attention to the history of patterns they elicited; mention the reasoning behind the selection of a means of representation or choice of examples; and so on.

Table 1 presents a summary of the key activities in PE and supporting concepts/activities corresponding to the Social Web discussed in this section.

Concerns of Communicating the Elements of Pattern Engineering via the Social Web

The integration of patterns within the Social Web is not absolute and has its share of shortcomings. In this section, the scope and limitations of the Social Web for communicating the elements of PE is examined and in doing so specific examples are provided.

Cost to Actors

The issue of cost impacts all actors. The increase in the number of mature, open source implementations of the technologies underlying the Social Web has contributed to control of development costs and thereby to the reduction of entry barrier for aspiring pattern producers. There is still the non-trivial cost of time and effort in learning, especially due to the fact that there is currently no single provider of these implementations and that the interaction (including the user interface functionality) varies tremendously from one application to another.

However, for pattern consumers, the technological infrastructure of the Social Web remains largely exclusive: for optimal operation, it assumes high-speed Internet connection, state-of-the-art operating systems, and latest user agents with up-to-date capabilities on the client-side. This is not all free-of-cost for personal use. In general, ‘Rich Internet Applications’ of the Social Web

Table 1. A mapping of activities in PE and concepts/activities associated with the Social Web

Pattern Engineering Activity	Social Web Concept/Activity
• Eliciting Pattern Instances	• Collaborative Researching • Brainstorming
• Assessing the Viability of the Pattern	• Brainstorming
• Eliciting Domain Knowledge (Sociological Approach)	• Social Scheduling
• Locating Expert Body of Knowledge for Authoring a Pattern	• Folksonomy • Social Bookmarking • Syndication
• Selecting Expert Body of Knowledge for Authoring a Pattern	• Brainstorming
• Participating in Writers' Workshop	• Social Scheduling
• Locating the Means for Representing and Presenting Patterns	• Folksonomy • Social Bookmarking • Syndication
• Selecting the Means for Representing and Presenting Patterns	• Brainstorming
• Authoring a Pattern	• Collaborative Authoring
• Assuring the Quality of a Pattern	• Social Web Application
• Representing and Presenting a Pattern	• Mashup • Microformat
• Publishing a Pattern	• Social Web Application
• Shepherding	• Collaborating • Social Networking
• Evaluating the Quality of a Pattern	• Collaborating
• Locating a Pattern	• Folksonomy • Microformat • Social Bookmarking • Syndication
• Reading and Understanding a Pattern	• Folksonomy
• Selecting a Pattern	• Brainstorming
• Using a Pattern	• Blogging • Podcasting
• Reflecting on the Use of a Pattern	• Blogging • Feedback
• Mentoring a Pattern Author, Pattern Reader, or Pattern User	• Blogging • Podcasting • Responding to Feedback
• Maintaining a Pattern	• Social Web Application
• Interfacing a Pattern with Other Entities of Knowledge and Artifacts	• Social Web Application

are resource-intensive. For example, the mashups in which aggregation of information takes place on the client-side expect hardware and software capabilities that a pattern consumer may not have. As another example, the Software Engineering Radio podcasts are currently not streamed but are available only as download at file sizes that could be prohibitive to those on low bandwidth.

The issue of hardware and network cost is not perennial. If past statistical trends are any indicators, then it is likely to subside with time. The cost of hardware necessary for creating Social Web applications (like computers with fast processor speeds, and large storage space and memory; digital cameras; audio recorders; and so on) has been on decline in the past few years, and reflect a market trend that is expected to continue. The same applies to cost of Internet service. However, it also needs to be noted that the original vision of the Web is yet to be realized: a large segment of the world population does not (still) have access to the Internet in general (Press, 2004) and the Web in particular.

Decentralization and Distribution of Control

The technologies underlying the Social Web can be disruptive: they require change and, in some cases, radical departure from conventional approaches. For instance, the Social Web is not meant for pattern authors working individually in isolation.

Any form of multi-node interaction over a distributed network where a node could be either a human or a machine (like real-time collaboration among author and shepherd, resource sharing, or mashup) creates an indirection. Its success is based on the assumption that it works like a symphony: *all* the nodes involved are available, are in unison, interaction among them is transparent and timely, and so on. The past experience with the Web has shown that the relationship between the probability of success and the number of nodes is not linear (the former can decrease as the latter increases).

The 'transfer' of even some of the traditional control from server-side to client-side has its side-effects that need to be balanced. For example, the exposure of the PRP to the Social Web may lead to a perception by the pattern readers that a pattern is never 'complete' but in a 'perpetual beta' state, which contradicts the characteristic of 'timelessness' (Alexander, 1979) of a pattern. The transfer of control also faces a classical dilemma. For example, moderating feedback by retaining only complimentary messages from pattern consumers or time-delimiting the feedback could be perceived as bias on part of pattern producers, which is against the spirit of openness of the Social Web. In contrast, it is not automatic that all unfiltered feedback driven by 'citizen journalism' or 'user-generated content' contributes to enriching the description of a pattern, and therefore some degree of moderation is necessary.

In spite of several possible uses, the flexibility of blogs and Wikis comes with a price: they are known for 'noise' (including impertinent information), 'casual' writing (due to the presence of phonetic, 1337 style of writing, and frequent spelling and/or grammatical errors), and 'editing wars' (discussions that have morphed into endless debates that put personal interest before that of a pattern reader). These, however, can be attributed to human usage rather than to inherent limitations of the underlying technology. In any case, this impacts the quality of the description of a pattern, and is not favorable to either a pattern administrator ([–PE-CR-3]) or to a pattern reader ([–PE-CR-4]). A partial solution to this issue could be to (1) separation of the description of a pattern as provided by the pattern author from any annotations (which should be clearly labeled and managed as such), and (2) provision of multiple views of a pattern, including the option to suppress any annotations, to a pattern reader.

Folksonomy (as opposed to taxonomy) is an uncontrolled vocabulary, and the lack of terminological control can have linguistic implications due to synonymy, homonymy, and polysemy. In particular, classical issues associated with the natural language use of acronyms can surface. It is not automatic that all tags that are created by pattern consumers may be relevant to the context. For example, MVC has other known expansions like Marriott Vacation Club, the Missouri Valley Conference, and the Motor Vehicle Commission that are irrelevant to the notion of a pattern as discussed in this paper. Similarly, the tag 'Architecture' in the civil engineering sense is not relevant to the MVC pattern. Therefore, once again, to add long-term value, the tags associated with the description of a pattern need to be monitored and moderated.

Semiotic Quality of Representations of Patterns

The theory of semiotics is the field of study of signs in which the communication itself is viewed as interchange of signs. It is known that a pattern can be viewed as a second-order sign of a semiotic system (Buschmann, Henney, & Schmidt, 2007b). It is possible for a sign to have one or more representations.

From a semiotic viewpoint (Stamper, 1992), the quality of the representation of a pattern (and a member of the pattern space in general) can be viewed on six interrelated levels: physical, empirical, syntactic, semantic, pragmatic, and social. This paper focuses on the last two levels. Then, inspired by conventional quality modeling (Fenton & Pfleeger, 1997) and by [PE-CR-1] – [PE-CR-4], pragmatic and social levels can be decomposed further. The desirable pragmatic and social quality attributes of concern to a pattern producer include comprehensibility, legality, and maintainability, while the desirable pragmatic and social quality attributes of concern to a pattern consumer include accessibility, comprehensibility, credibility, performance, readability, reliability, and usability.

The quality of representations of a growing number of patterns and pattern languages on the Web is a concern. In recent years, various accessibility, performance, reliability, and usability issues with patterns and pattern languages made available on the Web have been reported (Dennis & Snow, 2006; Segerståhl & Jokela, 2006; Manolescu et al., 2007). The evaluations (Deng, Kemp, & Todd, 2005) of certain collections of patterns (Gaffar et al., 2003) lead to questions of the credibility of these collections and for which there are no trivial answers.

The technologies/applications underlying the Social Web do not by themselves contribute towards the improvement of many of the aforementioned quality attributes of representations of patterns. In fact, in some cases they can potentially lead to a detriment. For example, let use consider the case of accessibility and legality. The misuse of microformats by overriding the semantics of XHTML attributes or real-time applications based on AJAX that are exclusively visually-driven, mouse-input-only, can be unfavorable to accessibility (Cooper, 2007). The use of fixed fonts and certain colors can make the tags/tag clouds inaccessible to those with certain forms of visual disability. A similar argument holds for podcasts. XFN graphs can become unreadable as the number of nodes and vertices increase. Furthermore, this situation only gets exacerbated with the use of mobile devices to access the Social Web.

The rise of the Social Web has amplified the classical struggle between liberty and legality. The term 'free' in the notion of 'freely sharing' on the Social Web (analogous to the open source movement) stands for freedom. As the same time, the preservation of rights of the creators of digital work is an ongoing challenge in a distributed environment where laws (if any) can vary across jurisdictions, and patterns are no exception. The use of pattern mashups, similar to its predecessors like 'inclining images,' can be open to legal issues related to copyright infringement, irrespective of nature (intentional or inadvertent) of the motive. The issues related to absence of any clear terms of use, or lack of comprehension of misinterpretation of license by an average pattern reader, are yet to be satisfactorily addressed.

Therefore, if the aim is to reach broad, diverse, and global readership, both the pattern producers and the pattern prosumers need to exercise caution towards adopting new, unproven, techniques and technologies in representations of patterns. In particular, as the boundaries between actor classes fade, quality assurance will need to become a *shared* responsibility. In absence of a supervising authority and relying solely on an honor system, this can be admittedly difficult.

DIRECTIONS FOR FUTURE RESEARCH

It is still early to predict the outcome of the Social Web phenomenon in general and its impact on patterns in particular. The work presented in this paper can be extended in a few different directions that are briefly discussed next.

Social Network Analysis of the Pattern Community

The diversity and visibility of participants in public appearances such as events (conferences, meetings, and workshops) in different countries, postings on blogs, mailing lists, and newsgroups, and so on, indicates that the pattern community is thriving. A social network analysis (SNA) of the pattern community, as it continues to grow and morph into a dedicated social network, would be useful.

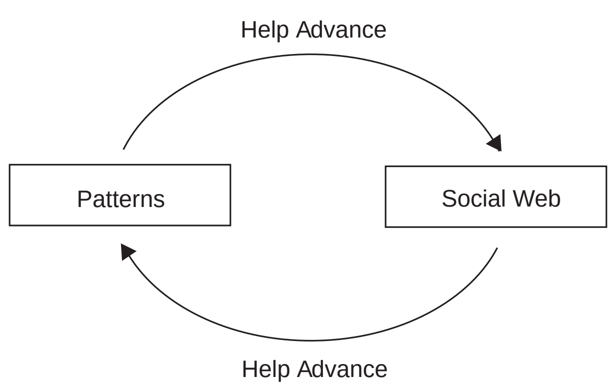
In particular, quantitative properties of the resulting graph such as centrality, closeness, clustering coefficient, cohesion, density, eigenvector centrality, and radiality, could be analyzed. This could help reveal certain relevant qualitative aspects of the network such as the relationships between actual actors; frequencies of use of specific patterns and pattern languages by certain actors; publications related to patterns and pattern languages recommended by people; demographical use of patterns; new domains of applicability of patterns; and so on.

Using Patterns for Social Web Applications

As shown in Figure 7, there is an apparent *symbiotic* relationship between patterns and the Social Web, one direction of which is explored by SW4PE in this paper.

The support for the other direction has been steadily increasing. The User Interface Design Patterns Library (http://ui-patterns.com/) provides, for instance, a TAG pattern and a TAG CLOUD pattern. There are patlets (Decker et al., 2006), patterns, and anti-patterns (Mader, 2008) available for making proper use of Wikis, which in turn could improve the collaboration between patterns authors and pattern readers. There are

Figure 7. The symbiotic relationship between patterns and the Social Web



design patterns for writing new microformats (Allsopp, 2007). There are also patterns available for computer-mediated communication in general (Schümmer & Lukosch, 2007) and for the design of Social Web applications in particular (O'Reilly, 2005) that can assist in the other direction. For example, APPLICATION SHARING, COLLABORATIVE SESSION, SHARED ANNOTATION, SHARED BROWSING, SHARED EDITING, and SHARED FILE REPOSITORY are patterns applicable to the Social Web context. It would be interesting to investigate the impact on the quality of collaboration and the quality of social (groupware) software that make use of these patterns. Finally, the design of Social Web applications typically corresponds to the PUBLISH-SUBSCRIBE architectural pattern.

The Convergence of the Social Web and the Semantic Web: Implications for Patterns

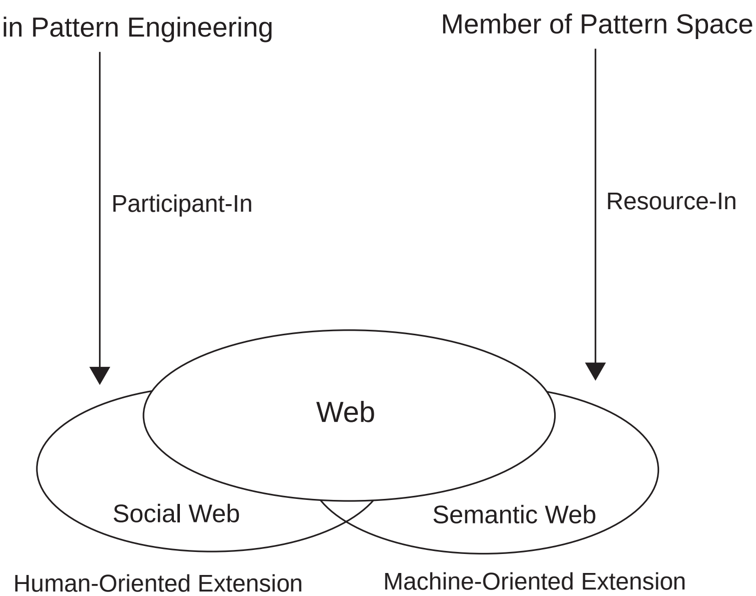
The Semantic Web has recently emerged as another perceived extension of the current Web that adds technological infrastructure for better knowledge representation, interpretation, and reasoning (Hendler, Lassila, & Berners-Lee, 2001). The Social Web efforts and the Semantic Web initiative are not competing but complementing, and need to co-exist. For the sustainability of the architecture of the Web, it is essential that the extensions of the Web evolve harmonically (Shadbolt, Hall, & Berners-Lee, 2006). For a unified view, the Social Web-specific efforts will need to take advantage of formalization (and thereby become more machine-oriented) and the Semantic Web-specific efforts will need to become more human-centric. This can be crucial for a future generation of PMS.

It would therefore be of interest to examine the synergies between the Semantic Web and the Social Web efforts, or as it is more commonly referred to by the pseudonym Web 3.0 (Lassila & Hendler, 2007), from the viewpoint of benefits and concerns to patterns. This, as the Figure 8 illustrates, is all the more significant since the actors in PE are participants in the Social Web and the members of the pattern space are resources in the Semantic Web.

Indeed, there are an increasing number of initiatives (Mika, 2007) that belong to the intersection of the Semantic Web and the Social Web, some of which may be relevant to PE. The Friend of a Friend (FOAF) and RELATIONSHIP are both Resource Description Framework (RDF) vocabularies: FOAF is used for expressing metadata about people, and their interests, relationships between them, the things they create, and activities they are involved in; RELATIONSHIP enriches FOAF by extending the types of relationships between people. As an example, the following is a combination of FOAF and RELATIONSHIP markup that represents the Myers-Briggs Type Indicator (MBTI) (Keirsey, 1998) of the shepherd John Smith and the relationship between John Smith and the author Steven Nash:

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-
     rdf-syntax-ns#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
 xmlns:rel="http://purl.org/vocab/relation-
     ship/">
  <foaf:Person>
    <foaf:name>John Smith</foaf:name>
     <foaf:mbox rdf:resource="mailto:john.
     smith@john.smith.ca"/>
           <foaf:myersBriggs>ESTJ</
     foaf:myersBriggs>
    <foaf:knows>
      <foaf:Person>
        <foaf:name>Steven Nash</foaf:name>
      </foaf:Person>
    </foaf:knows>
    <rel:mentorOf>
      <foaf:Person>
        <foaf:name>Steven Nash</foaf:name>
```

Figure 8. The actors in PE and the members of the pattern space belong to the human and machine extensions of the Web, respectively




```
      </foaf:Person>
    </rel:mentorOf>
  </foaf:Person>
</rdf:RDF>
```

The limitations of these vocabularies are evident as they do not represent precise relationships among actors of patterns. For example, the `mentorOf` element is only an approximation to a shepherd.

The notion of an ontology perhaps forms one of the most important layers in the Semantic Web architecture. A formal representation of a pattern language as an ontology in the Web Ontology Language (OWL) enables better opportunities for organization and inferencing than that is possible by conventional means (Kamthan & Pai, 2006b). For example, an ontological representation of a pattern language allows making the implicit relationships between patterns explicit, which can complement [3] of PRP. However, the focus in this study is on technical rather than on social aspects. The aforementioned means of social collaboration could be useful towards ontology engineering of patterns, which is conducted in a social context (Gruber, 2004) and requires a considerate human effort to reach a consensus. These ontologies could be enriched by more knowledge about the actors. For example, the Semantically-Interlinked Online Communities (SIOC) Project aims to use Semantic Web technologies to connect people using different (Social) Web applications (like e-mail, blog or Wiki). The SIOC Core Ontology Specification explicitly uses RDF, OWL, and FOAF.

The other possibilities for a confluence between the patterns, the Semantic Web, and the Social Web from a research viewpoint are: a pattern language map in SVG could be generated from an OWL ontology and annotated with pattern reader-oriented tags to make it amenable to the tools for the Social Web; the FOAF – XFN and FOAF – hCard microformat connections from the perspective of an ontology of pattern language could be further exploited; the places where an ontology falls short could be compensated by the creation of a microformat for representation of patterns in

XHTML; the implications of Semantic Wikis on communication of patterns from an ontological viewpoint could be examined; and so on.

Patterns on the Mobile Social Web

It would perhaps not be an overstatement to suggest that in the last few years the ability to access the Web via a wireless device has been remarkably successful (Stanoevska-Slabeva, 2003). The potential of mobile access to the Social Web has evidently inspired the notion of Mobile Social Web, or more commonly known by the pseudonym Mobile Web 2.0 (Jaokar, 2006; Golding, 2008).

Due to the inherent constraints of both hardware and software (Tarasewich, 2003), it is unlikely that the Mobile Social Web could ever become 'de facto' environment for communicating the elements of PE. However, it could still serve as a means to facilitate collaboration in the PRP, in particular that of patlets, pattern thumbnails, and small pattern language maps, and for syndication. Further exploration of the interplay between patterns and the Mobile Social Web, particularly from the viewpoint of an extension to Figure 7, would be of interest.

Extending the Scope of SW4PE

The arguments presented in this paper accentuating the prospects versus highlighting the concerns of communicating the elements of PE via the Social Web are not exclusive. They could apply to other members of the pattern space as well as to other similar situations. For instance, as the number of anti-patterns grows, assessing the viability of the Social Web in communicating anti-patterns would be worth investigating.

The world of patterns is not secluded from other entities of knowledge. Indeed, software design patterns are related to (in the sense that they are influenced by or influence), for example, aspects, software engineering principles, Application Programming Interfaces (APIs), and software frameworks (Garzas & Piattini, 2005). The Social Web has potential benefits for these other entities of knowledge that are related to patterns and, in doing, so would benefit patterns indirectly. The same applies to the relationships between software patterns and other reusable software artifacts such as those identified in the Zachman Framework for Enterprise Architecture (Zachman, 1987).

There are various knowledge areas in the Guide to the Software Engineering Body of Knowledge (SWEBOK) and the Software Engineering Education Knowledge (SEEK) that require human-to-human communication. Further studies on topics, such as, integrating Social Web technologies/applications in requirements engineering (Macaulay, 1993), collaborative conceptual modeling, or in a socio-constructivist approach to software engineering education in general and in a collaborative approach to open source course projects in particular (Kamthan, 2007), would also be of research interest.

CONCLUSION

The creation and transfer of knowledge that occurs in PE rests strongly on human-to-human communication. This human-orientation needs to be acknowledged explicitly, and the Social Web provides an open and global environment for doing so.

The Social Web opens new vistas for the actors of patterns and indeed for the PRP itself. It celebrates a critical feasibility issue facing the pattern producers—that the number of pattern producers is less than the number of pattern consumers, and due to practical limitations on resources (in terms of time and effort) to dedicate, can not realistically be expected to explicitly document every possible view of patterns—a resolution to which results from the 'collective intelligence' of the pattern consumers. This, however, is not free of cost.

It is likely that by appropriate use of technologies/applications, some of the concerns outlined in this paper can at least be avoided if not entirely eliminated. It should, however, be noted that although certain limitations of the Social Web as it pertains to communication of patterns are transient, others are more fundamental.

The potential benefits of the Social Web can outweigh the costs in the long-term if the expectations are realistic and if an effort is made to address the associated concerns. If the past experience with the use of the Internet and the Web is any indicator, the movement has traditionally been towards flexibility, usability, and universality. The success between the interplay of patterns and the Social Web is likely to depend on these invariant values.

In conclusion, the shared benefits of any socialization for the purpose of communication of patterns can only come to fruition with the *shared* sense of accountability and responsibility on part of all actors. This can come only with coordination between pattern producers and pattern consumers. Therefore, an optimistic but cautious use of the Social Web is an imperative.

ACKNOWLEDGMENT

The author would like to thank the reviewers for their feedback and suggestions for improvement.

REFERENCES

Alexander, C. (1979). The Timeless Way of Building. Oxford University Press.

Alexander, I. F. (2005). A Taxonomy of Stakeholders: Human Roles in System Development. International Journal of Technology and Human Interaction, 1(1), 23-59.

Alexander, C., Ishikawa, S., & Silverstein, M. (1977). A Pattern Language: Towns, Buildings, Construction. Oxford University Press.

Allsopp, J. (2007). Microformats: Empowering Your Markup for Web 2.0. Friends of Ed.

Appleton, B. A. (1997). Patterns and Software: Essential Concepts and Terminology. Object Magazine Online, 3(5), 20-25.

Ballard, B. (2007). Designing the Mobile User Experience. John Wiley and Sons.

Buschmann, F., Henney, K., & Schmidt, D. C. (2007a). Pattern-Oriented Software Architecture, Volume 4: A Pattern Language for Distributed Computing. John Wiley and Sons.

Buschmann, F., Henney, K., & Schmidt, D. C. (2007b). Pattern-Oriented Software Architecture, Volume 5: On Patterns and Pattern Languages. John Wiley and Sons.

Cooper, M. (2007). Accessibility of Emerging Rich Web Technologies: Web 2.0 and the Semantic Web. The 2007 International Cross-Disciplinary Workshop on Web Accessibility (W4A 2007), Banff, Canada, May 7-8, 2007.

Coplien, J. O. (1996). The Human Side of Patterns, C++ Report, 81-85.

Coplien, J. O. (2000). A Pattern Language for Writers' Workshops. In: Pattern Languages of Program Design 4. N. B. Harrison, B. Foote, & H. Rohnert (Eds.). Addison-Wesley.

Decker, B., Ras, E., Rech, J., Klein, B., & Hoecht, C. (2006). Using Wikis to Manage Use Cases: Experience and Outlook. Workshop on Learning Software Organizations and Requirements Engineering, Hannover, Germany, March 27-28, 2006.

Deng, J., Kemp, E., & Todd, E. G. (2005). Managing UI Pattern Collections. The Sixth ACM SIGCHI New Zealand Chapter's International Conference on Computer-Human Interaction: Making CHI Natural, Auckland, New Zealand, July 7-8, 2005.

Dennis, T., & Snow, K. (2006). Web Design Patterns Collection Technical Design. Center for Document Engineering Technical Report CDE2006-TR04. University of California, Berkeley, USA.

Ebersbach, A., Glaser, M., & Heigl, R. (2006). Wiki: Web Collaboration. Springer-Verlag.

Engelbart, D. C. (1995). Toward Augmenting the Human Intellect and Boosting our Collective IQ. Communications of the ACM, 38(8), 30-32.

Fenton, N. E., & Pfleeger, S. L. (1997). Software Metrics: A Rigorous & Practical Approach. International Thomson Computer Press.

Freeman, L. C. (2004). The Development of Social Network Analysis: A Study in the Sociology of Science. Empirical Press.

Gaffar, A., Sinnig, D., Javahery, H., & Seffah, A. (2003). MOUDIL: A Comprehensive Framework for Disseminating and Sharing HCI Patterns. CHI 2003 Workshop on Perspectives on HCI Patterns: Concepts and Tools, Fort Lauderdale, USA, April 6-7, 2003.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley.

Garzas, J., & Piattini, M. (2005). An Ontology for Microarchitectural Design Knowledge. IEEE Software, 22(2), 28-33.

Golding, P. (2008). Next Generation Wireless Applications: Creating Mobile Applications in a Web 2.0 and Mobile 2.0 World. John Wiley and Sons.

Gruber, T. (2004). Every Ontology is a Treaty. SIGSEMIS Bulletin, 1(3), 2004.

Hafiz, M., Adamczyk, P., & Johnson, R. E. (2007). Organizing Security Patterns. IEEE Software, 24(4), 52-60.

Harrison, N. B. (2000). The Language of Shepherding: A Pattern Language for Shepherds and Sheep. In: Pattern Languages of Program Design 4. N. B. Harrison, B. Foote, & H. Rohnert (Eds.). Addison-Wesley.

Harrison, N. B. (2003). Advanced Pattern Writing. The Eighth European Conference on Pattern Languages of Programs (EuroPLoP 2003), Irsee, Germany, June 25-29, 2003.

Hendler, J., Lassila, O., & Berners-Lee, T. (2001). The Semantic Web. Scientific American, 284(5), 34-43.

Henninger, S., & Corrêa, V. (2007). Software Pattern Communities: Current Practices and Challenges. The Fourteenth Conference on Pattern Languages of Programs (PLoP 2007), Monticello, USA, September 5-8, 2007.

Hess, C., & Ostrom, E. (2007). Understanding Knowledge as a Commons: From Theory to Practice. MIT Press.

Jaokar, A. (2006). Mobile Web 2.0: The Innovator's Guide to Developing and Marketing Next Generation Wireless/Mobile Applications. Futuretext.

Kamthan, P. (2007). On the Prospects and Concerns of Integrating Open Source Software Environment in Software Engineering Education. Journal of Information Technology Education, 6, 45-64.

Kamthan, P. (2008). A Situational Methodology for Addressing the Pragmatic Quality of Web Applications by Integration of Patterns. Journal of Web Engineering, 7(1), 70-92.

Kamthan, P. (2009). Pattern-Oriented Use Case Modeling. In: Encyclopedia of Information Science and Technology (Second Edition). M. Khosrow-Pour (Ed.). IGI Global.

Kamthan, P., & Pai, H.-I. (2006a). Knowledge Representation in Pattern Management. In: Encyclopedia of Knowledge Management. D. Schwartz (Ed.). Idea Group.

Kamthan, P., & Pai, H.-I. (2006b). Representation of Web Application Patterns in OWL. In: D. Taniar, & J. W. Rahayu (Eds.). Web Semantics and Ontology. Idea Group.

Kamthan, P., & Pai, H.-I. (2008). Using Patterns for Engineering High-Quality E-Commerce Applications. In: Business Web Strategy: Design, Alignment, and Application. M. Memmola & L. Al-Hakim (Eds). IGI Global.

Keirsey, D. (1998). Please Understand Me II. Prometheus Nemesis Book Company.

Lassila, O., & Hendler, J. (2007). Embracing "Web 3.0". IEEE Internet Computing, 11(3), 90-93.

Lea, D. (1994). Christopher Alexander: An Introduction for Object-Oriented Designers. ACM SIGSOFT Software Engineering Notes, 19(1), 39-46.

Leuf, B., & Cunningham, W. (2001). The Wiki Way: Quick Collaboration on the Web. Addison-Wesley.

Lieberman, B. A. (2007). The Art of Software Modeling. Auerbach Publications.

Macaulay, L. (1993). Requirements Capture as a Cooperative Activity. The First IEEE International Symposium on Requirements Engineering, San Diego, USA, January 4-6, 1993.

Mader, S. (2008). Wikipatterns: A Practical Guide to Improving Productivity and Collaboration in Your Organization. John Wiley and Sons.

Manolescu, D., Kozaczynski, W., Miller, A., & Hogg, J. (2007). The Growing Divide in the Patterns World. IEEE Software, 24(4), 61-67.

Meszaros, G., & Doble, J. (1998). A Pattern Language for Pattern Writing. In: Pattern Languages of Program Design 3. R. C. Martin, D. Riehle, & F. Buschmann (Eds.). Addison-Wesley, 529-574.

Mika, P. (2007). Social Networks and the Semantic Web. Springer-Verlag.

Nelson, T. H. (1982). Literary Machines. Mindful Press.

Nonaka, I., & Takeuchi, H. (1995). The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press.

O'Reilly, T. (2005). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. O'Reilly Network, September 30, 2005.

Polanyi, M. (1983). The Tacit Dimension. Peter Smith.

Press, L. (2004). The Internet in Developing Nations: Grand Challenges. First Monday, 9(4-5).

Rech, J. (2007). Podcasts about Software Engineering. ACM SIGSOFT Software Engineering Notes, 32(2).

Rising, L. (2000). The Pattern Almanac 2000. Addison-Wesley.

Schumacher, M. (2003). Security Engineering with Patterns: Origins, Theoretical Models, and New Applications. Springer-Verlag.

Schumacher, M., Fernandez-Buglioni, E., Hybertson, D., Buschmann, F., & Sommerlad, P. (2006). Security Patterns: Integrating Security and Systems Engineering. John Wiley and Sons.

Schümmer, T., & Lukosch, S. (2007). Patterns for Computer-Mediated Interaction. John Wiley and Sons.

Segerståhl, K., & Jokela, T. (2006). Usability of Interaction Patterns. CHI 2006 Conference on Human Factors in Computing Systems, Montréal, Canada, April 22-27, 2006.

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The Semantic Web Revisited. IEEE Intelligent Systems, 21(3), 96-101.

Shuen, A. (2008). Web 2.0: A Strategy Guide. O'Reilly Media.

Stamper, R. (1992). Signs, Organizations, Norms and Information Systems. The Third Australian Conference on Information Systems, Wollongong, Australia, October 5-8, 1992.

Smith, G. (2008). Tagging: People-Powered Metadata for the Social Web. New Riders.

Stanoevska-Slabeva, K. (2003). Towards a Reference Model for M-Commerce Applications. The Eleventh European Conference on Information Systems (ECIS 2003), Naples, Italy, June 16-21, 2003.

Tarasewich, P. (2003). Designing Mobile Commerce Applications. Communications of the ACM, 46(12), 57-60.

Weiss, M., & Birukou, A. (2007). Building a Pattern Repository: Benefiting from the Open, Lightweight, and Participative Nature of Wikis. Wikis for Software Engineering Workshop (Wikis4SE 2007), Montreal, Canada, October 21, 2007.

Wesson, J., & Cowley, L. (2003). Designing with Patterns: Possibilities and Pitfalls. The Second Workshop on Software and Usability Cross-Pollination, Zürich, Switzerland, September 1-2, 2003.

White, B. (2007). The Implications of Web 2.0 on Web Information Systems. In: Web Information Systems and Technologies. J. Filipe, J. Cordeiro, & V. Pedrosa (Eds.). Springer-Verlag, 3-7.

Winters, N., & Mor, Y. (2008). IDR: A Participatory Methodology for Interdisciplinary Design in Technology Enhanced Learning. Computers & Education, 50(2) 579-600.

Zachman, J. A. (1987) A Framework for Information Systems Architecture, IBM Systems Journal, 26(3), 276-292.

This work was previously published in the International Journal of Technology and Human Interaction, Vol. 5, Issue 2, edited by B. C. Stahl, pp. 36-62, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Section VIII
Emerging Trends

This section highlights research potential within the field of Web technologies while exploring uncharted areas of study for the advancement of the discipline. Chapters within this section highlight emerging semantic Web applications, Web personalization, and learning on the Web. These contributions, which conclude this exhaustive, multi-volume set, provide emerging trends and suggestions for future research within this rapidly expanding discipline.

Chapter 8.1
The Social Semantic Desktop:
A New Paradigm Towards Deploying the Semantic Web on the Desktop

Ansgar Bernardi
German Research Center for Artificial Intelligence (DFKI) GmbH, Kaiserslautern, Germany

Stefan Decker
National University of Ireland, Ireland

Ludger van Elst
German Research Center for Artificial Intelligence (DFKI) GmbH, Kaiserslautern, Germany

Gunnar Aastrand Grimnes
German Research Center for Artificial Intelligence (DFKI) GmbH, Kaiserslautern, Germany

Tudor Groza
National University of Ireland, Ireland

Siegfried Handschuh
National University of Ireland, Ireland

Mehdi Jazayeri
University of Lugano, Switzerland

Cédric Mesnage
University of Lugano, Switzerland

Knud Möller
National University of Ireland, Ireland

Gerald Reif
University of Lugano, Switzerland

Michael Sintek
German Research Center for Artificial Intelligence (DFKI) GmbH, Kaiserslautern, Germany

Leo Sauermann
German Research Center for Artificial Intelligence (DFKI) GmbH, Germany

ABSTRACT

This chapter introduces the general vision of the Social Semantic Desktop (SSD) and details it in the context of the NEPOMUK project. It outlines the typical SSD requirements and functionalities that were identified from real world scenarios. In addition, it provides the design of the standard SSD architecture together with the ontology pyramid developed to support it. Finally, the chapter gives an overview of some of the technical challenges that arise from the actual development process of the SSD.

DOI: 10.4018/978-1-60566-112-4.ch012

INTRODUCTION

A large share of everybody's daily activities centres around the handling of information in one way or the other. Looking for information, digesting it, writing down new ideas, and sharing the results with other people are key activities both in work as well as in manifold leisure activities. The abundance of PCs and the Web in today's world result in new numbers and qualities of information exchange and interaction which are seen both as chance and as threat by the users. Supporting personal and shared information handling is thus a highly requested but yet unsolved challenge.

In traditional desktop architectures, applications are isolated islands of data – each application has its own data, unaware of related and relevant data in other applications. Individual vendors may decide to allow their applications to interoperate, so that, e.g., the email client knows about the address book. However, today there is no consistent approach for allowing interoperation and a system-wide exchange of data between applications. In a similar way, the desktops of different users are also isolated islands – there is no standardized architecture for interoperation and data exchange between desktops. Users may exchange data by sending emails or uploading it to a server, but so far there is no way of seamless communication from an application used by one person on their desktop to an application used by another person on another desktop.

The problem on the desktop is similar to that on the Web – also there, we are faced with isolated islands of data and no generic way to integrate and communicate between various Web applications (i.e., Web Services). The vision of the SW offers solutions for both problems. RDF[a] is the common data format which builds bridges between the islands, and Semantic Web Service technology offers the means to integrate applications on the Web.

The Social Semantic Desktop (SSD) paradigm adopts the ideas of the SW paradigm for the desktop. Formal ontologies capture both a shared conceptualization of desktop data and personal mental models. RDF serves as a common data representation format. Web Services – applications on the Web – can describe their capabilities and interfaces in a standardized way and thus become Semantic Web Services. On the desktop, applications (or rather: their interfaces) will therefore be modelled in a similar fashion. Together, these technologies provide a means to build the semantic bridges necessary for data exchange and application integration. The Social Semantic Desktop will transform the conventional desktop into a seamless, networked working environment, by loosening the borders between individual applications and the physical workspace of different users.

By realizing the Social Semantic Desktop, we contribute to several facets of an effective personal information handling:

- We offer the individual user a systematic way to structure information elements within the personal desktop. Using standard technology to describe and store structures and relations, users may easily reflect and express whatever is important in their personal realm.
- Standardized interfaces enable the integration of all kinds of available desktop applications into the personal information network. Investments in programs, data collections, and hard-learned working styles are not lost but augmented and connected into a comprehensive information space.
- Based on the SW technology basis, all kinds of automated and semi-automatic support are possible, like, e.g., text classification services, image categorization, document relevance assessments, etc.
- The exchange of standard data formats between individual work spaces is supported not only on the technical level (e.g.,

standard communication protocols), but also on the semantic level (via sharing and alignment of ontologies and the corresponding annotated information elements). The integration with formal ontologies eases the sharing and understanding between different persons.

- Ultimately, we thus contribute to a solution for the initialization problem of the SW: As the individual user will receive immediate benefit from the semantic annotation within the personal workspace, the motivation is high to invest the necessary structuring and formalization work. As the standards used allow for an effortless sharing of such work, the amount of semantically annotated information which can be made available in the Web grows dramatically – which in turn makes it worthwhile to develop new SW-based services.

In this chapter we describe in detail the core components which are necessary for building a Social Semantic Desktop. We illustrate the necessary standard framework and describe the role and structure of the ontologies which support the spectrum from personal to social information handling. We outline the implementation decisions which need to be observed in order to realize a consequently ontology-oriented system, which is able to deal with the numerous flexibilities required within the Semantic Web. Finally, we show examples of the benefits obtained from the realization and use of an SSD.

The ideas and implementation principles presented in this chapter are distilled from our experiences in the NEPOMUK Project[b]. For each section we will describe the general motivation and principles and then give details on how the particular challenges have been solved in the NEPOMUK project.

BACKGROUND

The Social Semantic Desktop vision has been around for a long time: visionaries like Vannevar Bush (1945) and Doug Engelbart (1962) have formulated and partially realized these ideas. However, for the largest part their ideas remained a vision for far too long, since the necessary foundational technologies were not yet invented – figuratively speaking, these ideas were proposing jet planes when the rest of the world had just invented the parts to build a bicycle. Only in the recent years several technologies and research streams began to provide a foundation which will be combined and extended to realize the envisioned collaborative infrastructure of the SSD.

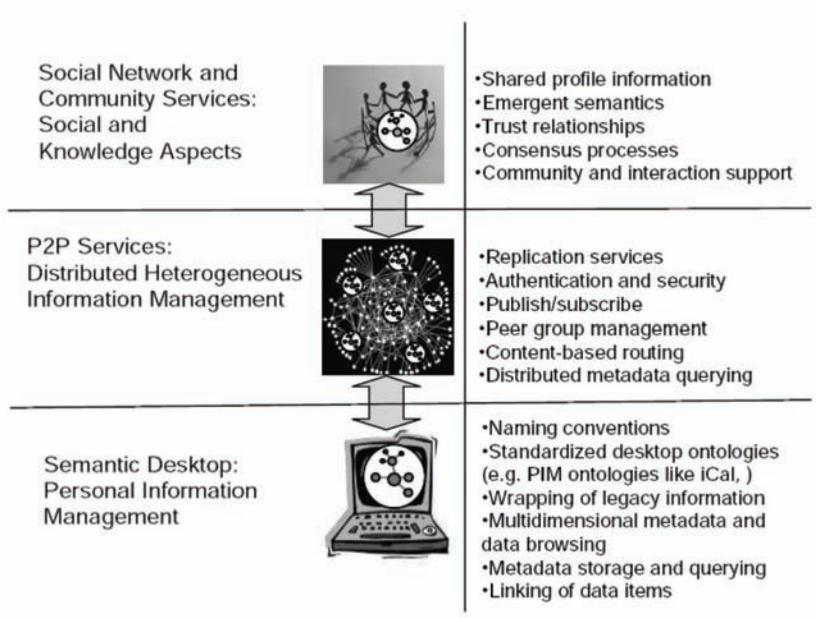
Figure 1 shows the highest-level architecture and connections between components of the SSD, i.e., the social networks, the P2P infrastructure, and the individual desktops. Traditional semantics, knowledge representation, and reasoning research are now interacting. While none of them can solve the problem alone, together they may have the explosive impact of the original Web:

The Semantic Web effort provides standards and technologies for the definition and exchange of metadata and ontologies. Available standard proposals provide ways to define the syntax (RDF) and semantics of metadata based on ontologies (Web Ontology Language – OWL (McGuiness et. al, 2004), RDF Schema – RDFS). Research covering data transfer, privacy, and security issues is now also under development.

Social Software maps the social connections between different people into the technical infrastructure. As an example, Online Social Networking makes the relationships between individuals explicit and allows the discovery of previously unknown relationships. The most recent Social Networking Sites also help form new virtual communities around topics of interest and provide means to change and evolve these communities.

P2P and Grid computing develops technol-

Figure 1. Component architecture of the Social Semantic desktop



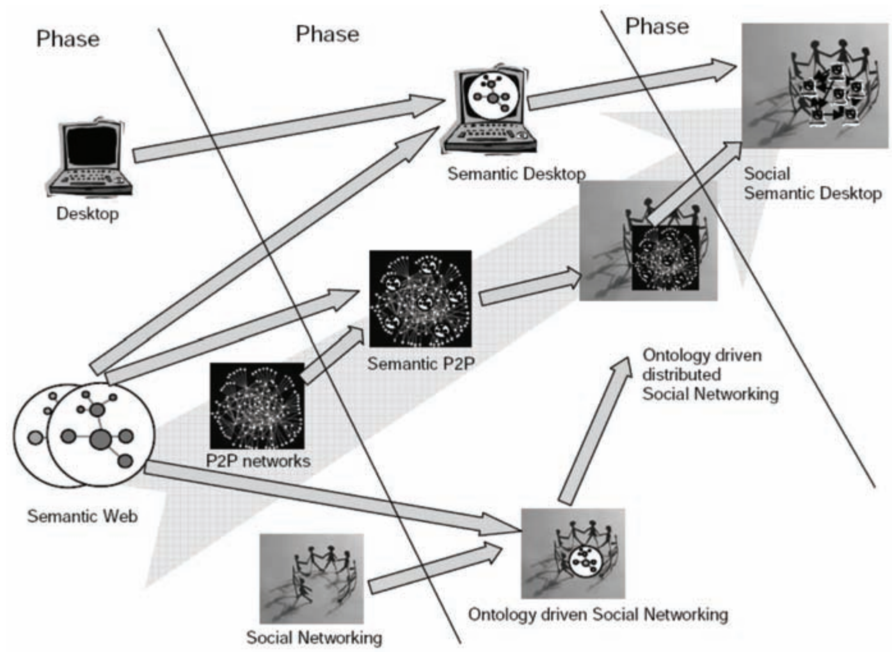
ogy to network large communities without centralized infrastructures for data and computation sharing. P2P networks have technical benefits in terms of scalability and fault tolerance, but a main advantage compared to central sites is a political one: they allow to build communities without centralized nodes of control, much as the Internet grew as fast as it did because it was based on reciprocity – it avoided political debate as to who gets to own big, expensive central facilities. Recent research has provided initial ways of querying, exchanging and replicating data in P2P networks in a scalable way.

By projecting the trajectory of current trends, we can simplify this picture by stating that next generation desktop applications will support collaboration and information exchange in a P2P network, connecting online decentralized social networks, and enabling shared metadata creation and evolution by a consensus process. The result of this process will be the Social Semantic Desktop. Figure 2 depicts the phases in which the relevant co-evolving technologies are combined to achieve the final goal, i.e., the realization of the Social Semantic Desktop.

SCENARIOS

Before we move on to the specific functionalities that a Social Semantic Desktop supports and discuss how they are implemented, we will first present some scenarios that will illustrate what an SSD is, how it will be used, and how it will change the way we do knowledge work. We chose the scenarios such that they illustrate the different dimensions of an SSD: Sect. *The Semantic Dimension* describes example usage that shows the use of semantics on the desktop, and Sect. *The Social Dimension* will show the social dimension of an SSD, i.e., the interaction between desktops of different users. The scenarios give an overview of what is possible and how the SSD presents itself to the user.

Figure 2. Phases towards the Social Semantic desktop



The Semantic Dimension

A typical use of a single Semantic Desktop is to organize ones data: files, emails, pictures, etc. Users are able to tag those information resources with concepts from a network of ontologies. The ontologies also contain relations (or properties, to use RDF terminology), which can be used to link information resources on the desktop. Organizing information resources in this way helps users to find what they are looking for quicker, and makes it possible for the Semantic Desktop itself to aid the user in their daily work. When a user first begins using the Semantic Desktop, many often-used concepts and properties are already present. E.g., there are basic concepts such as **Person**, **Meeting** or **Place**, and properties such as **knows** or **located-in**. Also, we can assume that useful things like an ontology of all countries are already in place. Then, as the need arises, users can extend the existing ontologies – e.g., they can add concepts for a particular meeting they attend or people they know, such as **Executive-Committee-Meeting-07/06/07**, **Jane-Doe** or **Hans-Schmidt**. The following two scenarios give examples of this kind of Semantic Desktop usage. We will use two imaginary users (*personas*[c]) flesh them out: Dirk, who works in a research group for some company in Germany, and Claudia, who is the group leader and his boss. Both Dirk and Claudia work on a project called Torque.

Organizing pictures (*Annotation*). Dirk just got back from his holidays in Norway, where he took a lot of pictures. Using his Semantic Desktop, he now wants to organize them, so that he can later find the pictures he wants to easier, generate photo albums for particular topics, etc. A lot of the important concepts probably already exist on his desktop, such as Norway, or the cities he has visited: **Oslo**, **Bergen** and **Trondheim**. Other concepts will be added by Dirk himself, such as **Holidays-in-Norway-2007** and tourist sights like **Preikestolen** or **Holmenkollen**. Since these concepts are more than just tags, Dirk can also say things *about* them, e.g., that **Holidays-in-Norway-2007** was a **Trip** and took place in **2007**, and that **Preikestolen** is a **Location** in **Norway**. Dirk even managed to take a picture of Prince Håkon and Princess Mette-Marit, so he creates two more concepts **Håkon** and **Mette-Marit**. There are many ways in which Dirk can link (or tag) his pictures to the relevant concepts – however, part of the Semantic Desktop are intuitive user interfaces, which hide most of the intricacies that go on under the hood from the user. E.g., Dirk might have an application that shows all the concepts that he is interested in the form of a tag cloud. Linking the pictures would then simply require him to drag them onto the desired concept in the cloud.

Planning a trip (*Context*). Later, Dirk finds out that he has to go on a work trip: a conference in Oslo. The Semantic Desktop assists him in planning and organizing this trip, through the notion of *context*. Dirk can create a new **Trip** object **Trip-to-OOC2007-Oslo** and tell his desktop that he is now in the context of this trip. This means that everything he does from this moment on will be interpreted as happening in that context, until he quits the context again. When he books a flight in his Web browser, the destination field will automatically be filled in with "Oslo", similarly the departure field. Afterwards, when he books a hotel room, he will be assisted similarly. Dirk will receive a number of email confirmations, such as the flight itinerary and booking confirmation for his hotel. These emails and their attachments will automatically be filed as belonging to the **Trip-to-OOC2007-Oslo** context, so that Dirk can easily find them again later. Once he knows his exact flight dates and where his hotel will be, he enters this information into his calendar, which is also context-aware and will therefore remember that these entries belong to Dirk's trip.

The Social Dimension

Users will have a lot of benefit from just using the Semantic Desktop on their own. However, by connecting to others, a number of additional possibilities arise.

Assigning tasks in a group (*Social Interaction*). In the previous scenario, Dirk found out he had to go on a business trip. In fact, he found out about this because he was notified by his boss Claudia, who also uses a Semantic Desktop. Claudia plans to travel to the OOC2007 conference in Oslo to present a research prototype her group has developed as part of the *Torque* project. She does not want to travel alone, so she first needs to find out who of her group members are available while the conference runs. Through the network of Social Semantic Desktops, her calendar application has access to the calendars (or parts of them) of all her contacts. She can ask the calendar to give her a list of all people in her group (**My-Research-Group**) who are working on the *Torque* project (**Torque-Project**) and are free when OOC2007 is on. She finds out that Dirk is free at the desired time. Just like Dirk in the previous scenario, she creates a **Trip-to-OOC2007-Oslo** object and makes it her current context. She also links the trip to the **Torque-Project**. Now, she creates a new **Task** object **Dirk-Prepare-Trip-To-OOC2007**, with a subtask **Dirk-Prepare-Presentation-Slides** and afterwards sends an email to Dirk, asking him to accompany her to the conference, book flights and hotel rooms, and prepare slides for the conference presentation. Her email and the task will of course be automatically linked to the proper context. Also, in this version of the scenario, Dirk no longer has to create the **Trip-to-OOC2007-Oslo** object himself – instead, it will be added to his Semantic Desktop automatically when he gets Claudia's mail.

FUNCTIONALITIES

In this section we describe a list of functionalities that are needed to support the scenarios mentioned above, as well as other scenarios developed during the NEPOMUK project. The SSD is a platform used to develop different kinds of social and semantic applications. These applications share

Table 1. Functionalities of the Social Semantic desktop

Desktop	Annotation, Offline Access, Desktop Sharing, Resource Management, Application Integration, Notification Management
Search	Search, Find Related Items
Social	Social Interaction, Resource Sharing, Access Rights Management, Publish/Subscribe, User Group Management
Profiling	Training, Tailor, Trust, Logging
Data Analysis	Reasoning, Keyword Extraction, Sorting and Grouping

common functionalities which must be supported by the SSD. We have divided them into five groups, which can be considered different aspects of the SSD. Tab. 1 shows the five different aspects and the individual functionalities within each group. Below we briefly describe the use of each functionality.

Desktop. At the *desktop* level, the semantic functionality common to most applications is the ability to add information about any resource. **Annotation** comprises the facilities to store and retrieve semantic relations about anything on the desktop. When Dirk annotates his photos from his trip, he does it from his most favorite photo application (such as Picasa or iPhoto), the annotations are then stored by the SSD. We name this functionality **Application Integration**; applications interact with the SSD by means of different services. When Dirk got notified about the trip to Oslo, this was an example of **Notification Management**. The SSD handles different kinds of mechanisms such as emails, RSS, or text messaging. When Dirk creates a new concept or even a new file on the SSD, the application he uses interacts with the **Resource Management** facilities of the SSD, creating the needed semantics according to the current context and setup. Some of the information Dirk needs when booking his trip are stored on Claudia's desktop. If she is not connected to the network, the **Offline Access** facility exports the relevant information to another desktop. **Desktop Sharing** is the ability for different users of the SSD to work on the same resources. Claudia might write a report of the trip together with Dirk: the resource management is done on Dirk's desktop, but Claudia can access and edit it remotely.

Search. The semantic network created on the desktop unleashes a whole new way of searching on the SSD. **Search** uses the semantic relations as well as social relations to retrieve relevant items. Once an item is found a user can also **Find Related Items**. For instance, when Dirk searches for a flight to Oslo, he can also search for related items and may find out that another company is actually cheaper, based on the experience of his social contacts.

Social. The SSD provides different means of **Social Interaction**, e.g., the embedding of semantic information in emails or text messaging, or the ability to annotate another user's resources. Some desktop level functionalities such as desktop sharing and offline access require the SSD to enable **Resource Sharing**. When Dirk and Claudia collaborate on the trip's report, Dirk might make it accessible to the whole group by adding it to a shared information space. When sharing resources or information on the network, the **Access Rights Management** of the SSD provides ways to define specific rights relations between users, groups and resources. The SSD's **User Group Management** system makes it easy for the rapid creation of new groups from a list of users. These groups can then be used to modify access rights or for resource sharing in a shared information space. E.g., some of Dirk's friends may have subscribed to get notifications of new pictures that Dirk annotates and makes available. The **Publish/Subscribe**

mechanism of the SSD facilitates the creation of feeds of relevant information.

Profiling. If enabled, the **Logging** functionality of SSD logs user activity, which may help to detect the current user's context. The *profiling* of the SSD can be done automatically by **Training**: the SSD learns to predict the user's behavior. The user can still **Tailor** the SSD's intelligent behaviors: some learned contexts can become irrelevant and may need to be re-adapted or removed. The notion of **Trust** on the SSD between people or information sources is also a result of the profiling of the desktop. Dirk might define that he trusts Claudia's information, or Claudia's SSD might learn that Dirk is a trustworthy source of information regarding the *Torque* project.

Data analysis. To support the training behaviors of the SSD or querying related items, the SSD provides different *data analysis* mechanisms such as **Reasoning**. For instance, when Dirk tags a picture with **Preikestolen** and **Norway**, the SSD may infer that Preikestolen is in Norway. This information can later be reused for search. **Sorting and Grouping** supports applications that perform search. The SSD returns items from many sources and people and sorts and groups these items regarding different criteria, using the semantics defined on these resources. The **Keyword Extraction** from resources such as text resources is useful for automatically tagging or summarizing.

ONTOLOGIES

Ontologies form a central pillar in Semantic Desktop systems, as they are used to model the environment and domain of the applications. The common definition of an ontology is "a formal, explicit specification of a shared conceptualization" (Gruber, 1995)

We distinguish four levels of ontologies for the SSD: *Representational*, *Upper-Level*, *Mid-Level and Domain*. The main motivation for having these layers is that ontologies at the foundational levels can be more stable, reducing the maintenance effort for systems committed to using them. A core principle of the Semantic Desktop is that ontologies are used for personal knowledge management. Each user is free to create new concepts or modify existing ones for his *Personal Information Model*. This modeling takes place on the domain-ontology level, but the user is of course free to copy concepts from the other layers and modify them to fit his or hers own needs. In order of decreasing generality and stability the four layers are:

Representation(al) Ontology. Representational ontologies (i.e., ontology definition languages) define the vocabulary with which the other ontologies are represented; examples are RDFS and OWL. The relationship of a representational ontology to the other ontologies is quite special: while upper-level ontologies generalize mid-level ontologies, which in turn generalize domain ontologies, all these ontologies can be understood as instances of the representational ontology. Concepts that might occur in the Representational Ontology level include: classes, properties, constraints, etc.

Upper-Level Ontology. "An upper ontology [...] is a high-level, domain-independent ontology, providing a framework by which disparate systems may utilize a common knowledge base and from which more domain-specific ontologies may be derived. The concepts expressed in such an ontology are intended to be basic and universal concepts to ensure generality and expressivity for a wide area of domains. An upper ontology is often characterized as representing common sense concepts, i.e., those that are basic for human understanding of the world. Thus, an upper ontology is limited to concepts that are meta, generic, abstract and philosophical. Standard upper ontologies are also sometimes referred to as foundational ontologies or universal ontologies." (Semy et. al, 2004) In the upper-level ontology you will find concepts like: **Person**, **Organization**, **Process**, **Event**,

Time, **Location**, **Collection**, etc.

Mid-Level Ontology. "A mid-level ontology serves as a bridge between abstract concepts defined in the upper ontology and low-level domain specific concepts specified in a domain ontology. While ontologies may be mapped to one another at any level, the mid-level and upper ontologies are intended to provide a mechanism to make this mapping of concepts across domains easier. Mid-level ontologies may provide more concrete representations of abstract concepts found in the upper ontology. These commonly used ontologies are sometimes referred to as utility ontologies." (Semy et. al, 2004). The mid-level ontologies may include concepts such as: **Company**, **Employer**, **Employee**, **Meeting**, etc.

Domain Ontology. "A domain ontology specifies concepts particular to a domain of interest and represents those concepts and their relationships from a domain specific perspective. While the same concept may exist in multiple domains, the representations may widely vary due to the differing domain contexts and assumptions. Domain ontologies may be composed by importing mid-level ontologies. They may also extend concepts defined in mid-level or upper ontologies. Reusing well established ontologies in the development of a domain ontology allows one to take advantage of the semantic richness of the relevant concepts and logic already built into the reused ontology. The intended use of upper ontologies is for key concepts expressed in a domain ontology to be derived from, or mapped to, concepts in an upper-level ontology. Mid-level ontologies may be used in the mapping as well. In this way ontologies may provide a web of meaning with semantic decomposition of concepts. Using common mid-level and upper ontologies is intended to ease the process of integrating or mapping domain ontologies." (Semy et. al, 2004). Domain ontologies consist of concepts like: **Group Leader**, **Software Engineer**, **Executive Committee Meeting**, **Business Trip**, **Conference**, etc.

Figure 3 shows how these four layers relate to the four ontologies created and used in the NEPOMUK Project. As detailed in Sect. "*Technology*", we were hesitant to make use of OWL for the representational ontology level in NEPOMUK, and in its place we developed the NEPOMUK Representational Language (Sintek et. al, 2007) (NRL). NRL defines an extension to the semantics offered by RDF and RDFS; the main contribution of NRL is the formalization of the semantics of named graphs. NRL allows multiple semantics (such as open and closed world) to coexist in the same application, by allowing each named graph to have separate semantics. The NEPOMUK Annotation Ontology (NAO) is a basic schema for describing annotations of resources, this is essentially a formalization of the tagging paradigm of Web2.0 applications. A specialized part of NAO is the NEPOMUK Graph Metadata schema (NGM) which allows the description of named graphs, defining meta-data properties such as the author, modification dates and version data.

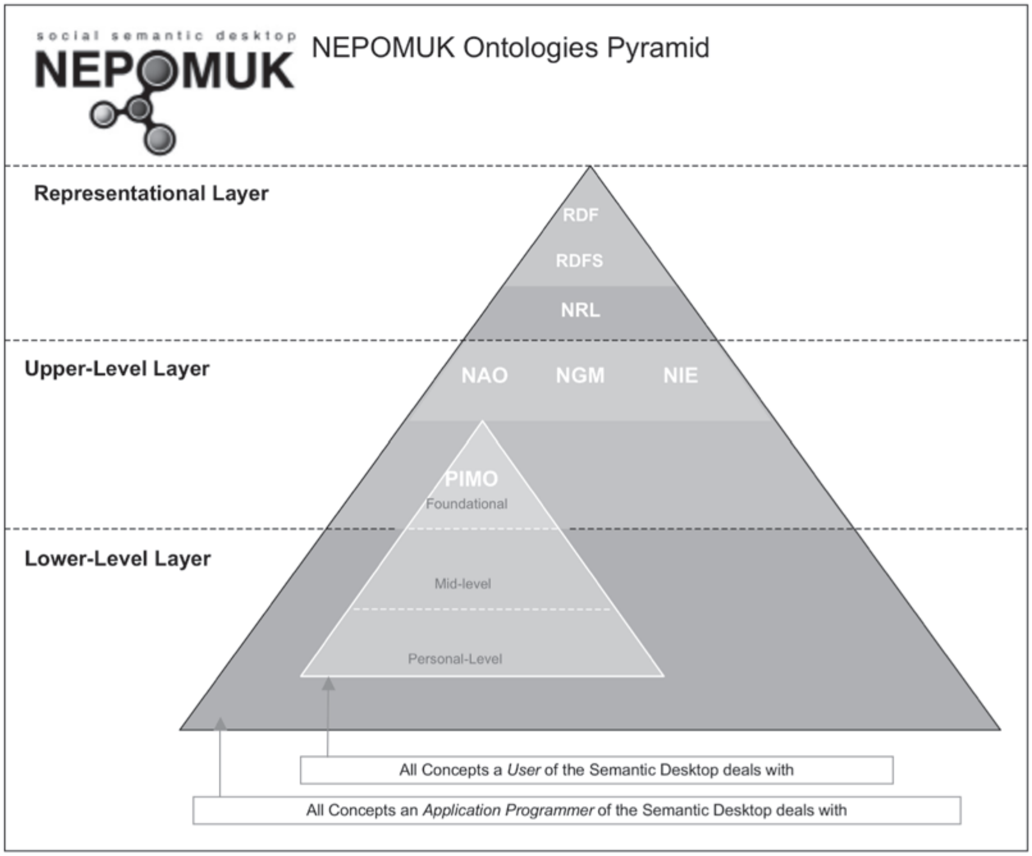
Finally, the NEPOMUK Information Elements ontology (NIE) contains classes and properties for describing objects found on the traditional desktop, such as files (Word documents, images, PDFs), address book entries, emails, etc. NIE is based on existing formats for file meta-data such as EXIF for image meta-data, MPEG7 for multimedia annotations, ID3 for music files, iCal, and others.

TECHNOLOGY

The Social Semantic Desktop deploys the Semantic Web on the desktop computer. Therefore, the technology stack proposed for the Semantic Web (the famous "Layercake"[d] adapted in Figure 4) is adopted for the SSD as well.

However, there are some specific considerations for the desktop scenario: everything on the desktop should identifiable by URIs. This is partially solved for files, where RFC1738[e] specifies the form of file:// URIs, but requires considerable

Figure 3. NEPOMUK ontology pyramid

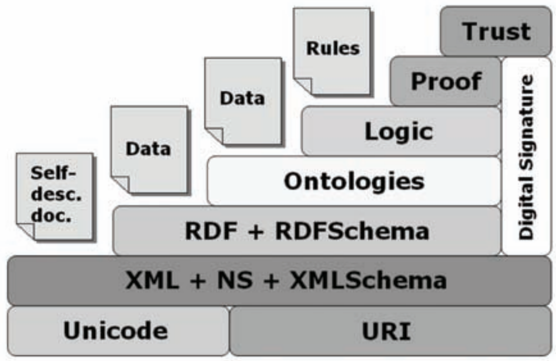


care for other applications which may not represent their data entities as individual files, such as address books or email clients.

Secondly, one can note that for the single desktop scenario there are fewer requirements on aspects such as trust, proof, and signatures. When one progresses to the Social Semantic Desktop, which involves interactions between many users, these aspects must be considered again.

In NEPOMUK we chose not to use the Web Ontology Language (OWL)(McGuiness et. al, 2004) as an ontology language, because of the challenge of dealing with (and implementing) OWL correctly; because our ontological modeling requirements were modest, and, most importantly, because OWL enforces an open-world view of the world, which did not seem to be appropriate for the (local) Semantic Desktop. In a World Wide Web context it is impossible for an application to read all available data, and an open-world assumption is natural, since additional data can be discovered at any moment. However, the open-world assumption makes it impossible to adopt negation as failure (Clark, 1978) which makes practical application development difficult and is also difficult to explain to the end-user. In the context of a local desktop application the situation is quite different, here it is perfectly possible to read all data available, and the closed world assumption makes much more sense. In place of OWL we developed our own ontology specification language called NRL, which uses the closed-world assumption. An additional RDF-based technology that we use widely, but which does not feature in the Semantic

Figure 4. The Semantic Web technology stack



Web stack is the concept of named graphs(Caroll et. al, 2005). This allows one to divide a larger RDF store into sets of RDF statements (graphs), where each is identified with a URI (the name). In this way it is possible to make meta-statements about each graph, such as provenance information. Named graphs thus become an alternative to RDF reification, which also allows making statements about other statements, but is harder to implement and creates a significant overhead. NRL does also allows applying different semantics for different named graphs, thus allowing us to integrate the local closed-world with the open-world of the extended Semantic Web.

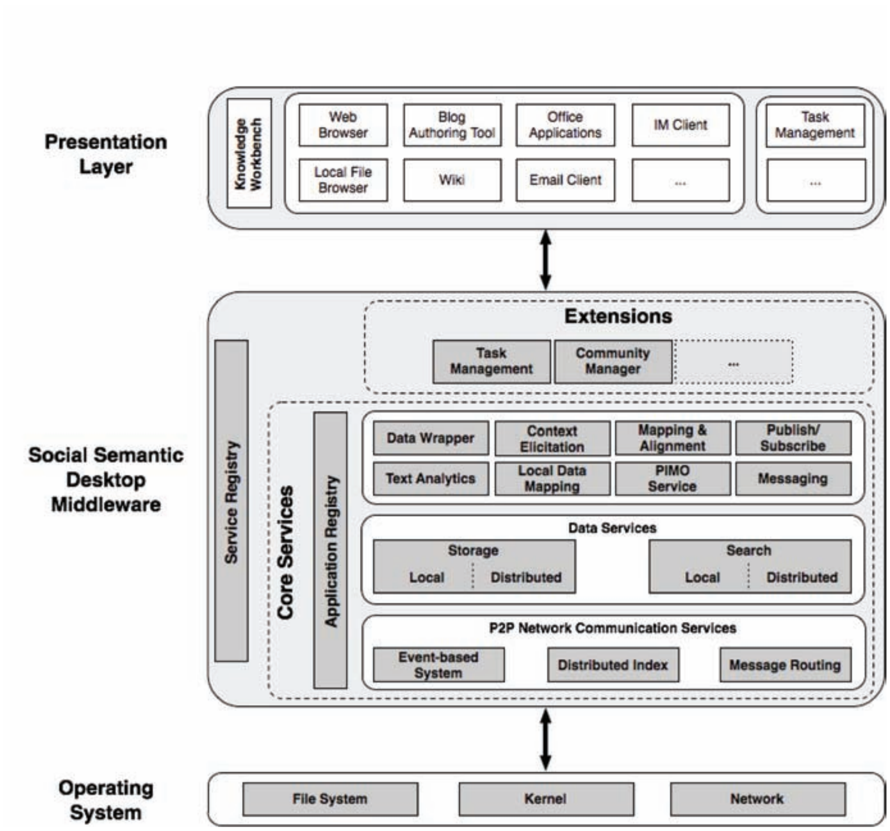
As noted previously, applications on the Semantic Desktop are analogous to services available on the Web. Each application will offer an interface for exposing the functionality it offers. Although a single desktop is not distributed, a network of SSDs is. It therefore suggests itself to adopt the Web Service stack of tools for inter-service communication for the Semantic Desktop: the Web Service Description Language (WSDL) [f] which is used for describing the interfaces of services offered, XML Schema (XSD)[g] which is used for primitive type definitions, and finally the Simple Object Access Protocol (SOAP)[h] which is used for the invocation of services. In Sect. "*Implementation and Engineering Principles*" we give further details on how these technologies work in relation to the Semantic Web technologies presented above.

ARCHITECTURE

In our vision, the standard architecture comprises a small set of standard interfaces which allow application developers to easily extend it and ultimately lead to an evolving ecosystem. Figure 5 depicts this set of interfaces transposed into services, together with their placement in the NEPOMUK architecture. The architecture has to reflect the two aspects of the scenarios introduced in Sect. "*Scenarios*", i.e., the semantic (which can operate on a single desktop) and the social aspect (which is relevant in a network of desktops). To cover these requirements and the functionalities discussed in Sect. "*Functionalities*", the SSD is organized as a Service Oriented Architecture (SOA). Each service has a well defined WSDL interface and is registered at the *Service Registry*. The social aspect of sharing resources over the network is enabled by the peer-to-peer (P2P) infrastructure of the architecture. In the following we present the services of the SSD.

The architecture, as show in Figure 5, is organized in three layers. Like current desktop systems, the desktop environment builds on top of the *Operating System* core, such as the file

Figure 5. Layered architecture of the Social Semantic desktop



system, kernel, and network environment. On the SSD the desktop environment is pooled in the *Social Semantic Desktop Middleware* Layer (SSD Middleware). The SSD Middleware groups the services of the SSD to be used in the Presentation Layer, which provides the user with SSD enabled applications that take advantages of the functionalities of the SSD.

The SSD is made up by individual desktops, which are organized in a P2P fashion. To support the communication between the peers, the SSD Middleware provides *P2P Network communication Services*. To enable information sharing between individual desktops, the RDF metadata of shared resources is stored in the *distributed index* of the P2P system. In NEPOMUK, the P2P system is based on GridVine (Aberer et. al, 2004), which in turn is built on top of P-Grid (Aberer et. al, 2003) and provides a distributed index with RDQL query supports.

Network Communication Services provide an *Event-based System*, which is responsible for the distribution of the events between the SSD peers. On the SSD, the event-based system is used to support the publish/subscribe system. Users as well as other services can use RDF to describe the kind of information they are interested in (e.g., new pictures of Norway become available, the status of a document changes to final, etc.). These subscriptions are stored in the distributed index of the P2P system. An event that was fired carries an RDF query as payload, which is matched against all subscriptions and triggers the notification of the subscriber. In addition, the *Messaging Routing* system uses RDF information to route messages to receiver.

The *Data Services* are responsible to control the insertion, modification, deletion, and retrieval of resources on the SSD. A resource can be a user, a document, a calendar entry, an email, and so on. It provides a service to store the RDF meta-data in the *Local Storage*. Resources and their RDF descriptions can either be added to the SSD manually, or the *Data Wrapper* or *Text Analysis* service extracts the information from desktop applications such as email clients or calendars. Data Wrappers are used to extract metadata form structured data sources (e.g., email headers, calendar entries, etc.). In NEPOMUK, data wrappers are implemented based on Aperture (Aperture, 2007). The Text Analysis service is used to extract metadata from unformatted text (e.g., email bodies, word processing documents, etc.). For local queries and for offline working the RDF metadata is stored in the *Local Storage*. If a resource is shared with other users in an information space, the meta-data is also uploaded to the distributed index of the peer-to-peer file sharing system. The *Search* service can either issue a *local* search in the local storage or a *distributed* search in the underlying P2P system.

Before new metadata can be added to the repository, we have to check whether this metadata describes resources that are already instantiated (i.e., an URI has been assigned) in the RDF repository. In this case, the URI of the resource should be reused, rather then creating a new one. This process is known as information integration (Bergamaschi et. al, 2001). The *Local Data Mapper* service takes over this responsibility in the SSD Middleware. E.g., the Data Wrapping service extracts contact information from the address book and stores the metadata in the repository. Since this is the first time information about the contacts is added to the RDF repository, a new URI is generated for each person. If later the Data Wrapping service extracts information from an email header, the Local Data Mapping service is responsible to lookup whether information about the sender of the email is already in the repository and reuse the corresponding URI instead of creating a new one (Sauermann et. al, 2006).

Ideally only one ontology exists for a domain of interest such as contact data, calendar events. In reality, however, we are faced with many ontologies of (partly) overlapping domains (e.g., FOAF and vCard for contact data, or different personal information models). Since individual users share information over the SSD, it is likely to happen that they use different ontologies for their annotations even when talking about similar domains. Therefore, the SSD Middleware provides a *Mapping & Alignment* service that can be used by other middleware services and services in higher layers to translate RDF graphs from a source ontology to a target ontology.

The SSD Middleware logs the actions a user performs on the resources on his desktop. The logged data is stored in the Local Storage and analyzed by the *Context Elicitation* service to capture the current working context of the user. The context can for example be used to adapt the user interface or to suggest meaningful annotations to the users, depending on the task they are currently working on.

As discussed in Sect. "*Technology*", the services on the SSD use RDF to exchange data. Therefore, services need the capability to generate and process RDF graphs. To simplify the handling of the RDF graphs, the *Ontology Service* provides an easy way to create and manipulate concepts in RDF graphs.

The *Publish/Subscribe System* allows users or other SSD services to subscribe to events on the SSD. The subscriptions are stored as RDF graphs in the distributed index. If an event occurs, the RDF query of the event is matched against the subscriptions. When the subscription, i.e., the RDF query, matches the event, the *Messaging* service looks up the preferred notification media (e.g., email, instant messaging, SMS) and delivers the messages. The Messaging System is further used for synchronous and asynchronous communication between SSD users.

The *Core Services* of the SSD Middleware comprise the services which provide the basic functionalities of the SSD. These services can be accessed via the SSD Application Programming Interface (API). If a developer wants to exploit the SSD Core Services to build his domain-specific application, he can do this as an *extension* of the SSD Middleware. An example for such an extension is the *Task Management* which provides functionalities such as creating, delegating, and manipulating of tasks. Finally, the *Application registry* allows applications from the Presentation Layer to register call back methods at the SSD Middleware if they need to be notified by SSD services, e.g., when a message arrives and has to be displayed to the user in an Instant Messaging Client.

The top layer of the architecture is the presentation layer. It provides a user interface to the services provided by the SSD, and is built using the SSD API. Many desktop applications are possible sources for resources that should be managed by the SSD. Therefore, each desktop application should integrate support for the SSD Middleware. Since this assumption does not hold for most of the current off-the-shelf applications, we developed plug-ins and add-ons to enable a seamless integration with existing applications. These plugins for example extract email or calendar data and add them as resources to the SSD. However, within NEPOMUK we also develop dedicated applications that make use of the SSD API directly, such as a semantic *Wiki* or *Blogging Tools*. (Möller et. al, 2006)

In addition, the *Knowledge Workbench* is the central place to browse, query, view, and edit resources and their metadata. In this way the Knowledge Workbench aims to replace current file management tools such as the MS File Explorer. If the SSD is extended by usage extensions, the application programmer also has to provide the corresponding user interface in the Presentation Layer (e.g., for Task Management, Community Management, etc.).

IMPLEMENTATION AND ENGINEERING PRINCIPLES

As detailed in Sect. "*Architecture*". we deem a Service Oriented Architecture (SOA) to be most suitable for the SSD framework. Furthermore, we decided to use the industry standard SOAP (Simple Object Access Protocol) for exchanging messages between our components. For traditional applications the names and structure of SOAP messages is specified using the Web Service Description Language (WSDL), which in turn uses XML schema data-types to specify the form of the objects being exchanged. However, since the formal modeling of the target domain using ontologies is the core idea of a Semantic Desktop application, the best-practices for SOAs are slightly different. In this section we will discuss some important differences from a traditional SOA system.[i] Basing a system architecture on underlying domain ontologies is similar in nature to Model Driven Architectures (MDA)[j]. However, on the SSD, ontologies take the place of UML models.

Working with RDF

Sect. "*Ontologies*" described the substantial effort that went into the modeling of our domains as ontologies in a formal language. These ontologies give us a very powerful and flexible modeling language, although the structure of instances of such ontologies at first sight seem much more constrained than complex XML schema data-types, the flexibility of RDF introduces some additional requirements for developers of components that should handle RDF instances:

- The structure of the RDF instances received may not be fully known at design time. This means one must take great care that the code does not break when encountering unknown properties in the data, and these unknown properties must also be preserved. In general, programming services for the Semantic Desktop is more like programming services for the web, rather than for traditional desktop applications, and one should follow the general rule of web-programming: "Be strict in what you send and tolerant in what you receive."

- Conversely, other services might not be aware of all the properties the local service uses. Therefore each service must be programmed to be tolerant of missing data and do their best with the data that was provided.

Passing Instances in Messages

Normally, when using SOAP in connection with WSDL and XML schema for data modeling, some mapping is used that will convert the XML schema definition to class definitions in the programming language of choice. Furthermore, stubs and skeletons will be generated for the service themselves, so that the details of communication are hidden. Programming against remote services is then indistinguishable from programming against a local object. However, when using services that pass instances for which the structure is defined by ontologies, the mapping is not so straight forward. Although interaction with RDF data can always be done on a completely general level using basic RDF APIs we are interested in facilitating the job of programmers consuming our services, and allowing them to work on a higher level than RDF triples. We identify three alternatives for programming web services where parameters are instances from an ontology:

- Starting with the ontologies, a number of tools[k] can be used to create a set of Java classes from the ontologies. The service interface is written using parameters of these types, and another tool is used to generate the WSDL and associated XML schema types from these. By sharing the URIs of the concepts in the ontologies with the URIs of the XML schema types, the semantics of messages is retained. The benefit of this approach is that much of the SOAP technology is retained, existing tools may be reused. Also, developers who are not familiar with Semantic Web technology will find that developing and using these services is unchanged from a normal Java environment. The main problem with this approach comes from the fact that ontologies are in general more dynamic than Java class definitions. In particular, as noted in Sect. "*Ontologies*", we expect the personal information models to change frequently. This approach requires a complete re-run of the whole tool chain and a recompile of the system when an ontology changes, as well as introducing some constraints on the ontologies.
- On the other end of the spectrum it is possible to bypass the parameter passing of SOAP all together, and rely more on the Semantic Web technology. Each method offered by a service will take a single RDF document (possibly including several named-graphs), and all the details about the message are given in these RDF graphs. An additional ontology for messages and parameters must be constructed, and some named-graph aware serialization (e.g., TriG or TriX[l]) of RDF is used to construct the XML SOAP messages. This approach was, for instance, used in the SmartWeb project[m]. The benefit of this approach is that the effort that has gone into modeling the ontologies is not duplicated for modeling objects. Also, the full expressivity of RDF may be used when modeling, as it not required that the instances fit into another representation. The backside to this flexibility is that it is significantly harder to program with RDF graphs than with simple Java objects, and both service developers and consumers need good

knowledge about RDF. One can of course envisage new tools that facilitate programming with such RDF messages, but since all the interesting details are hidden inside the RDF parameter, existing SOAP tools for development or debugging are no longer very useful.
- Finally, a hybrid approach of the two methods is possible. Here each method will retain multiple arguments, but each argument is represented by an RDF resource. We envisage two possibilities for doing this: either each parameter is given as a (*named-graph-uri*, *uri*) tuple pointing into an RDF document given as a special parameter; or, alternatively, each parameter is in itself an RDF graph plus the URI of the actual parameter (each RDF graph may contain several resources). The benefit of this method is that the changes in the ontology do no longer require a recompile of the system, while at the same time allowing slightly more compatibility with existing SOAP tools. The problem with this method remains that both client and server programmers need in-depth knowledge of RDF and the ontologies used.

Regardless of which of the three alternatives one chooses, it remains an important issue to make sure that the formal description of the services (i.e., the WSDL+XML Schema definitions) remain semantically correct and retain the pointers to the ontology concepts which the parameters represent. As mentioned, for the first approach this can be handled by well chosen URIs for the XMLSchema types. For the second and third approach the parameters have the form of simple string objects in both the WSDL definition and the SOAP messages, since the RDF serialization is represented as a string. However, both versions of WSDL available at the time of writing allow extensions to the WSDL format itself[n], and additional constraints on the type or form of the RDF instances contained inside the string parameters may be specified here. This is the approach taken by the *Semantic Annotation for WSDL and XML Schema* (SAWSDL) working group[o] and the NEPOMUK project makes use of their standard.

In this section we have considered a very lightweight approach to semantically enriching SOAP Web Services by passing RDF-based parameters. If a more powerful approach is required, the reader is advised to look into OWL-S[p] and the Web Service Modeling Language (WSML)[q], both defining much more sophisticated frameworks for Semantic Web Services.

RELATED WORK

In the following we review relevant research and development approaches for the Social Semantic Desktop. After providing a brief description, we discuss the lessons learned and state our conclusions.

Gnowsis (Sauermann, 2003) was among the first research projects targeting a Semantic Desktop system. Its goal is to complement established desktop applications and the desktop operating system with Semantic Web features, rather than replacing them. The primary focus of Gnowsis is on *Personal Information Management* (PIM). It also addresses the issues of identification and representation of desktop resources in a unified RDF graph. Gnowsis uses a Service Oriented Architecture (SOA), where each component defines a certain interface and it is available as an XML/RPC service.

The **Haystack** (Quan et. al, 2003) project presents a good example for an integrated approach to the SSD field. Inter-application barriers are avoided by simply replacing these applications with Haystack's own word processor, email client, image manipulation, instant messaging, etc. Haystack allows users to define their own arrangements and connections between views of information, thus making it easier to find information located in the personal space. The Haystack architecture can be split into two distinct parts: the Haystack Data Model (HDM) and the Haystack Service Model (HSM). The Data Model is the means by which the user's information space is represented, similar to what has been discussed in Sect. "*Ontologies*". The set of functionalities within Haystack is implemented by objects in the Haystack Service Model (HSM). Haystack has a standard three-tiered architecture, consisting of a user interface layer (the client), a server/service layer, and a database. Haystack was groundbreaking in terms of the dynamic creation of user interfaces, but the project ended before establishing any standards.

Another relevant personal information management tool is the **Semex System** (SEMantic EXplorer)(Dong et. al, 2005). Like other tools, it organizes data according to a domain ontology that offers a set of classes, objects and relationships. Semex leverages the Personal Information Management (PIM) environment to support on-the-fly integration of personal and public data. Information sources are related to the ontology through a set of mappings. Domain models can be shared with other users in order to increase the coverage of their information space. When users are faced with an information integration task, Semex aids them by trying to leverage data collected from previous tasks performed by the user or by others. Hence, the effort expended by one user later benefits others. Semex begins by extracting data from multiple sources and for these extractions it creates instances of classes in the domain model. It employs multiple modules for extracting associations, as well as allowing associations to be given by external sources or to be defined as views over other sets of associations. To combine all these associations seamlessly, Semex automatically reconciles multiple references to the same real-world object. The user browses and queries all this information through the domain model.

A similar idea is exploited by the **IRIS** Semantic Desktop(Cheyer et. al, 2005) ("Integrate. Relate. Infer. Share"), an application framework that enables users to create a "personal map" across their office-related information objects. IRIS offers integration services at three levels:

- Information resources (e.g., email messages, calendar appointments) and applications that create and manipulate them must be accessible to IRIS for instrumentation, automation, and query. IRIS offers a plug-in framework, in the style of the Eclipse architecture, where "applications" and "services" can be defined and integrated within IRIS. Apart from a very small, lightweight kernel, all functionality within IRIS is defined using a plug-in framework, including user interface, applications, back end persistence store, learning modules, harvesters, etc. Like Haystack, inter-application barriers do not exists, because all applications are made from scratch for IRIS.
- A Knowledge Base provides the unified data model, persistence store, and query mechanisms across the information resources and semantic relations among them. The IRIS user interface framework allows plug-in applications to embed their own interfaces within IRIS and to interoperate with global UI services, such as notification pane, menu toolbar management, query interfaces, the link manager, and suggestion pane.

DeepaMehta(Richter et. al, 2005) is an open source Semantic Desktop application based on the Topic Maps standard[r]. The DeepaMehta UI, which runs through a Web browser, renders Topic Maps as a graph, similar to concept maps. Information of any kind as well as relations between information items can be displayed and edited in the same space. The user is no longer confronted with files and programs. DeepaMehta has a layered, service oriented architecture. The main layer is the application layer, which offers various ways for the presentation layer to communicate with it via the communication layer (API, XML Topic Maps (XTM) export, EJB, SOAP). Finally, the storage layer holds all topics and their data either in a relational database or simply in the file system.

Other relevant projects include **Beagle++** (Brunkhorst et. al, 2005), a semantic search engine which provides the means for creating and retrieving relational metadata between information elements present on the destkop, **DBIN** (Tummarello et. al, 2006), which is similar to a file sharing client and connects directly to other peers, **PHLAT** (Cutrell et. al, 2006), a new interface for Windows, enabling users to easily specify queries and filters, attempting to integrate search and browse in one intuitive interface, or **MindRaider**[s], a Semantic Web outliner, trying to connect the tradition of outline editors with emerging SW technologies. The **MyLifeBits** project by Microsoft Research is a lifetime store of multimedia data. Though the system does not intent to be a SSD, one can learn from it how to integrate data, i.e., how to manage the huge amount of media and how to classify/retrieve the data(Gemmell et. al, 2002). It combines different approaches from HCI (Computer-Human Interaction) and information integration, while it lacks a conceptual layer beyond files. The **Apogée**[t] project deals with data integration in applications related to Enterprise Development Process (ECM). It aims at building a framework to create Enterprise Development Process-oriented desktop applications, independent from vendor or technologies. Finally, starting from the idea that everything has to do with everything, has a relationship with everything, **Fenfire**[u] is a Free Software project developing a computing environment in which you can express such relationships and benefit from them.

Although the systems we have looked at focus on isolated and complementary aspects, they clearly influenced the vision of the SSD presented here. Some of the architectural decisions made in the NEPOMUK project and presented in this chapter are similar to those of platforms like Haystack, IRIS, and DeepaMetha, e.g., in that we present a User Interface Layer, a Service and a Data Storage Layer. The modular architecture, also identified within the Haystack, SEMEX, and DeepaMetha systems, as well as the standardized APIs offer an easy way of introducing new components. Our approach guarantees that each component may be changed without affecting other components it interacts with. The interaction has to suffer only in the case in which the API of the component is modified. The NEPOMUK Architecture also provides service discovery functionalities: the NEPOMUK Registry providing a proper support for publishing and discovering the existing NEPOMUK Services by using a standard interface.

CONCLUSION

We presented the Social Semantic Desktop as a comprehensive approach to information handling. Oriented at the needs of knowledge workers, this approach centers around supporting the main information-oriented activities: The articulation of knowledge and the generation of new information items; the structuring, relating, and organization of information, and the sharing of formal and informal information within networks of co-operating people. From this, we derived key functionalities of the desktop, but also for search, social interaction, profile building, and data analysis.

Building the SSD relies on basic principles: Whatever appears within the personal workspace is treated as an information item. Content, relations, special services all refer to formal annotations of such information items, which in turn link between information items and personal information models. Unifying the flexibility and personal liberty of expressing whatever concepts seem relevant with the commitment to socially shared conceptualizations results in a layered hierarchy of ontologies which allow the necessary differences in stability, sharing scope, and formality. Integrating the tools of everyday information processing asks for an easy and flexible integration of existing desktop applications. Finally, the adoption of Semantic Web standard technology for representation and communication enables the easy transgression from personal annotated information to shared Semantic Web content.

Consequently, the architecture of the SSD combines standards-based data repositories with a rich middleware, which in particular allows for manifold service integrations and communications. On top of that, various presentation clients and specific applications support whatever activities are performed on the desktop. Such applications may be highly domain-specific, although core functionalities of knowledge work trigger standard applications, e.g., for document processing, task management, communication, etc.

The design decisions presented result in particular implementation and engineering principles; we outlined the adaptation to RDF, the service integration, and the message passing mechanisms in particular.

In summary, the SSD offers the basic technology and tools for everyday information processing by knowledge workers. In order to reach the intended wide acceptance and broad uptake, care was taken to make the central software components available under open source licenses, and to encourage the development and contribution of application-specific enhancements and adaptations. The concept of the SSD is promising and relies on a number of techniques which reach their maturity right now – consequently, a number of research and development projects are under way and contribute to the overall evolution of the concept.

Following the realizations described in this chapter, we see the SSD as a basis for the self-motivated generation of semantically annotated information, which will not only help the individual by allowing multitudes of specific services and support, but will also initiate a wide movement to populate the Semantic Web.

FUTURE RESEARCH DIRECTIONS

Although the ideas of the Social Semantic Desktop are based on solid foundations as presented here, the research areas surrounding this topic are still in their infancies. We will briefly discuss some of the pre-dominant challenges in the coming years:

Trust and Privacy. As pointed out in the Semantic Web technology stack presented earlier, a crucial component for any high-level Semantic Web service is the issue of trust and privacy. Trust touches on a wide range of issues, from the technical issues of cryptographic signatures and encryption, to the social issues of trust in groups and among individuals. These issues are all as valid for the Social Semantic Desktop as for the Semantic Web in general, or perhaps even more so, as people are less critical of putting personal data on their personal desktop.

User, group, and rights management. When a single personal Semantic Desktop allows easy sharing of information with the network of Social Semantic Desktops, determining access rights for this information becomes very important. The Social Semantic Desktop sets new requirements for distributed authentication, flexible group management, and fine-grained access rights, all the while remaining intuitive and unobtrusive for the end user.

Integration with the wider Semantic Web and Web 2.0. Currently we are talking about the Social Semantic Desktop as a network of Semantic Desktops built on the same standards. It is important to remember that the key benefit of Semantic technology is the easy access to integration with anyone using the same representational languages and ontologies. The growth of feature-rich Web applications is growing rapidly, and ensuring a strong bond between the Semantic Desktop and these services is a continuous challenge.

Ontologies and Intelligent Services. To date ontologies have been used to introduce a common vocabulary for knowledge exchange. On the Social Semantic Desktop ontologies are used to formalize and categorize personal information. This introduces many interesting issues around ontology versioning, ontology mapping, and ontology evolution. Furthermore, using ontologies with well-defined semantics will allow intelligent services to be built (e.g., using reasoning) that allow for much more than just browsing and (simple) searching.

User Evaluation. The underlying thesis of the whole (Social) Semantic Desktop effort is that the added semantics will improve productivity and enable new forms of cooperation and interaction which were not previously possible. In-depth empirical evaluation with real users of a Social Semantic Desktop systems are required to determine if this thesis really holds.

REFERENCES

Aberer, K., Cudré-Mauroux, P., Datta, A., Despotovic, Z., Hauswirth, M., Punceva, M., & Schmidt, R. (2003). P-Grid: A self-organizing structured P2P system. *SIGMOD Record*, *32*(3), 29–33. doi:10.1145/945721.945729

Aberer, K., Cudré-Mauroux, P., Hauswirth, M., & Pelt, T. V. (2004). Gridvine: Building Internet-scale semantic overlay networks. In S. A. McIlraith, D. Plexousakis, F. van Harmelen (Eds.), *The Semantic Web – ISWC 2004: Third International Semantic Web Conference*, 107-121. Springer Verlag.

Aperture: A Java framework for getting data and metadata, Last visited March 2007. http://aperture.sourceforge.net/.

Bergamaschi, S., Castano, S., Vincini, M., & Beneventano, D. (2001). Semantic integration and query of heterogeneous information sources. *Data & Knowledge Engineering*, *36*(3), 215–249. doi:10.1016/S0169-023X(00)00047-1

Berners-Lee, T., & Fischetti, M. (1999). *Weaving the Web – The original design and ultimate destiny of the World Wide Web by its inventor*. Harper San Francisco.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American.*

Brunkhorst, I., Chirita, P. A., Costache, S., Gaugaz, J., Ioannou, E., Iofciu, T., Minack, E., Nejdl, W., & Paiu. R. (2006). *The Beagle++ Toolbox: Towards an extendable desktop search architecture (Technical report)*, L3S Research Centre, Hannover, Germany.

Bush, V. (1945, July). *As we may think*. The Atlantic Monthly.

Carroll, J. J., Bizer, C., Hayes, P., & Sticker, P. (2005). Named graphs, provenance and trust, In A Ellis, T. Hagino (Eds.), *WWW 2005: The World Wide Web Conference*, 613-622.

Cheyer, A., Park, J., & Giuli, R. (2005, November 6). IRIS: Integrate. Relate. Infer. Share. In S. Decker, J. Park, D. Quan, L. Sauermann (Eds.), *Semantic Desktop Workshop at the International Semantic Web Conference*, Galway, Ireland, *175*.

Clark, K. L. (1978). Negation as failure. In J. Minker (Ed.), *Logic and Data Bases*, Plenum Press, New York, 293-322.

Cutrell, E., Robbins, D. C., Dumais, S. T., & Sarin, R. (2006, April 22-27). Fast, flexible filtering with PHLAT – Personal search and organization made easy. R. E. Grinter, T. Rodden, P. M. Aoki, E Cutrell, R. Jeffries, G. M. Olson (Eds.), *Proceedings of the 2006 Conference on Human Factors in Computing Systems,* CHI 2006, Montréal, Québec, Canada. ACM 2006, ISBN 1-59593-372-7.

Decker, S., & Frank, M. (2004, May 18). The networked semantic desktop. In C. Bussler, S. Decker, D., Schwabe, O. Pastor (Eds.), *Proceedings of the WWW2004 Workshop on Application Design, Development and Implementation Issues in the Semantic Web*, New York, USA.

Dong, X., & Halevy, A. Y. (2005). A platform for personal information management and integration. In M. Stonebraker, G. Weikum, D. DeWitt (Eds.), *Proceedings of 2005 Conference on Innovative Data Systems Research Conference*, 119-130

Engelbart, D. C. (1962). *Augmenting human intellect: A conceptual framework (Summary report)*, Stanford Research Institute (SRI).

Gemmell, J., Bell, G., Lueder, R., Drucker, S., & Wong, C. (2002, December 1-6). MyLifeBits: Fulfilling the memex vision. In *ACM Multimedia*, Juan-les-Pins, France, 235–238.

Gifford, D. K., Jouvelot, P., Sheldon, M. A., & O'Toole, J. W., Jr. (1991, October). Semantic file systems. In 13th *ACM Symposium on Operating Systems Principles*.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, *43*, 907–928. doi:10.1006/ijhc.1995.1081

Hendler, J. (2001, March/April). Agents and the SemanticWeb. *IEEE Intelligent Systems*, *16*(2), 30–37. doi:10.1109/5254.920597

McGuinness, D. L., & van Harmelen, F. (2004, February). *OWL Web Ontology Language Overview (Technical report)*. http://www.w3.org/TR/2004/REC-owl-features-20040210/.

Möller, K., Bojārs, U., & Breslin, J. G. (2006, June 11-14). Using semantics to enhance the blogging experience. In Y. Sure, J. Domingue (Eds.), *The Semantic Web: Research and Applications, 3rd European Semantic Web Conference*, ESWC 2006 Proceedings, Budva, Montenegro, 679-696.

Nelson, T. H. (1965). A file structure for the complex, the changing, and the indeterminate. In *ACM 20th National Conference Proceedings*, 84-100, Cleveland, Ohio.

Oren, E. (2006). An overview of information management and knowledge work studies: Lessons for the semantic sesktop. In S. Decker, J. Park, L. Sauermann, S. Auer, S. Handschuh (Eds.), *Proceedings of the Semantic Desktop and Social Semantic Collaboration Workshop (SemDesk 2006) at ISWC 2006*. Athens, GA, USA.

Quan, D., Huynh, D., & Karger, D. R. (2003). Haystack: A platform for authoring end user Semantic Web applications. In D. Fensel, K.P. Sycara, J. Mylopoulos (Eds.), *The Semantic Web – ISWC 2003: International Semantic Web Conference*, Proceedings, 738-753.

Richter, J., Völkel, M., & Haller, H. (2005). DeepaMehta – A Semantic desktop. In *Proceedings of the 1st Workshop on the Semantic Desktop - Next Generation Personal Information Management and Collaboration Infrastructure* at ISWC 2005, Galway, Ireland.

Sauermann, L. (2003). The Gnowsis – *Using Semantic Web Technologies to Build a Semantic Desktop*. Diploma Thesis, Technical University of Vienna, 2003.

Sauermann, L., Grimnes, G. A. A., Kiesel, M., Fluit, C., Maus, H., Heim, D., et al. (2006). Semantic desktop 2.0: The gnowsis experience. In I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, L. Aroyo (Eds.), *The Semantic Web – ISWC 2006: 5th International Semantic Web Conference*, Athens, GA, Proceedings.

Semy, S.K., Pulvermacher, M.K., & Obrst, L.J. (2004). *Toward the use of an upper ontology for U.S. Government and U.S. Military domains: An evaluation. (Technical report)*. MITRE, September 2004.

Sintek, M., van Elst, L., Scerri, S., & Handschuh, S. (2007). Distributed knowledge representation on the social semantic desktop: Named graphs, views and roles in NRL. In E. Franconi, M. Kifer, W. May (Eds.), *The Semantic Web – ESWC 2007: The 4th European Semantic Web Conference (ESWC 2007) Proceedings*.

Tummarello, T., Morbidoni, C., & Nucci, M. (2006). Enabling Semantic Web communities with DBin: An overview. In I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, L. Aroyo (Eds.), *The Semantic Web – ISWC 2006: 5th International Semantic Web Conference*, Athens, GA, Proceedings, 943-950.

ADDITIONAL READINGS

Current and recent research and development in the SSD domain has already been presented in Sect. "*Related Work*". However, one influence that has not been covered in this chapter so far, but is closely related to the idea of a Semantic Desktop is the concept of *Semantic File Systems* – file systems in which files are not organized hierarchically, but rather according to their metadata. The concept and an early implementation are described in detail in (Gifford et. al, 2001).

Finally, as another entry point for additional reading, we would like to point the reader to the series of *Semantic Desktop Workshops* which were co-located with the International Semantic Web Conferences in 2005[v] and 2006[w].

From a historical perspective, the most important references in the Social Semantic Desktop domain are those by Vannevar Bush (1945) and Doug Engelbart (1962) which we mentioned in Sect. "*Background*". Another important early influence is certainly Ted Nelson's work on hypertext (Nelson, 1965). A modern vision of those ideas is a paper by Decker and Frank (2004), which also coined the term "Semantic Desktop". Of course, any work that is based on the ideas of the Semantic Web is not complete without references to seminal papers such as (Berners-Lee et. al, 2001) or (Hendler, 2001). In fact, the original vision of the World Wide Web itself already contained the idea of an information space that would reach from "mind to mind" (Berners-Lee, 1999); a thought that is central to the SSD.

Most of the references given in this chapter are of a technical nature. However, one has to keep in mind that the SSD is a tool for *information management* and *knowledge work*, and thus psychological and sociological research into the nature of knowledge work in any form are relevant as well. Oren (2006) provides a detailed overview of literature in this field, with the intention of applying the lessons learned to the development of the Semantic Desktop.

ENDNOTES

[a] RDF: http://www.w3.org/RDF/

[b] The NEPOMUK Project is supported by the European Union IST fund, grant FP6-027705

[c] Within the NEPOMUK Project, these personas were created by distilling typical users from a series of interviews and evaluations with our use-case partners.

[d] Tim Berners-Lee talk, XML and the Web: http://www.w3.org/2000/Talks/0906-xml-web-tbl/

[e] RFC1738: http://tools.ietf.org/html/rfc1738

[f] WSDL: http://www.w3.org/TR/wsdl

[g] XML Schema: http://www.w3.org/XML/Schema

[h] SOAP: http://www.w3.org/TR/soap

[i] In this chapter we make the assumption that a modern object-oriented programming language like Java will be used for implementation, but observations and solutions are equally valid for most other languages.

[j] MDA: http://www.omg.org/mda/

[k] RDFReactor: http://wiki.ontoworld.org/wiki/RDFReactor; RDF2Java: http://rdf-2java.opendfki.de; Elmo: http://openrdf.org, etc.

[l] TriG/TriX: http://www.w3.org/2004/03/trix/

[m] SmartWeb: http://www.smartweb-project.de/

[n] Language Extensibility in WSDL1: http://www.w3.org/TR/wsdl#_language and in WSDL2: http://www.w3.org/TR/wsdl20#language-extensibility

[o] SAWSDL: http://www.w3.org/TR/sawsdl/

[p] OWL-S: http://www.daml.org/services/owl-s/

[q] WSML: http://www.wsmo.org/wsml/

[r] ISO/EIC 13250:2003: http://www.y12.doe.gov/sgml/sc34/document/0129.pdf

[s] MindRaider: http://mindraider.sourceforge.org/

[t] Apogée: http://apogee.nuxeo.org/

[u] Fenfire: http://www.fenfire.org/

[v] SemDesk2005: http://tinyurl.com/yuxpld

[w] SemDesk2006: http://tinyurl.com/2hqfak

APPENDIX: QUESTIONS FOR DISCUSSION

Q: I prefer to handle my photo collection in a web 2.0 photo sharing environment. Is this compatible with the Social Semantic Desktop? May I keep the work I have invested here?

A: Yes. Every photo in your collection can be reached via a specific URI, thus it can be handled as a particular information item in the SSD. You might implement a suitable wrapper to transfer local annotations from your SSD onto the photo sharing platform, if you intend to disclose this information.

Q: The Social Semantic Desktop presupposes that everything is an information item. What about entities which are not information but real-world objects? Can I manage them in the SSD and add comments about them, e.g., about my friend's cat?

A: The solution is easy: Just create an appropriate description of the real world object within your SSD, thus creating an URI for the object in question. Let's say you create an instance of the class pet in your SSD (assuming you have this category within your SSD) and describe it as 'well-known house cat'. Then you can link this instance to, e.g., a photo of the animal, or you add an 'owns' link which connects it to the URI of your friend, and so on. Making an arbitrary object re-appear as a formal instance within the SSD models is often called 're-birthing', btw.

Q: Think about scenarios you encounter every day, and where the SSD can make your work easier.

A: The answer is of course a personal one, but for a typical knowledge worker (researchers, students, journalists, etc.) here are some example ideas:

- Show me related appointments when composing emails to a person, i.e., You also have lunch with Claudia next week.
- Show me previously viewed PDF documents on the same topic when researching on Wikipedia.
- Remember my meal and window preferences when booking flights.
- Remind me of my previous idea of combining topic A with topic B when reviewing my topic A notes.
- Let me connect an incoming email from a student to the colleague who introduced me to that student.

Q: What are the benefits of the Social Semantic Desktop compared to solution such as Microsoft Exchange server or the tight integration of applications on MacOSX? They also fulfil many of the functionalities required by the scenarios outline in this chapter.

A: The Social Semantic Desktop is different because of the standards used to build it. Firstly, by basing the representational layers of the Semantic Desktop on the existing (Semantic) Web standards we enable interoperability by a wide range of existing projects, and secondly, by creating new standards for desktop integration and data-formats we encourage future software developers to build on top of the Semantic Desktop. On the Semantic Desktop both the applications and the data encourages open access, and this exactly the opposite of the vendor lock-in that for instance Exchange Server aims for.

Q: Inspect the current state of the Semantic Web and the data available. What data-sources and/or ontologies do you think could be useful for integration with the Semantic Desktop?

A: The answer will of course change as the Semantic Web evolves, but at the time of writing relevant ontologies include:

- The Friend-of-a-Friend project – http://xmlns.com/foaf/spec
- The Description-of-a-Project schema – http://usefulinc.com/doap
- The Semantically Interlinked Online Communities project – http://siocproject.org/
- Dublin Core for basic meta-data – http://dublincore.org/
- Useful data-sources and/or web-services include:
- GeoNames for (reverse) geocoding – http://www.geonames.org/
- DBpedia for a Semantic Web view of Wikipedia – http://DBpedia.org/

This work was previously published in Semantic Web Engineering in the Knowledge Society, edited by J. Cardoso and M. Lytras, pp. 290-314, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 8.2
Explaining Semantic Web Applications

Deborah L. McGuinness
Tetherless World Constellation, Rensselaer Polytechnic Institute (RPI),Stanford University, KSL, USA

Vasco Furtado
University of Fortaleza, UNIFOR, Brazil

Paulo Pinheiro da Silva
University of Texas at El Paso (UTEP), USA

Li Ding
Tetherless World Constellation, Rensselaer Polytechnic Institute (RPI),Stanford University, KSL, USA

Alyssa Glass
Stanford University, KSL, USA

Cynthia Chang
Stanford University, KSL, USA

ABSTRACT

In this chapter, we introduce the concept of explanation for Semantic Web applications by providing motivation, description, and examples. We describe the Inference Web explanation toolkit that provides support for a broad range of explanation tasks ranging from explaining deductive reasoning, to information extraction, to hybrid integrated learning systems. We argue that an explanation solution such as the one we endorse is required if we are to realize the full potential of hybrid, distributed, intelligent Web agents that users can trust and use.

INTRODUCTION

Question answering on the Semantic Web (SW) typically includes more processing steps than database retrieval. Question answering can be viewed as an interactive process between a user and one or more intelligent software agents. Using queries, user preferences, and context, intelligent agents may locate, select and invoke services and, if necessary, compose these services to produce requested results. In other words, the web paradigm shifts from one where users mainly retrieve explicitly stated stored information to a paradigm where application results are answers to potentially complex questions that may require inferential capabilities in addition to information retrieval. Web applications with question answering capabilities may still use information retrieval techniques to locate answers, but they may also need to use additional semantics such as encoded term meanings to support additional methods of information access (such as targeted database queries or knowledge base queries) along with information manipulations (such as reasoning using theorem provers, or inductive or deductive methods). Examples of this new, more complex reality include the automatic composition of web services encoded in OWL-S or semi-automatic composition of services as provided by workflows. Ontology-enhanced search is another example of how Semantic Web technology can provide and is providing new directions for a category of "smart" search applications. Many other SW applications are emerging with a common theme of increasing knowledge and autonomy. This new context generates an additional requirement for effective use of SW applications by typical users: *applications must provide explanation capabilities showing how results were obtained*. Explanations are quickly becoming an essential component in establishing agent credibility (e.g., Glass et al, 2008) and result credibility (e.g., Del Rio and Pinheiro da Silva, 2007) by providing process transparency, thereby increasing user understanding of how results are derived. Explanations can also identify information sources used during the conclusion derivation process. In the context of the SW, explanations should be encoded in a way that they can be directly or indirectly consumed by multiple agents, including both human users and software systems.

In this chapter we describe explanation as a special kind of pervasive SW functionality, in the sense that a SW application may need to provide transparency concerning its results. We first analyze some distinct application paradigms in the SW context, and for each paradigm we identify explanation requirements. We then describe a general framework, called Inference Web (IW) (McGuinness and Pinheiro da Silva, 2004) that includes the Proof Markup Language (PML) (McGuinness, et al., 2007, Pinheiro da Silva, McGuinness, Fikes, 2006), a modularized ontology describing terms used to represent provenance, justifications and trust relations. IW includes a set of tools and methods for manipulating PML-encoded result justifications. Using Inference Web, and its PML interlingua, applications may provide interoperable and portable explanations that support intelligent, interactive application interfaces. After the description of the IW framework and the PML interlingua, we will exemplify how PML and IW have been used to explain the results and behaviors of a wide range of applications including intelligent personal agents, information extraction agents, and integrated learning agents.

A CONCEPTUAL FRAMEWORK FOR EXPLAINING RESULTS FROM SEMANTIC WEB APPLICATIONS

We investigate the correspondence between SW application paradigms and their explanation requirements.

Semantic Web Application Characterization

SW applications are geared to take advantage of vast amounts of heterogeneous data with potentially varying amounts of semantic markup. They concentrate on identifying and meaningfully combining available semantic markup in order to derive complex results. Below we briefly characterize the SW applications features considered important from an explanation perspective: collaboration, autonomy, and use of ontologies.

Collaboration

Collaboration requires agents to interact and share knowledge with the common goal of solving a particular problem. Collaboration raises issues concerning how to create, use, and share a combination of provenance, trust and reputation throughout distributed reasoning processes. Wikis, for example, are gaining popularity as collaborative tools for human agents, although they do not provide a precise infrastructure for recording and reusing provenance information. A *Semantic Wiki* is a wiki application enhanced with Semantic Web technologies that support wiki content annotation that goes beyond simple structured text and untyped hyperlinks. Semantic Wikis provide the ability to represent metadata about content, term meanings, and inter-relationships. Provenance support is typically somewhat limited, in both ordinary wikis and in semantic wikis, to keeping track of which author (if a login authentication process is included) made which updates and when.

Content Management Systems (CMS) are one of the most common uses of wikis for knowledge management. Semantic Wikis aim to enhance ordinary wikis by allowing users to make their internal knowledge more explicit and formal, enabling search methods that go beyond simple keyword search. In this case, provenance information may be included in these searching capabilities. Other collaborative systems are aimed at Personal Information Management (PIM) or community knowledge management. The ability to store project history, and to utilize tools that access and perform intelligent queries over this history, is one of the benefits brought by Semantic Wikis used for content management.

The collaborative characteristic is also prominent in applications developed via the integration of multi-agent systems and Semantic Web services. In this situation, collaborating agents are software programs such as digital assistants that manage electronic information. These collaborating agents can proactively engage in tasks on behalf of their users to find, filter, assess and present information to the user in a more appropriate manner (Maes, 1994). Several types of multi-agent applications have been developed such as office organization (Pyandath & Tambe, 2002); technical support (Sullivan et al. 2000); and information retrieval (Rhodes et al., 1996). Again, most of these collaborating agents provide little support for storing and retrieving provenance information about how they work internally, and in particular, they provide only limited access to information about how they collaborate. However, end user activities may require the integration of multi-agent systems and Semantic Web services. Personal agents may also need user models, to allow them to better perform tasks in compliance with user needs and preferences.

Distributed solutions for multi-agent problems can alternatively be represented using a reactive multi-agent architecture. In these domains, the individual agents have little autonomy. The “intelligence” used to solve problems comes from intensive inter-agent communication. This paradigm is typically used on the web, where heterogeneity and loosely-coupled distributed systems are common. Thus, interactions between agents or system components must not be rigidly specified at design time, but opportunistically built though the use of new services as they become available. Prior knowledge of such services is thus not necessary (and often not practical nor desirable). Instead, agents must discover services by accessing a *service description* that can be semantically described by means of ontologies in which descriptive expressions or concepts are attached to services.

Autonomy

An individual agent’s autonomy controls its ability to act independently. Barber and Martin (1999) consider an agent’s degree of autonomy with respect to a particular goal that the agent is actively pursuing. Within this context, they define the degree of autonomy to be (1) the degree to which the decision making process was used to determine how that goal should be pursued; and

(2) how free the agent is from intervention by other agents. Traditional web-based applications have very little autonomy, since they primarily take direct input from the user and retrieve information consistent with the query. For example, a typical web search engine's primary interaction mechanism is based on communication between the user and the search engine. The degree of autonomy of the search engine is said to be low because the user is required to reformulate and resubmit the query when the original query is not satisfactorily answered by the engine. In contrast with typical search engines, SW applications have more autonomy while pursuing goals. For example, online shopping agents have autonomy over how to find answers to shopping queries concerned with product location, price comparison, or rating information. ShopBot can make several autonomous decisions, such as which content sources to use, which services to call and compose, and how to enhance the query with background representation information, all in an attempt to answer the user's question as efficiently and usefully as possible. In general, the development of autonomous problem-solving software agents in the Semantic Web is increasingly gaining popularity.

Use of Ontologies

Semantic Web applications are increasingly using large amounts of heterogeneous semantic data from multiple sources. Thus, the new generation of Semantic Web applications must be prepared to address issues associated with data of varying quality. Intelligence in these large-scale semantic systems comes largely from the system's ability to operate effectively with large amounts of disparate data.. In this context, ontologies are used to support information integration as well as to identify inconsistencies between data coming from multiple sources. Ontologies are being used to provide declarative specifications of term meanings. Agents can then decide to use a term meaning as specified in a particular ontology, and when multiple agents decide to use the same definition of a term (for example by referencing the same term in the same ontology), they can communicate more effectively. Usage of the same term, now with the same meaning, helps improve consistency across applications.

Content search and context search are other typical uses of ontologies. In content search, search engines use background knowledge bases to enhance queries and thus improve results. When the background knowledge bases contain term definitions, semantic query engines may be able to retrieve answers that are inferred by the query, no longer restricting the search to exact user-provided terms. Search engines can go beyond statistical clustering methods, which while effective, have limitations largely associated with training data sets. In context search, search engines may consider the user's context when processing a search. For example, a search engine may utilize a user's geographic location as well as known preferences when retrieving answers. Information about geographic location and preferences may be encoded in background ontologies.

Ontologies describing domain knowledge, user preferences, and problem areas are often used in creating agents with reasoning capabilities. These ontologies are often used to establish a common vocabulary among multiple agents. Personal agents' learning capabilities are also important, as such capabilities can increase the agents' level of autonomy (e.g., the Cognitive Assistant that Learns and Organizes (CALO, 2008). Personal agents can act alone or communicate with others in order to accomplish their task; in these cases, ontologies describing communications protocols are also necessary.

Explanation Issues

Given these Semantic Web application features which impact the need for explanation, we identify a set of criteria for analyzing the required explanations. These criteria include such issues as whether explanations are expected to be consumed by humans or machine agents; varying characteristics of these agents; and the resulting types of explanations that should be provided.

Explanation Types

System transparency allows users to see how answers are generated and how processes within and among agents have evolved to support answer generation. Transparency allows users to access lineage information that often appears hidden in the complex Semantic Web network. Note that explanations should be viewed as a web of interconnected objects recording source information, source assertions and assumptions, intermediate results, and final results instead of as a single "flat" annotation. Results from Semantic Web applications may be derived from a series of information manipulation steps, each of which applies a primitive information manipulation operation, e.g., an inference or extraction rule, on some antecedents and produces a conclusion. Note that an information manipulation step may be any kind of inference and is not limited to those that are used in sound and complete reasoners. Thus this representation can handle statistical methods, standard logical inference, or even non-logical information transformation methods. A justification may be viewed as a transaction log of information manipulation steps. When a user requests a detailed explanation of what has been done or what services have been called, it is important to be able to present an explanation based on this justification. These transaction logs may be quite detailed, so it is also important to be able to provide explanations that are abstractions of these logs.

Another kind of explanation can be obtained from provenance metadata that contains annotations concerning information sources, (e.g., when, from where, and by whom the data was obtained). Provenance metadata connects statements in a knowledge base to the statement sources such as web pages and publications, including annotations about data collection or extraction methods. Criticality of provenance is evident. Users demand detailed provenance metadata before they will accept and believe answers (e.g., Cowell, et al, 2006; Del Rio and Pinheiro da Silva, 2007). In some settings such where an initial evaluation of usefulness is made, provenance metadata (e.g., source, recency, and authoritativeness) is the only information that users need.

Trust in the Semantic Web is another subject of growing importance in the explanation context. Trust representation, computation, combination, presentation, and visualization present issues of increasing importance for Semantic Web applications, particularly in settings that include large decentralized communities such as online social networks (e.g., McGuinness, et. al, 2006).

Human or Machine Consumption

Semantic Web applications typically require explanation for both human and machine consumption. Software agents require representation of justifications, provenance and trust in a standard format in order to enable interoperability. An interoperable justification specification can be used to generate explanations of an agent's reasoning process as well as of the sources used by the agent during the problem solving process. Explanations aimed at either humans or software agents can be generated from the internal justification, provenance, and trust representations. When the explanations are aimed at humans, the explanations must also include human computer interface (HCI) considerations. For instance, the display of an explanation may take into consideration the level of expertise of the user, e.g., expert or non-expert, as well as the context of the problem (e.g., Del Rio and Pinheiro da Silva, 2007a). HCI researchers have approached the explanation problem by proposing intelligent question-answering systems (e.g., Maybury, 2003), intelligent help systems (e.g., Lieberman and Kumar, 2005), and adaptive interfaces (e.g., Wagner and Lieberman, 2003).

Visualization Capabilities

Explanations can be viewed as Semantic Web metadata representing how results were obtained. In distributed settings such as the Web, representation interoperability is paramount. A variety of "user friendly" rendering and delivery modes are required to present information to different types of users in varying contexts. As explanations may need to be delivered to users with a variety of skill levels, visual representation must be flexible, manageable, extensible, and interoperable. Additionally, corresponding presentation modes need to be customizable and context-dependent, and need to provide options for abstract summaries, detailed views, and interactive follow-up support. We consider several possible presentation modes. Implemented interfaces for each of these views can be seen in McGuinness, et al, 2006.

Global View. The entire process of explanation may be presented via a graphical display of a justification graph. The idea is to provide a view of the global structure of the reasoning process used by a question answering system. Common issues include how portions of information composing the explanation will be presented (for example, whether they are displayed in an English translation of the justification encoding, or in the reasoner's native language); or whether to restrict the depth and width of the explanation graph (e.g., with using notions such as lens magnitude and width options in the Inference Web browser). A useful feature in these kinds of views is to provide clickable hot links to enable access to additional information.

Focused View. Merely providing tools for browsing an execution trace is not adequate for most users. It is necessary to provide tools for visualizing the explanations at different levels of granularity and focus, for instance, to focus on one step of the justification, and to display that step using a natural language template style for presentation. Further focus on explanations can be provided by suggested context-appropriate follow up questions.

Filtered View. Alternative options may also be chosen, such as seeing only the assertions (ground facts) upon which a given result depended; only the sources used for ground assertions; or only the assumptions upon which the result depended. Another possible view is the collection of sources contributing information used to derive the result. Some users are willing to assume that the reasoning is correct, and as long as only reliable and recent knowledge sources are used, they are willing to believe the result. Initially, these users may not want to view all the details of the information manipulations (but they do want the option of asking follow-up questions when necessary).

Abstraction View. Machine-generated justifications are typically characterized by their complexity and richness of details that may not be relevant or interesting to most users. Filtering explanation information and providing only one type of information (for example, only showing the information sources) are some of the strategies used to deal with the large volume of data in justifications. These strategies translate the detailed explanation into a more abstract and understandable one.

In fact, this diversity of presentation styles is critical for broad acceptance of SW results. As we have interviewed users both in user studies (e.g., Cowell, et al, 2006; Del Rio and Pinheiro da Silva, 2007; Glass, et al., 2008) and in ad hoc requirements gathering, it was consistently true that broad user communities require focus on different types of explanation information and on different explanation formats. For any user segment that prefers a detailed trace-based view, there is another complementary and balancing user segment that requires an extensively filtered view. This finding results in the design and development of the trace-based browser, the explainer with inference step focus, multiple filtered follow-up views, and a discourse-style presentation component.

Explanation Issues vs. Semantic Web Application Characteristics

Having independently considered facets of both complex Semantic Web contexts and requirements for successful explanations, we now address how these issues relate to each other, providing requirements for explaining a broader range of SW applications.

Explanation and Collaboration

Trust and reputation are important issues in the context of collaborative applications and have been studied in the context of traditional wikis like Wikipedia (e.g., McGuinness, Zeng et al., 2006). The advent of semantic wikis introduces new concerns and requirements in terms of explanation. Autonomy among SW agents is continuously increasing, and if users are expected to believe answers from these applications, SW applications must support explanations. This requirement becomes even more important when SW applications collaborate to generate complex results.

As personal agents mature and assume more autonomous control of their users' activities, it becomes more critical that these agents can explain the way they solve problems on behalf of humans. The agents must be able to tell the user why they are performing actions, what they are doing, and they must be able to do so in a trustable manner. Justifications and task processing explanations are essential to allow personal agents to achieve their acceptance goals. In addition, the learning skill presented by some personal agents amplifies the need for explanation since it introduces a degree of variability resulting from learning results. Justifications concerning agent's internal reasoning for learning new knowledge as well as explanations concerning usage of knowledge sources are examples of what must be explained.

Distributed reasoning requires explanation capabilities to help users understanding the flow of information between the different agents involved in a problem solving process. These capabilities also allow users to understand the process taken by the distributed problem solvers. Additionally, provenance explanations are of interest since users might want to know information about each one of the learners and problem solvers used, as well as wanting to know information about each source of information that was used. Issues of trust and reputation are particularly likely to modify user's trust in agents' answers.

Explanation and Autonomy

In applications for which the degree of autonomy is low (for instance, a Google-based search query), no explicit explanation is provided. One could assume that aspects of explanatory material are implicitly embedded in the answers. In such settings, the user needs to have enough information to understand the context of the answers (e.g., the links selected by the query engine represent an information retrieval response to the query, and the answers include links to the sites containing the information). It is assumed that explaining why a search engine has selected a set of links is implicitly understood by the user (for instance, the search engine considers the provided answers to be the best responses, with some suitable definition of best which may rely on reverse citations, recency, etc.). The existence of a ranking mechanism is fundamental for the success of the interaction process because query reformulation depends on that ability. Understanding the process that led the search engine to provide an answer to a query facilitates the process of query refinement.

Even applications with low degrees of autonomy may experience demand from users for some forms of explanation. Users may want to know how a search engine got its answers, for example, if the answers were selected using certain purchased keywords or other advertising promotions, or if the answers depended on out-of-date source material. The information needs to be presented in an understandable manner, for instance, by displaying answers using purchased keywords in a different style.

Justifications become even more important in applications with higher degrees of autonomy. Autonomous agents can follow complex inference process, and justifications are an important tool for them to provide understandable information to end users.

Explanations and Ontologies

Ontologies can be used effectively to support explanations for a wide array of applications, ranging from relatively simple search applications to complex autonomous problem solving. For example, consider a contextual database search agent which considers user preferences when answering queries. Explanations of why a given solution was provided in a given context are particularly important when the solution does not match the user's specified preferences. Similarly, explanations are important when a particular contextual query results in different answers in different contexts (for example, when answers are dependent on the user's geographic location).

INFERENCE WEB: AN ONTOLOGY-ENHANCED INFRASTRUCTURE SUPPORTING EXPLANATIONS

We now explore Inference Web in the context of addressing the problem of providing explanations to justify the results and behaviors of Semantic Web services and applications. IW provides tools and infrastructure for building, maintaining, presenting, exchanging, combining, annotating, filtering, comparing, and rendering information manipulation traces, i.e., justifications. IW services are used by agents to publish justifications and explanations for their results that can be accessible digitally – on the web, on a local file system, or distributed across digital stores. Justification data and explanations derived from justifications are encoded using terms defined by the Proof Markup Language (PML) justification, provenance, and trust ontologies. The PML ontologies are specified in OWL and are easily integrated with Semantic Web applications. The ontologies include terms such as sources, inference rules, inference steps, and conclusions as explained later.

PML is an on-going, long-term effort with several goals and contributions to explaining Semantic Web application results and behaviors. Our earlier version of PML focused on explaining results generated by hybrid web-based reasoning systems, such as the question answering systems of DARPA's High Performance Knowledge Base program and its subsequent Rapid Knowledge Formation program. The requirements obtained for this initial explanation phase were similar to explanation requirements gathered for expert systems where knowledge bases were generated from reliable source information and using trained experts. Information in these systems was assumed to be reliable and recent. Thus, agent users only needed explanations about information manipulation steps, i.e. how the results were derived in a step by step manner from the original knowledge base via inference. In this setting, explanations concerning information sources used to derive results were not required.

As automated systems become more hybrid and include more diverse components, more information sources are used and thus users are seldom in a position to assume that all information is reliable and current. In addition to information manipulation, users may need explanations about information provenance. Under certain circumstances, such as intelligence settings that motivated DTO's Novel Intelligence for Massive Data program, provenance concerns often dwarfed all others when explanations were required (Cowell, et. al., 2006).

As automated systems begin to exploit more collaborative settings and input may come from many unknown authoring sources, notions of trust and reputation may become more critical. Meta information may be associated with authoring sources such as "I trust Joe's recommendations" or "I trust population data in the CIA World Factbook"). In these situations the meta-information may be user authored. In other settings, trust or reputation information may be calculated using techniques such as link analysis or revision analysis (Zeng, et.al. 2006).

Our goal is to go beyond explanation for traditional knowledge-based systems, and instead address explanation needs in a wide range of situations. We have settings where three different aspects of explanation sometimes dominate to the point that the other aspects are of secondary consideration. We thus took on a rationalization and redesign of our original representation Interlingua so that it could be modular. We can now support applications that only desire to focus on provenance (initially or permanently ignoring issues related to information manipulation and trust.). While these applications may later expand to include those concerns, they need not import ontologies with terms defined for those situations.

Using PML

To illustrate how PML supports explanation generation, we use a simple wine agent scenario. While this example is intentionally oversimplified, it does contain the question answering and explanation requirements in much more complicated examples. We have implemented a wine agent (Hsu, McGuinness, 2003) that suggests descriptions of wines to go with foods. The agent uses PML as its explanation interlingua, and a theorem prover capable of understanding and reasoning with OWL and outputting PML (Fikes, et. al., 2003)). The agent is capable of making wine recommendations to coordinate with meal courses (such as "Tony's specialty"). Before customers choose to follow the agent's recommendation, they may be interested in knowing a description of Tony's specialty, so that they can evaluate if the suggested wine pairing meets their desires. In this scenario, they would find that Tony's specialty is a shellfish dish and the wine agent suggests some white wines as potential matches. The user may want to know how the description of the matching wine was produced, and if the wine agent used other sources of information, such as commercial online wine web sites or hand built backend databases.

In some intelligence settings, e.g., (Cowell, et. al., 2006, Murdock, et. al., 2006), users often want to ask questions about what sources were relied on to obtain an answer. In some military settings, e.g., (Myers, et. al., 2007), users often want to ask what the system is doing, why it has not completed something, and what learned information was leveraged to obtain an answer. In other settings, such as collaborative social networks, users may be interested in either reputation as calculated by populations or trust as stated and stored by users, e.g., (McGuinness, et. al., 2006b). These setting are further elaborated in the following section.

Our PML explanation ontologies include primitive concepts and relations for representing knowledge provenance. Our original version of PML (Pinheiro da Silva et al., 2003) provided a single integrated ontology for use in representing information manipulation activities, the extended version of PML (called PML 2) improves the original version by modularizing the ontologies and refining and expanding the ontology vocabulary. This also broadens the reach covering a wider spectrum of applications for the intelligence, defense, and scientific communities. The modularization serves to separate descriptive metadata from the association metadata to reduce the cost of maintaining and using each module. The vocabulary expansion refines the definition and description structure of existing PML concepts; and it also adds several new primitive concepts to enrich expressiveness. For example, instead of simply serializing a piece of information into a text string, PML uses the concept of information as the universal reference to any piece of data, and enables explicit annotation (for instance, of format, language, and character encoding) about the string that serializes the piece of information.

PML provides vocabulary for three types of explanation metadata:

- The provenance ontology (also known as PML-P) focuses on annotating identified-things (and in particular, sources such as organization, person, agent, services) useful for providing lineage.
- The justification ontology (also known as PML-J) focuses on explaining dependencies among identified-things including how one identified-thing (e.g., information) is derived from other identified-things (e.g. information, services, agents).
- The trust relation ontology (also known as PML-T) focuses on representing and explaining belief assertions.

Provenance Ontology

The goal of the provenance ontology (also called PML-P[a]) is to annotate the provenance of information, e.g., which sources were used, who encoded the information, etc. The foundational concept in PML-P is *IdentifiedThing*. An instance of IdentifiedThing refers to an entity in the real world, and its properties annotate its metadata such as name, description, creation date-time, authors, and owner. PML-P includes two key subclasses of IdentifiedThing motivated by knowledge provenance representational concerns: *Information* and *Source*.

The concept Information supports references to information at various levels of granularity and structure. It can be used to encode, for example, a formula in logical languages or a natural language text string. PML-P users can simply use the value of information's *hasRawString* property to store and access the content of the referred information as a string. They may optionally annotate additional processing and presentation instructions using PML-P properties such as *hasLanguage*, *hasFormat*, *hasReferenceUsage* and *hasPrettyNameMappingList*. Besides providing representational primitives for use in encoding information content as a string, PML-P also includes primitives supporting access to externally referenced content via *hasUrl*, which links to an online document, or *hasInfoSourceUsage*, which records when, where and by whom the information was obtained. This concept allows users to assign an URI reference to information. The example below shows that the content of a piece of information (identified by #info1) is encoded in the Knowledge Interchange Format (KIF) language and is formatted as a text string. The second example below shows that the content of information (identified by #info_doc1) can be indirectly obtained from the specified URL, which also is written in KIF language.

```
<pmlp:Information
rdf:about=”#info1”>

<pmlp:hasRawString>(type Tonys-
Specialty SHELLFISH)

h</pmlp:hasRawString>
<pmlp:hasLanguage rdf:resource=
“http://inferenceweb.stanford.
edu/registry/LG/KIF.owl#KIF”
/>  <pmlp:hasFormat>text</
pmlp:hasFormat>  </
pmlp:Information>

<pmlp:Information
rdf:about=”#info_doc1”>

<pmlp:hasURL>http://iw.stanford.
edu/ksl/registry/storage/
documents/tonys_fact.kif/
pmlp:hasURL>   <pmlp:hasLanguage
rdf:resource= “http://infer-
```

```
enceweb.stanford.edu/regis-
try/LG/KIF.owl#KIF" />  </
pmlp:Information>
```

The concept source refers to an information container, and it is often used to refer to all the information from the container. A source could be a document, an agent, or a web page, and PML-P provides a simple but extensible taxonomy of sources. The Inference Web Registry (McGuinness and Pinheiro da Silva, 2003) provides a public repository for registered users to pre-register metadata about sources so as to better reuse such metadata. Our current approach, however, does not demand a centralized or virtual distributed registry; rather, it depends on a search component that finds online PML data and provides search service for users' inquiry.

```
<pmlp:Document rdf:about="#STE">

<pmlp:hasContent
rdf:resource="#info_doc1"/>

</pmlp:Document>
```

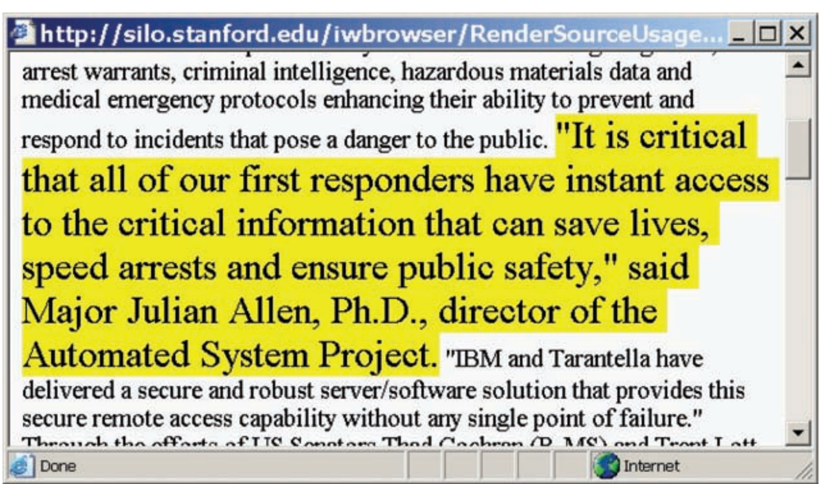
In particular, PML-P provides options for encoding finer grained references to a span of a text through its *DocumentFragmentByOffset* concept. This is a sub-class of Source and *DocumentFragment*. The example below shows how the offset information about #ST can be used to highlight the corresponding span of text (see Figure 1).

This type of encoding was used extensively in our applications that used text analytic components to generate structured text from unstructured input as explained below.

```
<pmlp:DocumentFragmentByOffset
rdf:about="#ST">

<pmlp:hasDocument
rdf:resource="#STE"/>

<pmlp:hasFromOffset>62</
pmlp:hasFromOffset>

<pmlp:hasToOffset>92</
pmlp:hasToOffset> </
pmlp:DocumentFragmentByOffset>
```

As our work evolved, a number of our applications demanded more focus on provenance. We became increasingly aware of the importance of capturing information about the dependency

Figure 1. Raw text fragment with highlighted segment used by text analytics components and represented in PML 2



between information and sources, i.e. when and how a piece of information was obtained from a source. PML 2 has a more sophisticated notion of *SourceUsage*. The encoding below simply shows how PML represents date information identifying when a source identified by #ST was used.

```
<pmlp:SourceUsage
rdf:about="#usage1">

<pmlp:hasUsageDateTime>2005-
10-17T10:30:00Z</
pmlp:hasUsageDateTime>

<pmlp:hasSource
rdf:resource="#ST"/>

</pmlp:SourceUsage>
```

Besides the above concepts, PML-P also defines concepts such as *Language*, *Inference-Rule*, and *PrettyNameMapping*, which are used to represent metadata for application processing or presentation instructions.

Justification Ontology

The goal of the justification ontology is to provide concepts and relations used to encode traces of process executions used to derive a conclusion. A justification requires concepts for representing conclusions, and information manipulation steps used to transform/derive conclusions from other conclusions, e.g., step antecedents.

A *NodeSet* includes structure for representing a conclusion and a set of alternative information manipulation steps also called *InferenceSteps*. Each InferenceStep associated with a NodeSet provides an alternative justification for the NodeSet's conclusion. The term NodeSet is chosen because it captures the notion that the NodeSet concept can be used to encode a set of nodes from one or many proof trees deriving the same conclusion. The URI of a NodeSet is its unique identifier, and every NodeSet has exactly one URI.

The term inference in InferenceStep refers to a generalized information manipulation step, so it could be a standard logical step of inference, an information extraction step, a simple computation process step, or an assertion of a fact or assumption. It could also be a complex process such as a web service or application functionality that may not necessarily be describable in terms of more atomic processes. InferenceStep properties include *hasInferenceEngine* (the agent who ran this step), *hasInferenceRule* (the operation taken in this step), *hasSourceUsage*, *hasAntecedentList* (the input of this step), and others.

PML2 supports encodings for several typical types of justifications for a conclusion. Three justification examples are as follows:

An unproved conclusion or goal. A NodeSet without any InferenceStep can be explained as an inference goal that still needs to be proved. Unproved conclusions happen when input information encoded in PML2 is provided to an agent.

```
<pmlj:NodeSet
rdf:about="#answer1">
<pmlp:hasConclusion
rdf:resource = "#info1" />  </
pmlp:hasConclusion>

</pmlj:NodeSet>
```

Assumption. The conclusion was directly asserted by an agent as an assumption. In this case, the conclusion is asserted by a source instead of being derived from antecedent information.

Direct assertion. The conclusion can be directly asserted by the inference engine. In this case, the conclusion is not derived from any antecedent information. Moreover, direct assertion allows agents to specify source usage. The following example shows that "'(type TonysSpecialty SHELLFISH)' has been directly asserted in Stanford's Tony's Specialty Example as a span of text between byte offset 62 and byte offset 92 as of 10:30 on 2005-10-17"

```
<pmlj:NodeSet
rdf:about="#answer2">
<pmlp:hasConclusion
rdf:resource="#info1" />

<pmlp:isConsequentOf>

<pmlp:InferenceStep
rdf:about="step2">

<pmlp:hasInferenceEngine
rdf:resource= "http://inference-
web.stanford.edu/registry/IE/
JTP.owl#JTP" />

<pmlp:hasInferenceRule
rdf:resource= "http://inference-
web.stanford.edu/registry/DPR/
Told.owl#Told" />

<pmlp:hasSourceUsage
rdf:resource="#usage1" />

</pmlp:InferenceStep>

</pmlp:isConsequentOf>

</pmlj:NodeSet>
```

TOOLS FOR MANIPULATING EXPLANATION IN PML

To address the need to support multiple visualization modes for explanation, Inference Web provides rich presentation options for browsing justification traces, including a directed acyclic graph (DAG) view that shows the global justification structure, a collection of hyperlinked web pages that allows step-by-step navigation, a filtered view that displays only certain parts of the trace, an abstracted view, and a discourse view (in either list form or dialogue form) that answers follow-up questions.

Global View. Figure 2 depicts a screen shot from the IW browser in which the *Dag* proof style has been selected to show the global structure of the reasoning process. The sentence format can be displayed in (limited) English or in the reasoner's native language, and the depth and width of the tree can be restricted using the lens magnitude and lens width options, respectively. The user may ask for additional information by clicking hot links. The three small panes show the results of asking for follow-up information about an inference rule, an inference engine, and the variable bindings for a rule application.

Focused View. In Figure 3a, our explainer interface includes an option to focus on one step of the trace and display it using an English template style for presentation. The follow-up action pull down menu then helps the user to ask a number of context-dependent follow-up questions.

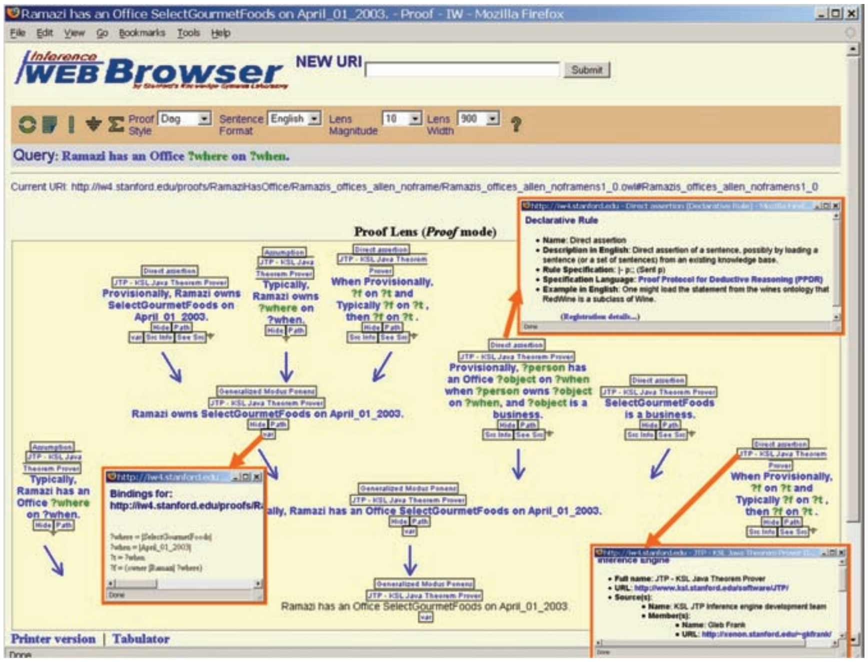
Filtered View. Figure 3b is the result of the user asking to see the sources.

Abstraction View. Inference Web approaches this issue with two strategies:

- Filter explanation information and only provide one type of information (such as what sources were used). This strategy just hides portions of the explanation and keeps the trace intact.
- Transform the explanation into another form. The IW abstractor component helps users to generate matching patterns to be used to rewrite proof segments producing an abstraction. Using these patterns, IW may provide an initial abstracted view of an explanation and then provide context appropriate follow-up question support.

The IW abstractor consists of an editor that allows users to define patterns that are to be matched against PML proofs. A matching pattern

Figure 2. Trace-oriented explanation with several follow-up question panes

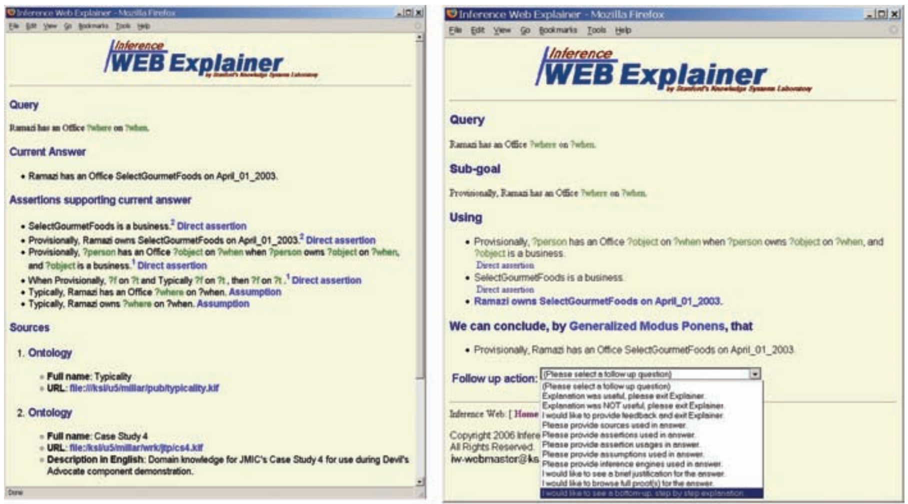


is associated with a rewriting strategy so that when a pattern is matched, the abstractor may use the rewriting strategy to transform the proof (hopefully into something more understandable). An example of how a proof can be abstracted with the use of a generic abstraction pattern is shown in Figure 4. In this case, the reasoner used a number of steps to derive that crab was a subclass of seafood. This portion of the proof is displayed in the *Dag* style in the middle of Figure 4 (inside the blue round-angled box). The user may specify an abstraction rule to reduce the multi-step proof fragment into a one-step proof fragment (class-transitivity inference) on the left side of Figure 4.

We are building up abstraction patterns for domain independent use, e.g. class transitivity as well as for domain-dependent use. It is an ongoing line of research to consider how best to build up a library of abstraction patterns and how to apply them in an efficient manner.

Discourse View. For some types of information manipulation traces, particular aspects or portions of the trace are predictably more relevant to users than others. Additionally, the context and user model can often be used to select and combine these portions of the trace, along with suggestions of which aspects may be important for follow-up queries. Particularly for these types of traces, IW provides a *discourse view*, which selects trace portions and presents them in simple natural language sentences. In this interaction mode, the full details of the inference rules and node structure are kept hidden from the user. Individual nodes, provenance information, and metadata associated with those nodes, are used as input for various explanation strategies, which select just the information relevant to the user's request and provide context-sensitive templates for displaying that information in dialogue form. This same information is also used to generate suggested follow-up queries for the user, including requests for additional detail, clarifying questions about the explanation that has been provided, and questions essentially requesting that an alternate explanation strategy be used.

Figure 3. (a) step-by-step view focusing on one step using an English template, and list of follow-up actions; (b) filtered view displaying supporting assertions and sources



CASE STUDIES: PML IN ACTION

We will describe four applications that are using the IW framework and PML for explaining semantic information and behavior. We selected four applications that can be categorized differently following the conceptual framework.

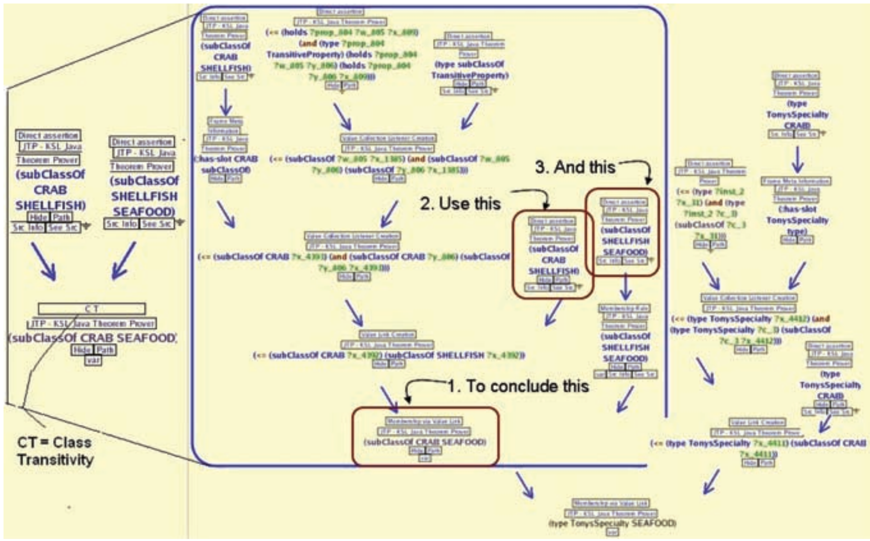
Cognitive Personal Assistants: CALO Example

IW and PML have been used by a DARPA-sponsored cognitive agent system called CALO that can be told what to do, reason with available knowledge, learn from experience, explain its recommendations, and respond robustly to surprise. The cognitive agent's actions are supported by justifications that are used to derive and present understandable explanations to end-users. These justifications reflect both how the actions support various user goals, and how the particular actions chosen by the agent were guided by the state of the world. More specifically, our approach to PML task justification breaks down the justification of a question about a particular task *T* into three complementary strategies, described here using terminology from SPARK (Morley & Myers 2004), the task engine used by CALO:

- **Relevance:** Demonstrate that fulfilling *T* will further one of the agent's high-level goals, which the user already knows about and accepts
- **Applicability:** Demonstrate that the conditions necessary to start *T* were met at the time *T* started (possibly including the conditions that led *T* to be preferred over alternative tasks)
- **Termination:** Demonstrate whether one or more of the conditions necessary to terminate *T* has not been met.

This three-strategy approach contrasts with previous approaches to explanation, most of which dealt with explaining inference (Scott et al. 1984, Wick & Thompson 1992). Previous approaches generally have not dealt with termination issues, and they also generally have not distinguished between relevance and applicability conditions. These are critical aspects of task processing and thus are important new issues for explanation.

Figure 4. Example of an abstraction of a piece of a proof



Behavior Justification in PML

In CALO context, PML documents contain encodings of *behavior justifications* using PML node sets. A task execution justification is always a justification of why an agent is executing a given task *T*. The final conclusion of the justification is a sentence in first order logic saying that *T* is currently being executed. There are three antecedents for this final conclusion, corresponding to the three strategies discussed above. Each antecedent is supported by a justification fragment based on additional introspective predicates.

It is important to note that all the task processing justifications share a common structure that is rich enough to encode provenance information needed to answer the explanation requests we have identified so far. By inspecting the execution state via introspective predicates, explanation components can gather enough provenance information to support a wide range of explanations.

Text Analytic Information Manipulations: KANI Example

KANI (Knowledge Associates for Novel Intelligence) (Welty, et. al., 2005, Murdock, et. al., 2006) is a DTO-sponsored intelligence analyst hybrid system that combines large scale information extraction with knowledge representation. In this section we focus on the relevance of provenance to support explanations of hybrid systems utilizing statistical and deductive inference.

In this setting, we can view all information manipulation steps in a PML justification as a kind of inference. We then generated a taxonomy of text analytic processes and tasks that can be viewed as inferences. The taxonomy was motivated by the need to describe and explain the dominant extraction tasks in UIMA[b], without overloading the system with more information than would be useful. One key was to generate a taxonomy that is adequate to accurately describe extraction task functionalities and simultaneously abstract enough to be able to hide details of the tasks from end users. Another key was to support explanations to end users of the integrated system, not authors of software components debugging their products.

We divided text extraction into three primitive areas: annotation, co-reference, and integration. We describe each briefly. Annotation tasks make assertions about spans of text that recognize a type or argument. Annotation inferences include:

1. **Entity recognition:** Determines that some span of text refers to an entity of a specified type. For example, a component could take the sentence "Tony Gradgrind is the owner of Tony's Foods" (the restaurant serving Tony's Specialty) and conclude that characters 0 to 14 of that sentence refer to some entity of type Person.
2. **Relation recognition:** Assigns a relation type to a span (e.g., a sentence describes a relation of type Owner).
3. **Relation annotation argument identification:** Determines and assigns values to the roles of a relation (e.g., a particular person is a participant in a given ownership relation instance).

Co-reference inferences utilize annotation inferences and further identify that multiple text spans actually refer to the same entity or relation.

1. **Entity identification:** Determines that a set of entity annotations refer to a particular instance.
2. **Relation identification:** Determines that a set of relation annotations refer to a particular relation instance.
3. **Extracted entity classification:** Determines that a particular co-referenced entity has a particular type. (e.g., the type of the entity referred to by "Gradgrind" is Person).
4. **Knowledge integration** inferences include mapping inferences providing access to provenance.
5. **Entity mapping:** Determines that an entity instance in the KB is derived from a set of entities and relation instances.
6. **Relation mapping:** Determines that a relationship in the target KB is derived from a set of entity and relation instances.
7. **Target entity classification:** Determines that an entity instance is an instance of an entity type in the target ontology.

We have registered these inferences in the IW registry and we use these information manipulation steps to explain all of the UIMA components used in our prototype system, which provides intelligence analyst support for analyzing documents and evaluating results of text statements.

Text Analytic Manipulation Descriptions

We use our taxonomy of text analytic manipulations in declarative descriptions encoding what was done to generate the extracted knowledge bases. UIMA generates a large extracted knowledge database containing its conclusions. We needed to take that as input (potentially augmented) and generate interoperable proof descriptions (a PML document) as an output.

The software component that produces PML documents for UIMA-based analysis processes begins with a specified result from a specified Extended Knowledge Database (EKDB) (e.g., TonyGradgrind is the Owner of TonysFoods). It follows the links in the EKDB from that conclusion back to the intermediate results and raw input that led to it. From these intermediate results, it is able to produce inference steps encoded in PML that refer to the corresponding tasks in the taxonomy. For example, if the EKDB records that characters 0 to 14 of some sentence were labeled as a Person and that this labeling was identified as specifying an occurrence of TonyGradgrind then the component would create an Entity Recognition inference step in PML for that labeling as well as coreference step for the result that the labeling is an occurrence of TonyGradgrind.

Transparent Accountable Data Mining: TAMI Example

TAMI (Weitzner, et. al., 2006) is an NSF-sponsored privacy-preserving system funded in the Cybertrust program. The idea is to provide transparency into the usage of data that has been collected, so that people may be able to see how data that has been collected about them has been used. In any accountable system, explanations are essential for providing transparency into the usage of information along with claims of compliance with privacy policies.

Usage policies are encoded concerning which organizations can use information for particular purposes. (The project specifically aims at usage instead of collection policies, so it is only use and reuse that is a topic for explanations). A transaction log is collected, which encodes data transfer information concerning transfers, policies, purposes, and organizations. Reasoning engines are used that evaluate the validity of transfer actions based on the encoded policies. These engines are instrumented to encode justifications for their determinations in PML, so that explanations can be provided about justified or unjustified transfers.

This system can be leveraged in a number of examples. One use case is in the explanation of justified or unjustified arrests. It is possible that data collected in compliance with rules for a particular purpose by an authorized agency may be reused to support a number of other conclusions. One prototype demonstration system in TAMI looks at arrests and then checks to see if they are justified according to their appropriate or inappropriate reuse of data that has been collected. Inference Web can then be used to explain why the system has determined that an arrest is legally justified or unjustified.

Integrated Learning Systems: GILA Example

GILA (Generalized Integrated Learning Architecture) is a DARPA-sponsored intelligent agent that integrates the results of multiple learners to provide intelligent assistant services. The initial domain is airspace control order deconfliction. GILA uses multiple independent learning components, a meta reasoning executive, and other components to make recommendations about ways to resolve conflicts in an existing airspace control order. In order to be operational, it must be able to explain its recommendations to end users and auditors. In addition, the explanations may be uses by learners and the meta executive to choose appropriate recommendations and assign credit and blame.

DISCUSSION

Explanation has been an active line of research since at least the days of expert systems, where explanation research largely focused on explaining rule-based systems. Today, explanation in rule systems is once again a research. Rule systems are now being integrated into hybrid settings, and now explanation must be done on both the rule components and the setting in which conclusions from those rule components are integrated and used. Also, theorem proving systems, such as Description Logic Reasoners, historically integrated explanation capabilities after usage increased and broadened. Early description logics that were broadly used, such as CLASSIC and LOOM provided some notion of explanation (e.g., McGuinness, 1996) in either insight into a trace or a proof theoretic-based approach to explanation. More recent explanation demands have inspired current generation tableaux-based DL reasoners to include some notion of explanation focusing on provenance, axiom usage, and clash detection (e.g., Parsia, et al, 2005, Plessers and Troyer, 2006). While all of these efforts are useful and important, today's explanation systems need to handle a much broader range of question answering styles and thus demand much more versatility and interoperability for their explanation infrastructure. Simultaneously, the infrastructure needs to be modular so that users with limited scope can support their applications without the burden of extra (unwanted) overhead. In our research on explaining provenance, we have recently modularized our explanation interlingua and the supporting background ontologies so that clients *only* interested in explaining provenance may use our infrastructure with the freedom of importing only the required modules.

Explanation requirements often arise in many settings that do not simply use standard deductive reasoning components. Our work, for example, has taken us into the realm of explaining text analytic components and a wide range of machine learning components. As a result, we have explored and are continuing to explore representation, manipulation, and presentation support for explaining systems that may use statistical, incomplete, and/or uncertain reasoning paradigms. Explanation research has also branched out into settings such as collaborative social networks, and we have engaged in research aimed particularly at explaining systems embedded in or leveraging large distributed communities. In many of the more recent research areas, we have found many requirements concerning trust, ranging from trust calculation to trust propagation, as well as presentation issues related to filtering by trust.

One relatively active area of provenance explanation is in the field of scientific applications. Increasingly, virtual collections of scientific data are being enabled by semantic technology (e.g., Virtual Observatories such as the Virtual Solar Terrestrial Observatory (McGuinness, et al, 2007). Such repositories are much more likely to be usable and to be used when provenance is maintained and available concerning where the data came from. More recently, there has been emphasis on additionally explaining the workflow from which it was produced. Thus, there is an emerging emphasis on explaining scientific provenance and workflow.

FUTURE RESEARCH DIRECTIONS

We have active research plans in a number of areas related to explanation.

1. **Learning.** Increasingly hybrid systems are depending on individual or multiple learning components to provide either ground facts or sometimes procedures. We are currently working multiple learning component authors to provide explanation components for learned information and learned procedures.
2. **Provenance.** The importance of provenance seems to be growing in many fields and we are focusing on providing relatively lightweight explanation solutions for provenance. We are also exploring special purpose needs of interdisciplinary scientific applications with respect to provenance.
3. **Trust.** Our current trust model is relatively simplistic and we are investigating ways of providing more representational primitives, methods for automatically suggesting trust ratings, and methods for intelligently combining and explaining combined trust values.
4. **Evaluation.** We have developed a PML validator that checks to see if an encoding is valid PML. We are extending that to provide an ontology evaluation module that not only checks for syntactic and semantic correct-

ness, but also reviews (and explains findings concerning) ontology modeling styles.

CONCLUSION

In this chapter, we have explored the growing field of explanation. We noted that as applications become more autonomous, complex, collaborative, and interconnected, the need for explanation expands. We presented a modular interlingua capable of representing explanations that focus on provenance, justifications, and trust. We also presented the Inference Web infrastructure for manipulating explanations in a wide range of application settings. We provided examples in a diverse set of domains showing different settings where explanations are required, and then described how Inference Web and PML are being used to meet these needs. We also presented a number of different presentation paradigms for explanations.

ACKNOWLEDGMENT

We have benefited greatly by working with a number of excellent collaborators including Bill Murdock, Chris Welty, and Dave Ferrucci from IBM and Andrew Cowell, Dave Thurman, and colleagues from Battelle on NIMD, Michael Wolverton, Karen Myers, David Morley from SRI on CALO, Danny Weitzner, Tim Berners-Lee, Lalana Kagal, Chris Hanson, Gerry Sussman, Hal Abelson, Dan Connolly, Sandro Hawke, Kay Waterman, and colleagues from MIT on TAMI, and a large contingent of collaborators on GILA including Ken Whitebread, Martin Hofmann, Phil DiBona, Steve Wilder from Lockheed Martin and collaborators in multiple universities on the project related to learners and meta reasoning. This work has been partially supported by contract numbers: 55-00680, PO TT0687676, 5710001895-2, 2003*H278000*000, HR0011-05-0019, and F30602-00-1-0579.

REFERENCES

Barber, K., & Martin, C. (1999, May 1). Agent autonomy: Specification, measurement, and ydnamic adjustment. In *Proceedings of the Autonomy Control Software Workshop at Autonomous Agents 1999* (Agents ’99), 8-15. Seattle,WA.

CALO. (2008). http://www.ai.sri.com/project/CALO

Cowell, A. J., McGuinness, D. L., Varley, C. F., & Thurman, D. A. (2006). Knowledge-worker requirements for next generation query answering and explanation systems. In the *Proceedings of the Workshop on Intelligent User Interfaces for Intelligence Analysis, International Conference on Intelligent User Interfaces* (IUI 2006), Sydney, Australia.

Del Rio, N., & Pinheiro da Silva, P. (2007, June). Identifying and explaining map imperfections through knowledge provenance visualization. *Technical report UTEP-CS-07-43a*, University of Texas at El Paso, El Paso, TX.

Del Rio, N., & Pinheiro da Silva, P. (2007a, November 26-28). Probe-It! Visualization support for provenance. In *Proceedings of the Third International Symposium on Visual Computing (ISVC 2007),* Lake Tahoe, NV/CA.

Dent, L., Boticario, J., McDermott, J., et al. (1992). A personal learning apprentice. In *Proceedings of the 10 National Conference on Artificial Intelligence*, San Jose, California: AAAI Press, pp. 96-103.

Dzbor, M., Motta, E., & Domingue, J. B. (2004). Opening up magpie via semantic services. In McIlraith et al. (eds), The Semantic Web - ISWC 2004, *Third International Semantic WebConference*. Hiroshima, Japan. *Lecture Notes in Computer Science*, 3298,Springer-Verlag.

Glass, A., McGuinness, D., & Wolverton, M. (2008). Toward establishing trrust in adaptive agents. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI'08)*, Gran Canaria, Spain. Also, KSL Technical Report KSL-07-04.

Guha, R., & McCool, R. (2003). Tap: A Semantic Web platform. *Computer Networks*, *42*(5), 557–577. doi:10.1016/S1389-1286(03)00225-1

Huynh, D., Mazzocchi, S., & Karger, D. (2005, November 6-10). Piggy bank: Experience the Semantic Web inside your Web browser. In Gil et al. (eds), *The Semantic Web - ISWC 2005, 4th International Next Generation Semantic Web Applications ISWC 2005*. Galway, Ireland. *Lecture Notes in Computer Science,* 3729 Springer-Verlag.

Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., & Saarela, S. (2005). MuseumFinland - Finnish museums on the Semantic Web. *Journal of Web Semantics*, *3*(2-3), 224. doi:10.1016/j.websem.2005.05.008

Lashkari, Y., Metral, M., & Maes, P. (1994). Collaborative interface agents. In *Proceedings of the 12 National Conference on Artificial Intelligence*. Seattle, WA: AAAI Press, pp. 444-450.

Lieberman, H., & Kumar, A. (2005, September). Providing expert advice by analogy for on-line help, *IEEE/ACM Conference on Web Intelligence & Intelligent Agent Technology*, Compiègne, France.

Lopez, V., Motta, E., & Uren, V. (2006, June 11-14). PowerAqua: Fishing the Semantic Web. In York Sure and John Domingue (eds.), *The Semantic Web: Research and Applications, 3rd European Semantic Web Conference*, *ESWC 2006*, Budva, Montenegro. *Lecture Notes in Computer Science 4011*, Springer, ISBN 3-540-34544-2.

Maes, P. (1994). *Agents that reduce work and information overload communications of the ACM*, *37*(7), 31-40.

Maybury, M. (2003). New directions on question and answering, *AAAI Spring Sysmposium*, TR-SS-03-07, Stanford, CA.

McGuinness, D., Fox, P., Cinquini, L., West, P., Garcia, J., Benedict, J. L., & Middleton, D. (2007a, July 22-26). The virtual solar-terrestrial observatory: A deployed Semantic Web application case study for scientific research. In *proceedings of the Nineteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-07).* Vancouver, BC, Canada.

McGuinness, D. L. (1996). Explaining reasoning in description logics. Ph.D. Thesis, Rutgers University. Technical Report LCSR-TR-277. Rutgers Department of Computer Science Technical Report Series.

McGuinness, D. L., Ding, L., Glass, G., Chang, C., Zeng, H., & Furtado, V. (2006a) Explanation interfaces for the Semantic Web: Issues and models. Presented in the *3rd International Semantic Web User Interaction Workshop (SWUI'06),* Co-located with the *International Semantic Web Conference*, Athens, Georgia, USA.

McGuinness, D. L., Ding, L., Glass, G., Chang, C., Zeng, H., & Furtado, V. (2006a) Explanation interfaces for the Semantic Web: Issues and models. Presented in the *3rd International Semantic Web User Interaction Workshop (SWUI'06),* Co-located with the *International Semantic Web Conference*, Athens, Georgia, USA.

McGuinness, D. L., Ding, L., Pinheiro da Silva, P., & Chang, C. (2007). A modular explanation interlingua. In the *Proceedings of the Explanation-aware Computing Workshop (ExaCt-2007)* co-located with the *Association for the Advancement of Artificial Intelligence*, Vancouver, BC.

McGuinness, D. L., & Pinheiro da Silva, P. (2004, October). Explaining answers from the Semantic Web: The inference Web approach. *Journal of Web Semantics*, *1*(4), 397–413. doi:10.1016/j.websem.2004.06.002

McGuinness, D. L., Zeng, H., Pinheiro da Silva, P., Ding, L., & Narayanan, D. & Bhaowal. M. (2006b, May 22). Investigations into trust for collaborative information repositories: A Wikipedia case study. *WWW2006 Workshop on the Models of Trust for the Web (MTW'06)*, Edinburgh, Scotland.

Morley, D., & Myers, K. (2004). The SPARK agent framework. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS-04)*, New York, NY.

Mota, E., & Sabou, M. (2006). *Next generation Semantic Web applications*, ASWC.

Murdock, J. W., McGuinness, D. L., Pinheiro da Silva, P., Welty, C., & Ferrucci, D. (2006, November 5-9). Explaining conclusions from diverse knowledge sources. In the *Proceedings of the Fifth International Semantic Web Conference*, Athens, Ga.

Parsia, B., Sirin, E., & Kalyanpur, A. (2005) Debugging owl ontologies. In the *Proceedings of the World Wide Web Conference*, pp. 633-640.

Plessers, P., & Troyer, O. D. Resolving inconsistencies in evolving ontologies. In the *Proceedings of the European Semantic Web Conference*, pp. 200-214.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., & Goranov, M. (2003). KIM – A Semantic Annotation Platform. In D. Fensel, K. Sycara, and J. Mylopoulos (eds.), *The Semantic Web - ISWC 2003, Second International Semantic Web Conference. Lecture Notes in Computer Science*, 2870, Springer-Verlag.

Pynadath, D. V., & Tambe, M. (2002). Electric elves: Adjustable autonomy in real-world multi-agent environments. In socially intelligent agents *– Creating relationships with computers and robots*. Kluwer Academic Publishers.

Rhodes, B. J., & Starner, T. (1996). Remembrance agent: A continuously automated information retrieval system. *Proceedings, First international Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology*. London, UK.

Schraefel, M. C., Shadbolt, N. R., Gibbins, N., Glaser, H., & Harris, S. (2004). CS AKTive space: Representing computer science in the Semantic Web. In *Proceedings of the 13th International World Wide Web Conference.*

Sullivan, D., Grosz, B., & Kraus, S. (2000). Intention reconciliation by collaborative agents. In *Proceedings of the Fourth International Conference on Multi-Agent Systems, IEEE Computer Society Press*, Boston, MA.

Wagner, E., & Lieberman, H. (2003, January). End-user debugging for electronic commerce. *ACM Conference on Intelligent User Interfaces*, Miami Beach.

Weitzner, D. J., Abelson, H., Berners-Lee, T., Hanson, C. P., Hendler, J., Kagal, L., et al. (2006). Transparent accountable inferencing for privacy risk management. *Proceedings of AAAI Spring Symposium on The Semantic Web meets eGovernment*. Stanford University, USA: AAAI Press Also available as MIT CSAIL Technical Report-2006-007 and Stanford KSL Technical Report KSL-06-03.

Welty, C., Murdock, J. W., Pinheiro da Silva, P., McGuinness, D. L., Ferrucci, D., & Fikes, R. (2005). Tracking information extraction from intelligence documents. In *Proceedings of the 2005 International Conference on Intelligence Analysis (IA 2005)*, McLean, VA, USA.

ADDITIONAL READINGS

Explanation Infrastructure:

Diverse Explanation Presentation Paradigms. Deborah L. McGuinness, Li Ding, Alyssa Glass, Cynthia Chang, Honglei Zeng and Vasco Furtado. Explanation Interfaces for the Semantic Web: Issues and Models. Presented in *the 3rd International Semantic Web User Interaction Workshop(SWUI'06)*, Co-located with the International Semantic Web Conference, Athens, Georgia, USA, November 6, 2006.

(2004, October). Foundational paper: Deborah L. McGuinness and Paulo Pinheiro da Silva. Explaining Answers from the Semantic Web: The Inference Web Approach. *Journal of Web Semantics*, *1*(4), 397–413. doi:10.1016/j.websem.2004.06.002

Explanation Interlingua:

Newest version: McGuinness, D.L.; Ding, L., Pinheiro da Silva, P., and Chang, C. A Modular Explanation Interlingu. Proceedings of the 2007 Workshop on Explanation-aware Computing (ExaCt-2007), Vancouver, Canada, July 22-23, 2007.

(2006). Original version: Paulo Pinheiro da Silva, Deborah L. McGuinness and Richard Fikes. A Proof Markup Language for Semantic Web Services. *Information Systems*. Volume 31 [Previous version, technical report, Knowledge Systems Laboratory, Stanford University.]. *Issues (National Council of State Boards of Nursing (U.S.))*, *4-5*(June-July), 381–395.

Explanation and Trust Requirements Studies:

In Cognitive Assistant Settings. Glass, A.; McGuinness, D.L.; Wolverton, M. Toward Establishing Trust in Adaptive Agents. International Conference on Intelligent User Interfaces (IUI'08), Gran Canaria, Spain, 2008.

In Intelligence Settings. Cowell, A.; McGuinness, D.L.; Varley, C.; Thurman, D. Knowledge-Worker Requirements for Next Generation Query Answering and Explanation Systems. In the Proceedings of the Workshop on Intelligent User Interfaces for Intelligence Analysis, International Conference on Intelligent User Interfaces (IUI 2006), Sydney, Australia. 2006.

Selected Applications

Explaining Data Mining and Data Usage. Weitzner, D.J.; Abelson, H.; Berners-Lee, T.; Hanson, C.P.; Hendler, J.; Kagal, L.; McGuinness, D.L.; Sussman, G.J.; Waterman, K.K. Transparent Accountable Data Mining: New Strategies for Privacy Protection. Proceedings of AAAI Spring Symposium on The Semantic Web meets eGovernment. AAAI Press, Stanford University, Stanford, CA, USA, 2006.

Explaining Intelligence Applications. Christopher Welty, J. William Murdock, Paulo Pinheiro da Silva, Deborah L. McGuinness, David Ferrucci, Richard Fikes. Tracking Information Extraction from Intelligence Documents. In Proceedings of the 2005 International Conference on Intelligence Analysis (IA 2005), McLean, VA, USA, 2-6 May, 2005.

Explaining Task Processing in Learning Settings. McGuinness, D.L.; Glass, A.; Wolverton, M.; Pinheiro da Silva, P. Explaining Task Processing in Cognitive Assistants that Learn. Proceedings of the 20th International FLAIRS Conference (FLAIRS-20), Key West, Florida, May 7-9, 2007.

Explaining Text Analytics. J. William Murdock, Deborah L. McGuinness, Paulo Pinheiro da Silva, Christopher Welty and David Ferrucci. Explaining Conclusions from Diverse Knowledge Sources. The 5th International Semantic Web Conference (ISWC2006), Athens, Georgia, USA, November 5th - 9th, 2006.

Explanation, Trust, and Collaborative Systems:

Deborah, L. McGuinness, Honglei Zeng, Paulo Pinheiro da Silva, Li Ding, Dhyanesh Narayanan, and Mayukh Bhaowal. Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study. WWW2006 Workshop on the Models of Trust for the Web (MTW'06), Edinburgh, Scotland, May 22, 2006.

Victor, P., Cornelis, C., De Cock, M., & Pinheiro da Silva, P. Towards a Provenance-Preserving Trust Model in Agent Networks. Proceeding of the WWW'06 Workshop on Models of Trust for the Web (MTW'06), Edinburgh, Scotland, May 22, 2006.

Victor, P., Cornelis, C., De Cock, M., & Pinheiro da Silva, P. (to appear). Gradual Trust and Distrust in Recommender Systems. *Fuzzy Sets and Systems*.

Zaihrayeu, I., Pinheiro da Silva, P., & McGuinness, D. L. IWTrust: Improving User Trust in Answers from the Web. Proceedings of 3rd International Conference on Trust Management (iTrust2005), Springer, Rocquencourt, France, 2005.

Zeng, H., Alhossaini, M., Ding, L., Fikes, R., & McGuinness, D. L. Computing Trust from Revision History. The 2006 International Conference on Privacy, Security and Trust (PST 2006) Markham, Ontario, Canada October 30 -- November 1, 2006.

ENDNOTES

[a] The OWL encoding of PML-P is available at: http://iw.stanford.edu/2006/06/pml-provenance.owl

[b] http://www.research.ibm.com/UIMA/

This work was previously published in Semantic Web Engineering in the Knowledge Society, edited by J. Cardoso; M. Lytras, pp. 1-24, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 8.3
A New System for the Integration of Medical Imaging Processing Algorithms into a Web Environment

José Antonio Seoane Fernández
Artificial Neural Networks and Adaptative Systems Group, Spain & University of Corunna, Spain

Juan Luis Pérez Ordóñez
Center of Medical Informatics and Radiological Diagnosis, Spain & University of Corunna, Spain

Noha Veiguela Blanco
Artificial Neural Networks and Adaptative Systems Group, Spain & University of Corunna, Spain

Francisco Javier Novóa de Manuel
Center of Medical Informatics and Radiological Diagnosis, Spain & University of Corunna, Spain

Julián Dorado de la Calle
University of A Coruña, Spain

ABSTRACT

This chapter presents an architecture for the integration of various algorithms for digital image processing (DIP) into web-based information systems. The proposed environment provides the development of tools for intensive image processing and their integration into information systems by means of JAVA applets. The functionality of the system is shown through a set of tools for biomedical application. The main feature of this architecture is that it allows the application of various types of image processing, with different computational costs, through a web browser and in a transparent and user-friendly way.

DOI: 10.4018/978-1-60566-002-8.ch022

INTRODUCTION

The rapid advance of the medical imaging field is revolutionizing medicine. Technologies such as computed axial tomography (CT Scan), magnetic resonance imaging (MRI), Helicoidal CT Scan, and the fusion of CT Scan and positron emission tomography (PET), all provide an effective map of the human anatomy in a non-invasive manner.

Clinical practice usually relies on computing techniques to simplify the diagnosis of the medical expert. Medical imaging is not restricted to the visualization of anatomical structures, it is also used for diagnosis, surgical planning, simulation, radiotherapy planning, etc. These applications enable the clinicians to virtually interact with the anatomical structures and as such achieve the knowledge that enhances their performances. All the aforementioned techniques belong to a discipline known as Digital Image Processing (DIP). Traditionally, the medical DIP applications were carried out in expensive work-stations provided by the CT or PET machine supplier. These kinds of applications have certain drawbacks, such as administration and maintenance, which make them unsuitable for some environments.

The current trend in software development is the creation of applications that can be integrated into a Web environment and enjoy advantages such as placement independence, centralized application maintenance, and the use of firewalls without changing the filter rules. Web applications have proliferated due to their rapid learning and easy use as well as their personalization capability that provides a user-friendly interface. This trend towards Web developments is also being introduced into the medical field, to the detriment of the traditional clinical applications. DIP-related applications have high computational costs and therefore hospitals have to invest heavily in computing equipment in order to provide the clinicians with powerful mainframes. At this point, it seems logical to differentiate between algorithms of low and high computational cost.

It should be borne in mind that, as DIP is not a recent discipline, there exist libraries that include different algorithms for digital image processing. Already implemented algorithms should therefore be reused in new developments.

State of the Art

There currently exists a wide range of applications that allow the digital processing of medical images by means of a Web browser. The following list represents applications with two common factors: the DIP is processed at the client and the implemented algorithms tend to have a low computational cost.

- RAIM Java is a DICOM (Digital Imaging and Communication in Medicine) image viewer for biomedical imaging that was developed by the Biomedical Digital Imaging Center of UDIAT-CD S.A. (http://www.cspt.es/webcsptcastella/udiat/default.htm). This viewer was developed with Java technology and can therefore be used in almost any computer and graphic operative system. Since the visual display was conceived as an applet, it has to be executed within a Web browser; this allows the images to be processed in various ways, such as change of visualization window (Window-Level), rotation, scale, etc.
- CHILI: Digital radiology (http://www.chili-radiology.com) is a set of software components oriented towards tele-radiology and PACS (Picture Archiving and Communication Systems). It is a product from CHILI GmbH (Germany) and was developed in cooperation with the German Cancer Research Center and the *Steinbeis Transferzentrum Medizinische Informatik* company. CHILI WEB is one of its products and is composed by the CHILI/Web Server with the CHILI/Web Client. The CHILI/Web Server first receives the images of the modalities through the DICOM protocol and later stores them in a relational data base. The CHILI/WEB Server can work with an existing PACS. The CHILI/

Web Client is a platform-independent program developed with Java.

- RemotEye(http://eng.neologica.it/prodotti/remoteye) is a DICOM display developed by NeoLogica.it. which allows the visualization of DICOM imaging though Internet and offers the possibility of performing certain types of digital image processing such as geometrical transformations, changes of brightness and contrast, etc. It can easily be integrated with the PACS.
- MagicWeb/ACOM.Web (http://www.medical.siemens.com) was developed by Siemens and provides the publication of images and reports for an entire welfare centre with the aim of allowing clinicians to consult from different places. One of its characteristics is the optimization of imaging visualization, since images can be seen with different brightness and contrast levels, augmentation, compression, cinema mode, filters, etc.

On the other hand, there also exists a series of applications that use intensive processing for the performance of different algorithms and rely on complex computing developments, since they can not be executed within a Web browser. The following list represents various developments with complex computational DIP requirements.

- A good option for the development of distributed applications is the CORBA interface for communication among remote objects. This technology allows not only the integration of various languages on different machines but also a total interconnection among all the applications. An example of the use of this technology is the Image Processing Tool (http://imageprocess.sourceforge.net), a distributed processing system based on open-source software that uses a client-server architecture. The main shortage of this software is that, inherently to the use of CORBA, two machines cannot communicate through the Internet due to possible intermediate firewalls.
- The IRMA (image retrieval in medical applications) (http://phobos.imib.rwth-aachen.de/irma/index_en.php) is a distributed system developed by the Computing Department of the University of Aachen (Germany). This system provides not only uniform access to different modalities of medical imaging but also the application of different types of distributed processing algorithms to the images. The proposed architecture has a client-server structure and involves a central database, a task planner, and several processing daemons. The planner distributes the processing load among the different processing daemons installed at processing stations, which provides a low cost and high performance system. The main disadvantage of such a system is the communication between the client and the server on TCP-IP through non-standard protocols, and the non-multiplatform nature of the client.
- The Diamond Eye (Burl et al, 1999) architecture was developed by NASA laboratories and created initially to recover images and use them for data-mining tasks. This architecture involves the use of a Web browser to gain access to images stored in a database through a Java Applet. The architecture enables the client to access and process the images independently from the platform. The data-mining operations with a high computational cost are performed in a network of workstations (NOW) of Sun UltraSparc II. The requirements of the client are handled to the server, which executes in parallel the algorithm at the processing network by means of message-passing libraries [9].
- Distributed architectures are a good alternative to the supercomputers in terms of

processing power when executing high-cost computational tasks. Another architecture based on message passing is the proposal of Li, Veeravalli and Ko, (Li et al, 2003) that uses the PVM (Parallel Virtual Machine) library (http://www.csm.ornl.gov/pvm) to pass communication messages to and from the processors of various workstations. This approach is based on the division of the image into several parts in order to be distributed among the network nodes and processed afterwards; the approach of He, Wu, Liu and Zheng (He et al, 2003) explores the best way to distribute the data among the different processors.

Among all the aforementioned architectures, only he Diamond Eye developed by NASA achieves distributed processing using the Web browsers as front-end by means of Java applets. However, this system is very specific and does not allow the integration of different types of tools. A framework is therefore needed for the integration of heterogeneous tools into an information system that can support high computational costs through the Internet.

PROPOSED ARCHITECTURE

Considering the vast amount of existing medical imaging modalities and their respective processing needs, the proposed architecture had to be independent from both software and hardware and integrate several image processing algorithms. Our architecture provides access to the processing services through a Web interface. The present work proposes a framework for the development of tools that are able to provide a medical team with all the processing power required for high cost visualization, planning, and simulation tasks by means of a remote server. These tasks can be performed independently from a physical location, since only an Internet connection and browser are needed.

This section describes, from two different viewpoints, a solution for the integration of processing algorithms for biomedical digital imaging into Web systems. The first perspective is that of the physical placement of each element, whereas the second perspective focuses on the logical structure of the architecture.

Physical Framework Architecture

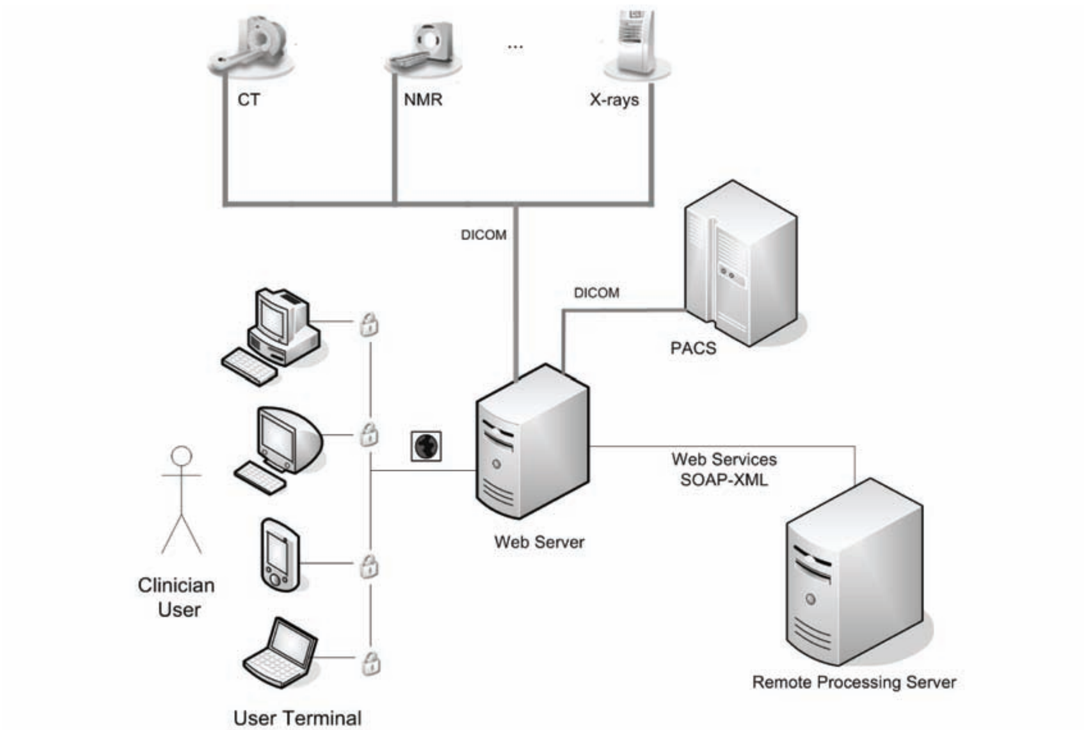
The main parts of the architecture are the following: the user terminal, the web server, and the remote processing server (Figure 1).

The user terminal is located at the Web browser and is integrated in an information system, whose data could be used for digital image processing operations. Once the user has selected the images that will be processed, the tools and the possible environments for digital processing are shown. An applet is executed when selecting a tool; this applet involves a display and a simple processing kernel and shows the images and processing options.

The Web Server is in charge of four fundamental tasks: it has to communicate with the kernel of the display for the transfer of images and other patient-associated data, it communicates the image information system for independently retrieving images and its associated information from various data sources, it performs the digital processing of the images with high computational requirements (remote processing), and it gives access to the different remote processing servers.

The use of remote processing servers offers the following advantages: the distribution of computational loads related to the digital processing, the integration of algorithms developed with different languages, the calculus transparency (how and where) for the developer thanks to Web access, and lastly, a higher processing capability. The architecture provides a uniform access through Web services and has a high scalability, which

Figure 1. Physical framework architecture



means that multiple queries can be received and processed at the cluster and that results can rapidly be returned by means of the distributed computation paradigm (Foster, 1999).

Logical Framework Architecture

The tools and components are defined so as to obtain a homogeneous and configurable interface.

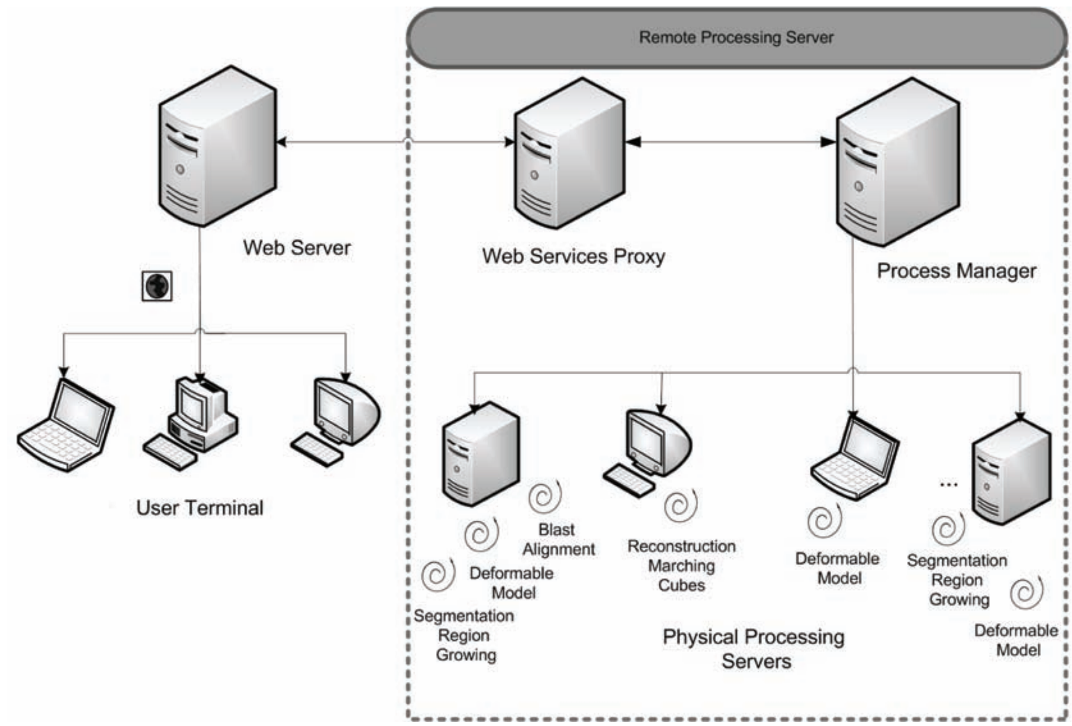
Tools: A tool comprises the set of specific functions that are needed to work with a certain type of image, i.e for each possible environment for digital image processing. For instance, the functions for working with hemodynamics imaging are not the same than the ones used for oncology imaging; however, there might be common functions for both tools. Each tool has one or more components.

Components: Each component performs a specific function in the system, ranging from information management to digital image processing, with optional forms for the modification of processing parameters that are visualized at the applet. An example of a component is the function that changes the brightness and contrast of an image. The components have two different parts attending the place where they are executed: one of them inside the applet of the client's Web browser, and the other inside the application server, where it processes and manages the remote information. When building these two parts, interfaces must be defined for the good functioning of the parts.

Each component usually works with one or more DIP, such as for instance the segmentation of a hemodynamics image. Depending on the complexity of such algorithms, they will be located inside the applet, inside the application server or inside the remote processing server.

Local processing: If the processing algorithms have simple computational requirements, their execution is performed at the kernel of the client

Figure 2. Remote processing server detail



applet, located at the user terminal. This type of processing is known as local processing.

Application server processing: This second type of processing is especially indicated for algorithms of medium complexity, which will be executed by the application server. Each component is divided into two parts: the local part at the user browser and the remote part at the application server. The algorithm parameters are established at the local part of the component. When executing the processing, a request is sent to the application server through the tool, the kernel of the display, and the applet. After arriving to the servlet -where the component that has to be executed is identified- the remote component of the algorithm will be invoked to execute the algorithm and return the result.

Remote server processing: Finally, the remote processing servers are used for load distribution and to provide support to the algorithms that are developed with various programming languages. In this case, the remote part of the component -located at the application server- sends the request to the Web service, where the algorithm is executed. The result is sent back to the remote part and, from there, to the display. The different parts of the remote processing server can be observed in Figure 2.

A remote processing server model with three parts is built as follows: the first part, the web services proxy, collects the external processing web requests that come from the information system. These requests are sent to another server where the processing manager is located and where they are directed in accordance with both, the needed processing type and the original object that made the request to the web service. The third part is represented by the physical remote processing servers, which are able to perform one or more different intensive processing algorithms. The remote object manager instantiates a remote processing object on the subsequent server in charge of executing the needed processing algorithm, keeping the result at the local memory of the machine where the algorithm is being executed, and returning that result when needed.

There are two reasons for dividing the remote processing server into three parts. Firstly, processing is usually quite costly and may collapse the reception, so we must separate the reception of the web server queries from the server that performs the intensive processing Secondly, the objects that are instantiated on the web server have no state, since every time a remote call is made from a web server, an object is instantiated; this object disappears as soon as it returns (web services behaviour). This fact reduces the functionality, because it does not allow the interaction with the remote object, but also acts as a call to a function that replies with a result. This architecture provides the web services with a virtual state that allows us to carry out several calls on the same object to assign parameters, modify them during the execution, and test the processing status.

The web services proxy implements only one processing method. The parameters of this method are an identifier of the object that invokes the web service, the operation type realised at the processing, and a list of parameters needed for performing that operation. This method invokes an operation from the remote object hosted at the processing manager that controls the processing objects. Distributing the processing manager among several machines provides a higher failure tolerance and a simpler maintenance, since each machine carries out only one function.

The object manager instantiates the processing objects that are needed to execute the processing algorithm required by the web services proxy; it also distributes the load among the different physical servers and it manages the security of the transactions. At the first call of an algorithm, the object manager instantiates a new object on one of the processing servers and then all the processing calls are directed towards that object.

Finally, the processing servers execute one or more processing algorithms. The remote object manager determines which physical processing server is more suitable for executing the new processing, according to the algorithms that every server has and the processor load at that moment.

IMPLEMENTATION

This section describes the technologies that were used to implement the architecture. The user terminal was developed with Java applets because of their multiplatform nature and also because it is a common solution for providing a web environment with processing capacities (Laird et al, 2003). In this part, where the local processing takes place, the algorithms must be developed with Java and with the option of using the advanced API for image processing of JAI (Java Advanced Imaging) (http://java.sun.com/products/java-media/jai).

The application server was developed with the J2EE platform, because it is portable, scalable, and safe, and because it uses open standards. The implementation that was used for J2EE is the Apache Tomcat 5.0.28. Since in this part the remote processing can be performed, we can integrate the developed processing algorithms that use either Java or C. Link libraries known as Java Native Interface (JNI) are used for algorithms developed with C.

The communication with the remote processing servers was established by means of web services, whose hardware and software are platform independent, which widens the range of implementation possibilities for the various remote processing servers. The algorithms developed with Java or C could be integrated using the same scheme for the web server. If the algorithm is developed with another language, or if the used technology is different, the algorithm's developing framework will only have to be supported by web services.

The validity of the remote processing system is proved by implementing two remote processing servers, one developed with Apache AXIS Java libraries, the other with a .NET Framework. In this case, the previously proposed remote processing server was implemented. An ideal processing scenario would imply three or more machines for the processing server: the first machine could have Internet Information Services (IIS), to collect the requests from the application server, or a stand-alone application that may require this type of processing by means of web services; the second machine would host the processing manager, and the remaining machines would house the specific processing servers. The technology used to inter-communicate the proxy with the object manager and the latter with the processing servers is .NET Remoting over TCP. In all these servers .NET framework must be installed. The use of .NET consists in integrating the developed algorithms by means of any .NET-supported language such as Visual Basic, C#, Visual C++, etc.

The standard network topology implies one or more J2EE application servers that respond to the requests of the client's applets. The remote processing server will need one or more machines with Internet Information Services (IIS) to serve the web services requests. These are the only machines that would have to be provided with an external connection. Security will be implemented by means of SSL certificates provided by the IIS (Seely, 2002), which warrant the Internet transaction security. On the one hand, the LAN network, which contains the IIS, has a machine provided with a *ProcessManager* for client control and the processing types of the system. The related data of previous versions will remain in a data base and as such allow several machines to provide this service. On the other hand, there exists a network of physical processing servers where every machine can execute one or more algorithms of intensive processing.

Integration into an Information System

The architecture was designed for to be easily integrated into an image information system. In this case, it was integrated into a biomedical imaging information system known as Web-SMIIS that is part of the SMIIS project (Pereira, 2003). The SMIIS is essentially a PACS for the retrieval of information from any modality that might support the DICOM standard, and the subsequent storage of that information into a data base for consultation. The SMIIS was developed by the Centre for Medical Informatics and Radiological Diagnosis (IMEDIR) of the University of A Coruña (Spain).

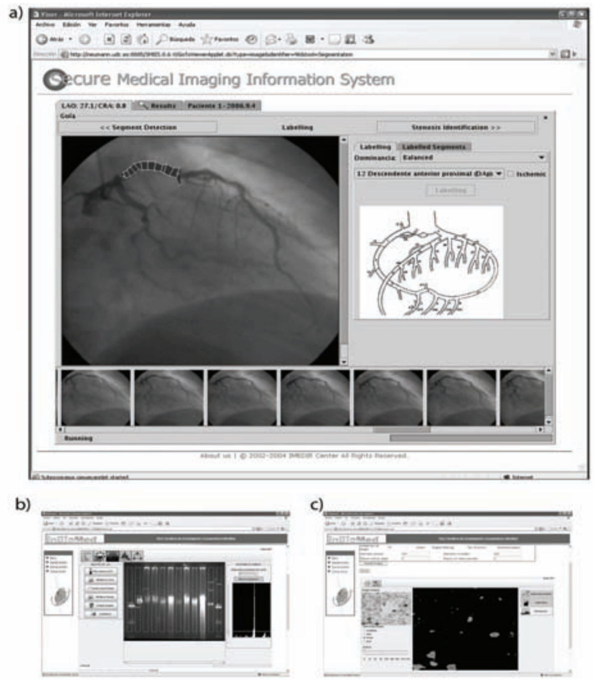
The integration between the two systems is focused on a point where the architecture has to obtain the images. This place depends on each information system. In the Web-SMIIS, the facade "*QueryFacadeDelegateFactory*" is the access to the SMIIS.

Also, this architecture could be integrated into any PACS DICOM thanks to a module that allows the extension of the system's functionality in order to achieve its compatibility with the DICOM standard and, more specifically, with the storage query/retrieve services.

RESULTS

This architecture was implemented in two Spanish hospitals, the *Complejo Hospitalario Universitario Juan Canalejo* and the *Instituto Médico Quirurgico San Rafael* in A Coruña. The first is the most important hospital of the city and a centre of reference for a geographic area of more than 500.000 people. The second is one of the most important private hospitals of A Coruña. The processing architecture is currently being validated by a group of 7 medical experts of the Hemodynamics Service of the Juan Canalejo Hospital, which carries out 3.000 catheterisms

Figure 3. (a) Stenosis detection tool (b) electrophoresis gel analysis tool (c) inmunohistochemical count tool



each year. The processing architecture is currently available in the radiology service of the San Rafael hospital.

Validation protocols are long and complex. The hemodynamics unit of the Juan Canalejo Hospital validated a total of 63 angiographs in groups of 2 to 3 members; the San Rafael Hospital validated 213 radiology studies (TC, MRI, medical ultrasonography, etc.) in groups of 3 members each.

We developed two tools for the semi-automatic analysis of angiographs in collaboration with the clinicians of the Hemodynamics Unit of Juan Canalejo in order to establish a score pattern for coronary stenosis. The first tool uses for segmentation tracking techniques (O'Brien and Ezquerra, 1994) (NEzquerra et al, 1998), the second tool uses morphological operators and techniques of region growth (Haris, 1999)(Kirbas, 2003). Figure 3a shows a screen capture of this tool.

This architecture was also used for tool development in the cooperative thematic research network INBIOMED (Pérez et al, 2005). With the collaboration of the Pharmacology Group from the University of Santiago de Compostela (Spain), a tool was developed for the analysis of protein and DNA electrophoresis gel (Figures 3b). It allows the application of low cost image preprocessing algorithms such as low-pass filtering, image-enhance filter, image rotation operator, etc., as well as the semiautomatic detection and analysis of the lanes and bands that contain this kind of image in order to obtain the protein weight represented by the relative position of the bands in each lane. Another tool that was developed by the INBIOMED network provides a fast and user-friendly cell count for immunohistochemical images using adaptative thresholding (Chow and Kaneko, 1972)(Chan et al, 1998) (figure 3c), applying not only low-cost image processing to the client side, but also algorithms with a high computational cost that are executed at a separate

Table 1. Hospital and validation data

	CHU Juan Canalejo	IMQ San Rafael
Beds	1.430	148
Total annual external consultations	672.295	44.310
Emergencies	179.101	7.949
Total annual admissions	44.814	4.605
Interventions	29.924	6.045
Implantation and Validation	Hemodynamics Service	Radiology Service
Service Volume	2.896 annual catheterisms	34.700 radiological tests
Validation team	7 persons	3 persons
Number of Tests	63 Validated studies	231 Validated studies

processing server that communicates with the system through Web services. (Table 1)

Finally, we developed a tool for the segmentation and subsequent reconstruction of three-dimensional medical image volumes. It uses algorithms with a high computational cost such as 3D region growth to segmentate the desired regions and marching cubes to reconstruct and visualize these regions in a Java applet on an internet browser. Both techniques are implemented in the remote processing server because of their high computational cost.

Table 2 shows the test results for the proposed architecture. Three high computational cost tests focused on development: cell counting in high resolution color images, segmentation of a skull in a CT set of 106 512x512 slices, using region growth, and reconstruction of this segmentation with marching cubes algorithms (Lorensen and Cline, 1987) (the processing time does not count image loading time). The tests were carried out according to three approaches: an applet, a stand-alone application, and the proposed architecture (10 processing units), with one, five, and twenty requests. A normal PC (Pentium IV 3GHz with 1GB RAM) was used in the one-machine test.

CONCLUSION

We propose a system that allows medical experts to analyze patient data through a web browser by integrating processing algorithms of biomedical imaging into web-based information systems.

The design is based on the use of both design and architectural patterns and provides the

Table 2. Test results

	Cell counting	Region growing	Marching cubes
Applet	40.8 s	(not enough memory)	(not enough memory)
Stand-alone application	12.6 s	19.3 s	5.6 s
Proposed architecture 1 request	15.3 s	22.1 s	6.8 s
Proposed architecture 5 request	22.1 s (4.4 s/req)	26.8 s (5.3 s/ req)	8.2 s (1.64 s/ req)
Proposed architecture 20 request	47.6 s (2.3 s/ req)	56.1 s (2.8 s/ req)	18.2 s (0.91 s/ req)

simple integration of this kind of architecture with information systems carried out under a "model-view-controller" paradigm such as SMIIS. Also, the design of new processing tools based on the "view-tool-component" model allows the developers of digital processing algorithms to easily integrate them and obtain new components and tools.

The remote processing server allows the incorporation of the architecture into new types of distributed systems, since it provides the developer with a façade that hides the placement, the functioning, and the situation of each specific remote processing server; it also provides a common interface and only one entry point for all the processing algorithms implemented in the system.

Another advantage of the remote processing server architecture is its communication with the application server by means of web services, because this implies remote processing through the Internet without any port redirection problems, firewalls, etc.; it also means that the application server may be located in a different place from the network of processing servers.

From the point of view of the user, we have achieved transparency with regard to the execution of the process. The clinician carries out data analysis and processing, regardless of whether the processing is simple and can be done at the local machine with the Web browser, or it is expensive and implies the joint work of several equipments.

ACKNOWLEDGMENT

This work was supported in part by the Spanish Ministry of Education and Science (Ref TIC2003-07593, TIN2006-13274), Ministry of the Treasury (PROLIT/SP1.E.194103), the INBIOMED network (Ref G03/160) financed by the Carlos III Health Institute (FIS-PI061524), and grants from the General Directorate of Research of the Xunta de Galicia (Ref. PGIDIT03-PXIC10504PN PGIDIT04-PXIC10503PN, PGIDIT04-PXIC10504PN). The work of Juan L. Pérez is supported by an FPI grant (Ref. BES-2006-13535) from the Spanish Ministry of Education and Science.

REFERENCES

Burl, M. C., Fowlkes, C., Roden, J., Stecher, A., & Mukhtar, S. (1999). Diamond Eye: A distributed architecture for image data mining. Presented in *SPIE Thirteenth International Symposium On Aerospace/Defense Sensing, Simulation and Controls*.

Chan, F. H. Y., Lam, F. K., & Hui, Z. (1998). Adaptive thresholding by variational method. *IEEE Transactions on Image Processing*, *7*(6), 468–473. doi:10.1109/83.661196

Chow, C. K., & Kaneko, R. (1972). Automatic boundary detection of the left ventricle from cineangiograms. *Computers and Biomedical Research, an International Journal*, *5*, 443–452. doi:10.1016/0010-4809(72)90070-5

Ezquerra, N., Capell, S., Klein, L., & Duijves, P. (1998). Model-guided labelling of coronary structure. *IEEE Transactions on Medical Imaging*, *17*(3), 429–441. doi:10.1109/42.712132

Foster, I., & Kesselman, C. (1999). *The grid blueprint for a new computing infrastructure*. Morgan Kaufmann Publishers.

Haris, K., Efstratiadis, S., Maglaveras, N., Pappas, C., Goruassas, H., & Loruidas, G. (1999). Model-based morphological segmentation and labelling of coronary angiograms. *IEEE Transactions on Medical Imaging*, *18*(10), 1003–1015. doi:10.1109/42.811312

He, X., Wu, Q., Liu, D., & Zheng, L. H. (2003). A Distributed and Parallel Edge Detection Scheme within Spiral Architecture. In *Proceedings of Visualization, Imaging and Image Processing 393* (pp 303-311).

Kirbas C. and Queq F. (2003). *A review of vassel extraction techniques and algorithms*

Laird S.P., Wonng J.S.K., Schaller W.J. Ericksin B.J., de Groen P.C. (2003). Desing and implementation of an internet-based medical image viewing system. *The Journal of Systems and Software. Col. 66*, 167-181.

Li, S. L., Veeravalli, B., & Ko, C. (2003). Distributed image processing on a network of workstations. *International Journal of Computers and Applications*, *25*, 2.

O'Brien, J., & Ezquerra, N. (1994), Automated segmentation of coronary vessels in angiographic image sequences utilizing temporal, spatial and structural constraints. *Visualization in Biomedical Computing*.

Pereira, J. (2003), Nuevo Enfoque en el desarrolo de un sistema de información de imágenes Médicas. *Una herramienta de apoyo a la toma de decisión clínica. Doctoral Thesis*. Universidade da Coruña.

Perez, D., Crespo, J., Anguita, A., Pérez, J. L., Dorado, J., & Bueno, G. (2005). Biomedical Image Processing Integration Through INBIOMED: A Web Services-Based Platform. Lecture Notes in Bioinformatics. *LNBI*, *3745*, 34–43.

Seely, S. (2002), *Seguridad http y servicios web en ASP.NET*. Retrieved from http://www.microsoft.com/spanish/msdn/articulos/archivo/111002/voices/httpsecurity.asp

KEY TERMS

Client-Server Architecture: The most fundamental distributed architecture. A client-server architecture is simply a client process that request services from a server process.

Common Object Request Broker Architecture (CORBA): Is a distributed object architecture defined by the Object Management Group. This architecture provides an interface that invokes other CORBA objects across a network.

Digital Imaging and Communication in Medicine (DICOM): A standard format and application protocol developed by the NEMA (National Electrical Manufacturers Association) to communicate systems over TCP. This protocol allows the integration of PACS, workstations, and TC, MNR, and other image scanners.

J2EE: Java 2 Enterprise Edition is a widely used platform for server programming in Java language, used to deploy distributed multi-tier Java software running in an application server. It is also known as Java EE in versions 1.5 and following.

Java Advanced Image (JAI): Is an image-processing toolbox, developed by Sun, that provides an object-oriented interface for the support of high-level programming models that allow images to be easily manipulated in Java applications.

Marching Cubes: A computer graphics algorithm for the extraction of a polygonal mesh of a set of volumetric data.

.Net Remoting: A distributed-object architecture by Microsoft to develop distributed applications over Microsoft platforms.

Network of Workstations (NoW): Is a computer network that connects several computer workstations with special software forming a cluster, to act as a distributed supercomputer on a building-wide scale.

Parallel Virtual Machine (PVM): Is a software package that allows a heterogeneous collection of computers hooked together by a network to be used as a single large parallel computer.

Picture Archive and Communication System (PACS): A storage system composed by different computers and networks dedicated to the storage and retrieval of medical images.

Platform-Independent: An application that can be run on many different server platforms, e.g. Java.

Region Growing: A segmentation technique based on the similarity of adjacent pixels. A region is started with a single pixel (seed pixel) and the adjacent pixels are added to the current region if they are similar to the region.

SSL: A secure socket layer is a secure protocol that provides secure communication over the internet based on cryptographic techniques. At present it is also known as TLS (Transport Layer Security)

Transmission Control Protocol/Internet protocol (TCP-IP): Is the basic family of network protocols for the Internet.

Web Services: Software system designed to support interoperable machine to machine interaction over web. It uses SOAP protocol and XML messages to receive request and offer responses.

This work was previously published in Handbook of Research on Distributed Medical Informatics and E-Health, edited by A. Lazakidou; K. Siassiakos , pp. 885-890, copyright 2009 by Medical Information Science Reference (an imprint of IGI Global).

Chapter 8.4
Social Media Marketing:
Web X.0 of Opportunities

Lemi Baruh
Kadir Has University, Turkey

ABSTRACT

In recent years social media applications, which enable consumers to contribute to the world of online content, have grown in popularity. However, this growth is yet to be transformed into a sustainable commercial model. Starting with a brief overview of existing online advertising models, this chapter discusses the opportunities available for advertisers trying to reach consumers through social media. The chapter focuses on viral marketing as a viable option for marketers, reviews recent viral marketing campaigns, and offers recommendations for a successful implementation of social media marketing. In conclusion, the author examines future trends regarding the utilization of the emerging Semantic Web in marketing online.

DOI: 10.4018/978-1-60566-368-5.ch004

INTRODUCTION

The brief history of the World Wide Web is filled with stories of unprecedented commercial success as well as shattered dreams of hopeful online entrepreneurs. It should not be surprising that, just as their predecessors, Web 2.0 and social media also bring about important questions regarding their sustainability. On the one hand, since 2006, social media sites have been growing in number and popularity (Boyd & Ellison, 2007). For example, according to comScore, a leading Internet information provider, as of December 2007 Facebook had close to 98 million unique visitors, and Fox Interactive Media, including MySpace, had more than 150 million. Similarly, recent years have seen a phenomenal growth in the popularity of weblogs (blogs): in 2007 every day, 175,000 new blogs were added to an estimated 67 million blogs that were already up and running (as cited in Rappaport, 2007). On the other hand, skeptics voice their belief that social media, despite their current popularity, may not have the staying power (“MySpace, Facebook and Other Social Networking Sites,” 2006).

An important component of skeptics’ concerns about the sustainability of social media pertains to the fact that there are no agreed upon ways of monetizing the rising popularity of social media (Allison, 2007; Hall, 2007). Perhaps, the most telling example of this problem is Facebook. Despite having a market value of around $15 billion, Facebook’s 2007 revenue was $150 million (McCarthy, 2008) – a considerably small share of the $21 billion online advertising industry. Then, the question of whether social media will be more than just a fad boils down to advertisers’ ability to utilize the unique opportunities presented by social media. Although advertisers and social media entrepreneurs are yet to agree on a marketing model for social media, recent discussions point to several important requirements that a successful model should accommodate. Given the decentralized architecture of the Internet in general and social media in particular, a central tenet of these recent debates concerns the relative merits of more conventional advertising methods and word of mouth (or word of “mouse”) based marketing approaches that cede control to the consumers.

In the light of these debates, this chapter will start by summarizing online advertising methods. After this brief summary, the chapter will focus on the opportunities and challenges for online marketers that are brought about by the development of social media. Finally, the chapter will discuss viral marketing and integrated marketing communication principles to provide a roadmap for realizing the financial and marketing potential of Web 2.0.

BACKGROUND

Online Advertising

In its most traditional sense, *advertising* is defined as a paid form of communication appearing in media, usually with the purpose of reaching a large number of potential customers. Since 1993, when CERN announced that the World Wide Web would be available to anyone free of charge, advertisers experimented with different methods of reaching consumers online. Unsurprisingly, the first reaction of advertisers was to treat the World Wide Web as a natural extension of traditional media, such as newspapers and television. And, just as in conventional mass media, early online advertising methods, such as banners, pop-ups and interstitials, were characterized by intrusiveness and adoption of a one-way stimulus-response model within which information flows from the advertiser to the customer (McCoy, Everard, Polak, & Galletta, 2007; Rappaport, 2007).

However, even in the early years of online advertising, signs of what was to come in interactive marketing were revealed. Shortly after banners became a popular online advertising method in 1994, keyword-activated “smart banners” were introduced. What set smart banners apart from their predecessors was that the contents of the banners were personalized in response to the search words entered by the users. As such, smart banners were one of the first examples of how content variability in new media (Manovich, 2001) can be utilized to customize information to consumers’ needs (Faber, Lee & Nana 2007).

Customization and Message Congruence in Interactive Media

As noted by several researchers, content variability and the consequent ability to customize content according to the needs of the consumer are made possible by the interactive capabilities of new media (Baruh, 2007; Faber et al., 2007).

Two important characteristics of interactive media are the ability to facilitate a two-way flow of communication and the related ability to track and store every bit of information about how consumers use a system (McCallister & Turow, 2002). Real-time information about how consumers use a medium, especially when combined with other data through data mining, enables marketers to extract profiles about individuals that can then be used to tailor messages and products.

The ultimate aim of this process is to target different consumer groups with specific messages that can tie a product to their informational needs, lifestyles or predispositions. Extant literature on online targeting suggests that consumers will be more receptive to messages that are tailored as such (McCoy et al. 2007; Robinson, Wyscocka, & Hand, 2007). To a large extent, this higher receptivity is the result of being able to promote the "right" product, at the "right" time and place and the "right" tone. A case in point that supports these research findings is the success of Google's AdWords, which accounts for 40% of online advertising spending. The premise of AdWords is that the marketers can reach motivated consumers by providing them with contextual advertising messages congruent with their online keyword searches (and presumably, their interests). Similarly, a widely known feature of online vendors such as Amazon.com is their customized product recommendation systems. The recommendation system these online vendors utilize is based on a data mining system called *market-basket analysis* (also called *association discovery*). The premise of this system is that the marketer can create cross-selling opportunities by identifying the product types that a customer would be interested in (e.g., microwave popcorn) on the basis of other products that he or she has already purchased or is purchasing (e.g., a DVD movie). As such, what the market-basket analysis algorithm does is to identify product clusters that are purchased together or sequentially using the product purchasing history of customers whose tastes are similar to a specific customer.

ONE STEP FURTHER: WEB 2.0 OF OPPORTUNITIES

Customization and Data from Social Media

As can be inferred from the discussion above, collecting information about consumers is an important prerequisite of customizing advertising messages in accordance with the informational needs and lifestyles of consumers. Certainly, data about individuals' online media consumption and purchasing behavior, especially when combined with other sources of data such as credit history, provide marketers with an unprecedented capability to not only determine which customers to target (and avoid) but also when and how to target them.

Within this context, social network sites, such as Facebook, MySpace or LinkedIn, have a potential to extend what is already a large pool of data about consumers. Such social network sites are designed to allow users to create personal profiles and connect with other users, friends or strangers. And through the creation and perennial updating of their profiles, users of social network sites actively participate in the dissemination of information about themselves (Andrejevic, 2007; Solove, 2007). The types of information users of social network sites disclose include: information about their hobbies, interests, likes and dislikes, whom they associate with, a dinner they had a couple of days ago and, sometimes, disturbingly private details about their social and sexual lives. Blogs, another highly popular form of social media, are no different from social network websites. As Solove (2007) points out, any topic, any issue and any personal experience are fair game for more than 60 million bloggers around the world.

The massive quantities of data that social media users reveal online are not left untapped by media companies and marketers. For example, MySpace has recently begun an effort to mine data from its more than 100 million users in order to better target advertising messages. Named as MySpace HyperTargetting, the system initially began mining data about general interest categories, such as sports and gaming, and is now further dividing interests into thousands of subcategories (Morrisey, 2007).

The Community Touch

An important point to note with respect to the types of data available in social media is that the digital goldmine of information is not simply a more detailed version of data collected about consumers' interests and behaviors in other forms of interactive media. Rather, in social media, the available data contain unprecedented details about the network affinities of users. The data about the network affinities of users can be utilized at two levels. First, through the "tell me about your friends and I'll tell you about yourself" principle, marketers can make further refinements to consumers' profiles based on the interests shared by members of the communities they belong to. Secondly, information about the communities that an individual belongs to can be used to identify the paths through which they can be reached.

Recent marketing techniques devised by online vendors and social media outlets illustrate how information about social affinities can be used to reach consumers. For example, Amazon.com's iLike application, a music service that markets new music and concerts to interested listeners, works by scanning the music libraries of its subscribers. The service connects like-minded listeners and promotes new music to users through add-ons such as Facebook's iLike widget. Similarly, Facebook's own Beacon platform tracks purchases Facebook users make on partnering online vendors and then informs users' networks about the recent purchase (Klaassen & Creamer, 2007; Thompson, 2007; Tsai, 2008). In addition to leveraging existing social networks to disseminate marketing messages, some software applications, for example, Stealth Friend Finder automatically generate massive and targeted Facebook Friend Requests to directly connect with the consumers.

Web 2.0 of Opportunities: Viral Marketing in Social Media

These examples of social targeting pinpoint the direction that marketing in social media can take. Rather than being an advertising distribution system, Beacon is a viral marketing tool that lets community members know what their co-members have purchased. In other words, with the Beacon system, the consumer, through the publication of his/her purchasing decision, assumes the role of an influencer. Subramani and Rajagopalan (2003) suggest that consumers may assume such a role either passively or actively. In the passive form, the consumer spreads the word simply by using or purchasing a product (as is the case when an e-mail from a Blackberry user contains a message saying the e-mail was sent using a Blackberry account). On the other hand, active viral marketing requires that consumers participate in the message dissemination process by contacting other potential customers (Clow & Baack, 2007).

An important criticism of passive viral marketing systems in social media is that they fail to utilize an important characteristic of Web 2.0 in general and social media in particular. Instead of being a passive consumer of readily available content, Web 2.0 users are participants in both the creation and dissemination of content. Accordingly, despite utilizing social graphs to target messages more effectively, the "your friend just bought this book from Amazon.com" message is nevertheless an advertising method that affords the consumer very little power as a potential source of influence (Anderson, 2006; Windley, 2007).

Considered from this perspective, a more appropriate way of utilizing the viral potential of social media users is to invite them to actively participate in promoting the product. First, existing research shows that close to a quarter of users of online social networks, such as Facebook, use these sites to influence other users (Webb, 2007). Second, as evidenced by Facebook users' negative reaction to Beacon, social network sites are relatively intimate environments and advertising intrusion (especially given an overall mistrust for advertising messages) is not welcome (Clemons, Barnett, & Appadurai, 2007; Gillin, 2007; Hall, 2007). In contrast, 94% of online social network users find product recommendations from friends to be at least very worthwhile to listen to (MacKeltworth, 2007). This finding is not surprising since recommendations coming from friends, family members, or colleagues are more likely to be trustworthy and relevant to one's needs (Clemons et al., 2007). In fact, according to a recent survey, along with the reputation of the manufacturer, recommendations from friends and family members are the biggest factor that influences purchasing decisions made by individuals (Klaassen & Creamer, 2007). Third, thanks to synchronous connections between multiple users, a computer-mediated word of mouth can reach a larger number of people than word of mouth in the brick and mortar world.

As briefly mentioned before, in addition to these three important advantages of inviting social media users to actively disseminate marketing messages, product information, or recommendations, social media also provide marketers with an unprecedented capability to identify the individuals who would be the best candidates in a social network to act as viral marketers. Domingos (2005) suggests that in addition to actually liking a product, a suitable viral marketing candidate should have high connectivity and should be powerful as a source of influence. Using social network analyses (Hanneman & Riddle, 2005; Scott, 2000; Wasserman & Faust, 1995), data regarding personal affiliations and social network memberships can be utilized to identify opinion leaders ("hubs") who are central to and powerful in a given network.

Recently, there have been several attempts to apply social network analysis to social media to identify social network influencers. For example, Spertus, Sahami, and Büyükkökten (2005) used network data from Orkut.com to identify members who could be used to recommend new communities to users. Similarly, in a study of Flickr and Yahoo360 networks, Kumar, Novak and Tomkins (2006) were able to distinguish between passive users and active inviters that contributed to the extension of the network. And recently, MySpace announced that it is constructing an "influencer" option for advertisers who could be interested in reaching users with active and large networks. To identify potential influencers, MySpace will use data regarding users' group memberships and interests, their friends' interests, level of network activity in a given network and other factors (Morrissey, 2007).

The Integrated Marketing Communications Perspective

In 1976, Wayne DeLozier suggested that marketing communication was a process of creating an integrated group of stimuli with the purpose of evoking a set of desired responses. According to this integrated marketing communications perspective, which has been adopted by many companies since the 1980's, rather than being considered in isolation from one another, each component of the marketing mix should be coordinated to present a unified image to consumers.

Considered from this perspective, fulfilling the viral marketing promise of Web 2.0 and social media requires that the viral marketing effort be part of a greater scheme of corporate communications. In other words, rather than merely focusing on spreading the word, the viral marketing effort should fit the brand personality (Webb, 2007). A particular case illustrating this point is the "Top This TV Challenge" campaign of Heinz®. In this campaign, Heinz® invited consumers to produce 30-second TV commercials for Heinz® Ketchup and submit the commercials on YouTube. The winner of the contest, determined first by a panel of judges and then by the votes of consumers, was awarded $57,000 and a chance to get the commercial aired on national television. The premise of the campaign was not only that it fit the "fun" brand image of Heinz® Ketchup but also that the consumers would play a crucial role in disseminating Heinz Ketchup's name. Just as intended, many of the 4,000 qualified contestants who posted their videos on YouTube (as required) also created MySpace and Facebook pages promoting their own videos and consequently Heinz Ketchup.

Another example illustrating the connection between viral marketing and an integrated marketing communications approach that provides a fit between the marketing campaign and the organizational image is the "Download Day" organized by Mozilla Firefox in June 2008. Mozilla is a not for profit organization that is mostly known for its Firefox Web Browser (a challenger of the market leader, Internet Explorer). The organization is a self-proclaimed open source project that publicly shares the source codes of their own software for the development of new Internet applications. Unlike its major competitors, such as Internet Explorer and Safari, the Firefox Web Browser is positioned as an "organic browser" that has been developed through a collaborative process whereby thousands of software developers – the majority of which are not employed by Mozilla – contribute to the software. Likewise, the dissemination of Firefox largely relies on volunteers "spreading" the software.

In June 2008, Mozilla created a Download Day event to promote the third version of its Firefox Web Browser. The purpose of the Download Day was to set a world record in the number of downloads in 24 hours. To inform would-be users about the event, Mozilla heavily utilized social media and viral marketing. Following the initial announcement, the word of mouth about the Download Day first quickly spread through social news aggregators such as Digg™ and Reddit.com. Then, the links in the social news aggregators forwarded interested users to the Download Day homepage. In addition to asking individuals to pledge to download Firefox on Download Day and providing an update on the number of individuals who pledged to download, the homepage also invited them to engage in viral marketing by inviting their social networks to the event via Facebook, Bebo and MySpace, promoting the event on microblogging Twitter-like websites or organizing "Download Fests" on university campuses.

These two examples provide important insights regarding the criteria for a successful viral marketing campaign online (and in social media):

1. **Campaign-Organizational Image Congruence:** In the Download Day example, the event, the promoted goal (setting a world record) and the method of dissemination of the information of the event (through social media) were in line with Mozilla's overall image as a non-corporate, decentralized and innovative organization that relies on volunteers and users for its promotion as well as software development. Similarly, the "Top This TV Challenge" campaign of Heinz® fits the "fun" brand image of Heinz® Ketchup.
2. **Inciting Virality and Buzz:** This is the key for creating a pull rather than inducing a push in an organization's marketing campaign. An attractive event (in this case a world record setting event) or a message is a crucial component in developing an organic viral marketing process. The ability to create buzz through the event will also increase the chances that the viral marketing campaign will supplement other marketing communication goals: such as, providing

material for other promotional efforts or getting coverage in traditional media—the latter being especially important for Firefox given that Mozilla does not have a centrally controlled advertising budget to spend on conventional media. For example, the overwhelming interest in the Top This TV Challenge (with 5.2 million views) also helped create publicity for the company in the mainstream media and prompted Heinz® to repeat the challenge in 2008.

3. **Getting Consumers to be Personally Invested:** Mozilla's Download Day emphasized not only the possibility of a world record but that the consumers would be an integral part of this unique success. In this case, the prospects of being a part of a possible Guinness World Record-setting activity may have increased the chances that consumers identify with (and are invested in) not only the product or the brand but also the success of the campaign. Perhaps, for the contestants in the Heinz® Top This TV Challenge, the personal investment was even higher because their own success (in terms of getting enough votes to win the contest) partly depended on the popular votes they would get from other consumers.
4. **Creating Levels of Viral Involvement:** In terms of options available for viral marketing, social media not only expand the available options but also create the possibility of multiple levels of viral involvement. For example, in the Heinz® Top This TV Challenge, the level of viral activity of a contestant that promotes his/her video will naturally be higher than a regular YouTube user who happens to come across a challenger's video that is worth sharing with friends. The Mozilla's Download Day event, on the other hand, systematically utilized the social media (and other venues) to create tiers of consumer involvement. For example, an enthusiastic Firefox user could go as far as organizing a download festival whereas a regular user of Facebook or MySpace could invite friends to pledge for the download on the Mozilla's Download Day homepage.

FUTURE TRENDS

As discussed in the preceding sections, a central tenet of the debates regarding the marketing potential of social media pertains to the balance that needs to be struck between the efficiency of automatic recommendation systems and the organic involvement created by the real community touch of viral marketing campaigns that invite consumers to actively participate in the dissemination of the marketing messages. On the one hand, systems such as Facebook's Beacon platform and MySpace's "influencer" option promise to deliver large-scale, automated word of mouth that can expand the reach of viral marketing campaigns. However, the perceived intrusiveness of such systems, as well as their tendency to use consumers as passive hubs to automatically relay marketing messages, may call into question the true virality of such advertising efforts, consequently reducing their appeal for consumers.

Recent discussions regarding "Web 3.0" and the future of the Internet may point to the direction that this uneasy relationship between virality and automatic customization may take. Despite frequent disagreements regarding the definition of Web 3.0, an increasing number of commentators have started to use the concept interchangeably with the Semantic Web – a set of technologies that enable software agents to understand, interpret and extract knowledge from information, making it possible for them to complete "sophisticated tasks for users" (Berners-Lee, Hendler & Lassila, 2001). Michael Bloch provides a simple example explaining how the Semantic Web would work:

You want to go out to dinner…and your car is in the shop… You issue a command for the agent to search for a restaurant serving Indian food *within a 10-mile radius…You want a restaurant that has a 4 star rating issued by a well-known restaurant critic. Furthermore, you want the table booked and a cab to pick you up from your place. Additionally you want a call to be made to your phone once that's all done; but you don't want to be disturbed by the call as you'll be in a meeting - just for the reservation details added to your phone organizer. (Bloch, 2007)*

As this example suggests, the Semantic Web is more than a compilation of web pages. Rather, it is a network of systems and databases that can communicate with each other to perform tasks on an individual's behalf. Moreover, as recent developments suggest, the Semantic Web will have the potential for subtler customization of information in accordance with the cognitive (and perhaps emotional) styles/needs of consumers. For example, an article by Hauser, Urban, Liberali and Braun (forthcoming) from MIT's Sloan School of Management announces an algorithm that uses clickstream data to morph the website content and format to the cognitive style of its users.

As evidenced by recently developed semantic web advertising applications (such as *Semantic-Match*™ – a semantic advertising platform that utilizes a natural language processing algorithm to understand content and sentiments and target advertising accordingly), when applied to online advertising, semantic capabilities can enhance customization, decrease errors that are associated with keyword targeted advertising and provide a more conversational interaction between the advertiser and the consumer. With respect to viral marketing, such advancements in language processing and customization can address an important shortcoming of passive virality by making it more personal. Whereas social network analyses aid the identification of hubs that can act as active viral marketers, improvements in natural language processing can prove beneficial in terms of understanding the communicative processes and dynamics within a social network. This information can help the marketing organization create different strategies to reach various potential hubs, create levels of viral involvement depending on the depth and the context of the communicative processes between network members, and customize the webpage that potential customers arrive at as a result of the viral marketing effort.

CONCLUSION

In recent years, Web 2.0 applications that enable web users to contribute to the world of online content have grown in popularity. In 2008, the Top 10 most frequently visited web site list of Alexa Internet – a web traffic information service – consistently included several social media sites: namely, YouTube, MySpace, Facebook, Hi5, Wikipedia and Orkut.com (2008). Despite their popular appeal, however, many of the Web 2.0 initiatives are still struggling to turn their popularity into financial success.

What is important to note is that when it comes to monetizing social media, there are no magic formulas. However, as explained above, the interactive nature of social media, combined with consumers' participation in the creation and dissemination of information, make viral marketing a viable candidate to fulfill the promise of a Web 2.0 of opportunities. In contrast to impersonal advertising methods that consumers do not trust and find intrusive, viral marketing through social media has the potential to be a personal, personable, participatory and trustworthy. source of information. Nonetheless, this should not be taken for granted that any and all viral marketing efforts in social media would be successful.

Extant literature suggests that there are certain prerequisites to a successful implementation of a viral marketing campaign in social media. First, as Webb (2007) suggests, because the company is going to have to rely on consumers to push the message, the message (and the product) should be worth pushing. Second, as consumers grow more suspicious of traditional advertising methods, marketers engaging in viral marketing in social media should pay the utmost attention to keeping viral marketing free from centralized interference that can damage its credibility. For example, Coplan (2007) notes that to remain credible, consumer marketers should be "honest about their opinions good and bad, open about their affiliation – and unpaid" (p. 26). This second prerequisite of success in social media marketing is closely related to the third one: In the world of consumer marketers, companies should learn to "cede control to customers" (cited in Poynter, 2008, p. 12). Partially, this means that viral marketing may be mixed with negative word of mouth and backlash (Gillin, 2007; Giuliana, 2005). At the same time, both positive and negative word of mouth should be considered as an opportunity to engage in a conversation with customers. For example, recently Cadbury PLC decided to relaunch Wispa (a chocolate bar discontinued in 2003) as a response to demands from 14,000 Facebook members (Poynter, 2008). Finally, as evidenced by the recent negative public reaction to the inadequate privacy protection on Facebook, marketers should be aware of the relatively intimate nature of social network sites.

REFERENCES

Alexa.com. (2008). *Global top 500.* Retrieved July 6, 2008, from http://www.alexa.com/site/ds/top_sites?ts_mode=global&lang=none

Allison, K. (2007). Facebook set for a delicate balancing act. *Financial Times (North American Edition)*, 8.

Anderson, C. (2006). *The log tail: How endless choice is creating unlimited demand*. London: Random House Business Books.

Andrejevic, M. (2007). *iSpy: Surveillance and power in the interactive era*. Lawrence, KS: University Press of Kansas.

Baruh, L. (2007). Read at your own risk: Shrinkage of privacy and interactive media. *New Media & Society*, *9*(2), 187–211. doi:10.1177/1461444807072220

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *American Scientist.* Retrieved June 3, 2008, from http://www.sciam.com/article.cfm?id=the-semantic-web.

Bloch, M. (2007, July 28). *The Semantic Web–Web 3.0*. Retrieved June 3, 2008, from http://www.tamingthebeast.net/blog/online-world/semantic-web-30-0707.htm

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. doi:10.1111/j.1083-6101.2007.00393.x

Clemons, E. K., Barnett, S., & Appadurai, A. (2007). *The future of advertising and the value of social networks: Some preliminary examinations*. Paper presented at the 9th International Conference on Electronic Commerce, Minneapolis, MN.

Clow, K. E., & Baack, D. (2007). *Integrated advertising, promotion, and marketing communications*. Upper Saddle River, NJ: Pearson Prentice Hall.

ComScore, Inc. (2008). *Top global Web properties*. Retrieved February 19, 2008, from http://www.comscore.com/press/data.asp

Coplan, J. H. (2007). Should friends pitch friends? *Adweek*, *48*, 26–26.

DeLozier, M. W. (1976). *The marketing communications process*. London: McGraw Hill.

Domingos, P. (2005). Mining social networks for viral marketing. *IEEE Intelligent Systems*, *20*(1).

Faber, R. J., Lee, M., & Nan, X. (2004). Advertising and the consumer information environment online. *The American Behavioral Scientist*, *48*(4), 447–466. doi:10.1177/0002764204270281

Gillin, P. (2007). *The new influencers: A marketer's guide to the new social media*. Sanger, CA: Quill Driver Books.

Giuliana, D. (2005). *Alternative marketing techniques for entrepreneurs*. Retrieved January 3, 2008, from http://www.scribd.com/doc/35013/Alternative-Marketing-Techniques-for-Entrepreneurs

Hall, E. (2007). Study: Popularity of social networks hampers ad growth. *Advertising Age*, *78*(31), 18.

Hanneman, R., & Riddle, M. (2005). *Introduction to social network methods*. Retrieved December 17, 2007, from http://www.faculty.ucr.edu/~hanneman/nettext/C10_Centrality.html

Hauser, J. R., Urban, G. L., Liberali, G., & Braun, M. (forthcoming). Website morphing. *Marketing Science*. Retrieved July 4, 2008, from http://web.mit.edu/hauser/www/Papers/Hauser_Urban_Liberali_Braun_Website_Morphing_May_2008.pdf

Klaassen, A., & Creamer, M. (2007). Facebook to add shopping service to its menu. *Advertising Age*, *78*(44), 39–40.

Kumar, R., Novak, J., & Tomkins, A. (2006). *Structure and evolution of online social networks*. Paper presented at the 12th International Conference on Knowledge Discovery in Data Mining, New York.

MacKelworth, T. (2007). *Social networks: Evolution of the marketing paradigm*. Retrieved March 12, 2008, from http://www.amacltd.com/pdf/SocialNetworksWhitePaper.pdf

Manovich, L. (2001). *The language of new media*. Cambridge, MA: MIT Press.

McAllister, M. P., & Turow, J. (2002). New media and the commercial sphere: Two intersecting trends, five categories of concern. *Journal of Broadcasting & Electronic Media*, *46*(4), 505–515. doi:10.1207/s15506878jobem4604_1

McCarthy, C. (2008, February 1). *Report: Facebook raises '08 revenue projection*. Retrieved March 6, 2008, from http://www.news.com/8301-13577_3-9862792-36.html

McCoy, S., Everard, A., Polak, P., & Galletta, D. F. (2007). The effects of online advertising. *Communications of the ACM*, *50*(3), 84–88. doi:10.1145/1226736.1226740

Morrissey, B. (2007). Social network ads: Too close, too personal? *Adweek*, *48*, 11–11.

Poynter, R. (2008). Facebook: The future of networking with customers. *International Journal of Market Research*, *50*(1), 11–12.

Rappaport, S. D. (2007). Lessons from online practice: New advertising models. *Journal of Advertising Research*, *47*(2), 135–141. doi:10.2501/S0021849907070158

Robinson, H., Wyscocka, A., & Hand, C. (2007). Internet advertising effectiveness: The effect of design on click-through rates for banner ads. *International Journal of Advertising*, *26*(4), 527–541.

Scott, J. P. (2000). *Social network analysis: A handbook*. London: Sage Publications.

Solove, D. J. (2007). *The future of reputation: Gossip, rumor, and privacy on the Internet*. New Haven, CT: Yale University Press.

Spertus, E., Sahami, M., & Büyükkökten, O. (2005). *Evaluating similarity measures: A large-scale study in the Orkut social network*. Paper presented at the 11th International Conference on Knowledge Discovery in Data Mining, Chicago, IL.

Subramani, M. R., & Rajagopalan, B. (2003). Knowledge-sharing and influence in online social networks via viral marketing. *Communications of the ACM*, *46*(12), 300–307. doi:10.1145/953460.953514

Thompson, R. J. (2007). Can't skip this: Consumers acclimating to Internet ads. *Brandweek*, *48*, 5. MySpace, Facebook and other social networking sites: Hot today, gone tomorrow? (2006, May 3). *Knowledge@Wharton*. Retrieved April 24, 2007, from http://knowledge.wharton.upenn.edu/article.cfm?articleid=1463

Tsai, J. (2008). Facebook's about-face. *Customer Relationship Management*, *12*(1), 17–18.

Wasserman, S., & Faust, K. (1995). *Social network analysis: Methods and applications*. Cambridge, MA: Cambridge University Press.

Webb, G. (2007, October/November). A new future for brand marketing. *The British Journal of Administrative Management*, 13-15.

Windley, P. (2007). *The fine line between advertising and recommendations*. Retrieved December 12, 2007, from http://blogs.zdnet.com/BTL/?p=7134

Zarsky, T. Z. (2004). Desperately seeking solutions: Using implementation-based solutions for the troubles of information privacy in the age of data mining and the Internet Society. *Maine Law Review*, *56*, 13–59.

KEY TERMS AND DEFINITIONS

Content Variability: *Content variability* refers to the notion that new media objects can exist in an infinite number of variations. This characteristic of new media is the result of the digital coding of content and consequently the modular nature of information.

Data Mining: Data mining is a technologically driven process of using algorithms to analyze data from multiple perspectives and extract meaningful patterns that can be used to predict future users behavior The market basket analysis system that Amazon.com uses to recommend new products to its customers on the basis of their past purchases is a widely known example of how data mining can be utilized in marketing.

Interactive Media: *Interactive media* is a catch-all term that is used to describe the two-way flow of information between the content user and the content producer. In addition to enabling consumers to actively participate in the production of content, interactive media also allow for the collection of real time data, which can later be used for content customization.

Semantic Web: The *Semantic Web* refers to a set of design principles, specifications, and web technologies that enable networked software agents to understand, interpret and communicate with each other to perform sophisticated tasks on behalf of users.

Social Network Analysis: Social network analysis is a research methodology utilized in research to investigate the structure and patterns of the relationship between social agents. Examples of sources of relational data include: contacts, connections, and group ties which can be studied using quantitative methodologies.

Social Network Sites: Social network sites are web-based systems that enable end-users to create online profiles, form associations with other users, and view other individuals' profiles. Examples of social network sites include: Match.

com, MySpace, Facebook, Orkut, Hi5, Bebo and LinkedIn.

Viral Marketing: *Viral marketing* refers to a form of word of mouth marketing that relies on consumers relaying product information, a marketing message or a personal endorsement to other potential buyers.

Web 2.0: Introduced in 2004, during a conference brainstorming session between O'Reilly Media and MediaLive International, *Web 2.0* refers to the second generation of web-based content. Rather than merely pointing to technological changes in the infrastructure of the Internet, the concept of Web 2.0 underlines the notion that end-users can do much more than consume readily available content: The user of Web 2.0 also plays a key role in the creation and the dissemination of content. Popular examples include: video-sharing and photo-sharing sites, such as YouTube and Flickr; social network sites, such as Orkut, MySpace and Facebook; and Weblogs (blogs).

This work was previously published in Handbook of Research on Social Interaction Technologies and Collaboration Software; Concepts and Trends, edited by T. Dumova; R. Fiordo, pp. 33-44, copyright 2010 by Information Science Reference (an imprint of IGI Global).

Chapter 8.5
Web Content Recommendation Methods Based on Reinforcement Learning

Nima Taghipour
Amirkabir University of Technology, Iran

Ahmad Kardan
Amirkabir University of Technology, Iran

ABSTRACT

Information overload is no longer news; the explosive growth of the Internet has made this issue increasingly serious for Web users. Recommender systems aim at directing users through this information space, toward the resources that best meet their needs and interests. In this chapter we introduce our novel machine learning perspective toward the web recommendation problem, based on reinforcement learning. Our recommendation method makes use of the web usage and content data to learn a predictive model of users' behavior on the web and exploits the learned model to make web page recommendations. Unlike other recommender systems, our system does not use the static patterns discovered from web usage data, instead it learns to make recommendations as the actions it performs in each situation. In the proposed method we combined the conceptual and usage information in order to gain a more general model of user behavior and improve the quality of web recommendations. A hybrid web recommendation method is proposed by making use of the conceptual relationships among web resources to derive a novel model of the problem, enriched with semantic knowledge about the usage behavior. The method is evaluated under different settings and it is shown how this method can improve the overall quality of recommendations.

DOI: 10.4018/978-1-60566-306-7.ch011

INTRODUCTION

The amount of information available on-line is increasing rapidly with the explosive growth of the World Wide Web and the advent of e-Commerce. Although this surely provides users with more options, at the same time makes it more difficult to find the "relevant" or "interesting" information from this great pool of information. This problem is commonly known as *information overload*: The state of having too much information to make a decision or remain informed about a topic. To address the problems caused by information overload, recommender systems have been introduced (Resnick & Varian, 1997). These systems can be defined as the personalized information technologies used to predict a user evaluation of a particular item (Deshpande & Karypis, 2004) or more generally as systems that guide users toward interesting or useful objects in a large space of possible options (Burke, 2002).

Recommender systems have been used in various applications ranging from predicting the products a customer is likely to buy (Shany *et al.*, 2005), movies, music or news that might interest the user (Konstan *et al.*, 1998; Zhang & Seo, 2001) and web pages that the user is likely to seek (Cooley *et al.*, 1999; Fu *et al.*, 2000; Joachims *et al.*, 1997; Mobasher *et al.*, 2000a), which is also the focus of this chapter. Web page recommendation is considered a user modeling or web personalization task (Eirinaki *et al.*, 2004). One research area that has recently contributed greatly to this problem is web mining. Most of the systems developed in this field are based on web usage mining which is the process of applying data mining techniques to the discovery of usage patterns form web data (Srivastava *et al*., 2000). These systems are mainly concerned with analyzing web usage logs, discovering patterns from this data and making recommendations based on the extracted knowledge (Fu *et al.*, 2000; Mobasher *et al.*, 2000a; Shahabi *et al.*, 1997; Zhang & Seo, 2001). One important characteristic of these systems is that unlike traditional recommender systems, which mainly base their decisions on user ratings on different items or other explicit feedbacks provided by the user (Deshpande & Karypis, 2004; Herlocker *et al.*, 2000), these techniques discover user preferences from their implicit feedbacks, e.g. the web pages they have visited. More recently, systems that take advantage of domain knowledge, e.g. a combination of content, usage and even structure information of the web, have been introduced and shown superior results in the web page recommendation problem (Li & Zaiane, 2004; Mobasher *et al.*, 2000b; Nakagawa & Mobasher, 2003).

In this chapter we will introduce a different machine learning perspective toward the web recommendation problem, which we believe is suitable to the nature of the problem and has some intrinsic advantages over previous methods. Our recommendation method falls in the category of methods that aim at supporting user's short-term information needs on a single website by recommending web pages to the user based on their navigation, such as previous works presented in (Mobasher *et al.*, 2000a,b; Li & Zaiane, 2004; Nakagawa & Mobasher, 2003). The proposed recommendation method makes use of the web usage and content data to learn a predictive model of users' behavior on the web and exploits the learned model to make web page recommendations to the users (Taghipour *et al.*, 2007; Taghipour & Kardan, 2007; Taghipour & Kardan, 2008). We model the recommendation process as a Reinforcement Learning (RL) problem (Sutton & Barto, 1998) or more specifically a Q-Learning problem. For this purpose we have devised state and action definitions and rewarding policies, considering common concepts and techniques used in the web mining domain. Then we train the system using web usage logs available as the training set, by adapting a variation of Q-learning algorithm. Our recommendation method differs from the previous methods in which the purpose was to find explicit and static patterns or rules, e.g. association rules or clusters of similar sessions, from the data. Here the system learns to make recommendations, i.e. predictions of interesting web pages, as the actions to perform in each situation (state). The choice of reinforcement learning was due to several reasons: It provides a framework appropriate to the nature of web page recommendation problem, mainly due to the concept of delayed reward or temporal difference in RL. Also, due to the characteristics of this type of learning and the fact that we are not making decisions explicitly from the static patterns discovered from the data, it provides us with a system which can potentially be constantly in the learning process and hence does not need periodic updates and can adapt itself to changes in website structure and content and more importantly to the new trends in user behavior.

We begin by introducing our method for web recommendations from web usage data (Taghipour *et al.*, 2007), i.e. usage logs available at web servers. Although the mentioned technique has shown promising results in comparison to common techniques like collaborative filtering and association rules, an analysis of the system's performance, shows how this method still suffers from the problems commonly faced by other usage-based techniques, such as the inability to generalize the learnt usage-based model which might results various problems such as the low coverage of recommendations or the "new item" problem commonly faced in collaborative filtering recommendations (Burke, 2002). To address these problems and to enhance our solution furthermore, we make use of the conceptual relationships among web pages and derive a novel model of the problem, enriched with semantic knowledge about the usage behavior (Taghipour & Kardan, 2008). We use existing methods to derive a conceptual structure of the website (Godoy & Amandi, 2005). Then we come up with new definitions for our states, actions and rewarding functions which capture the semantic implications of users' browsing behavior. Our new hybrid, i.e. usage- and content-based, model for the web page recommendation problem shows the flexibility of the reinforcement learning framework for the recommendation problem and how it can be extended to incorporate other sources of information. We evaluate our method under different settings and show how this method can improve the overall quality of web recommendations.

BACKGROUND

Web Recommender Systems

Recommender systems have been developed using various approaches and can be categorized in various ways (Burke, 2002). From an architectural point of view, recommendation generation approaches fall into two main categories: memory-based and model-based (Breese *et al.*, 1998). Memory based approaches memorize all the previous historical data, e.g. ratings, and make use of this data in the recommendation generation phase. Therefore, these techniques are more prone to scalability issues. Model-based approaches, on the other hand, use the available data to learn a model for recommendation. In these systems the computationally expensive learning phase is performed offline and hence they generally tend to scale better than memory based systems. It should be noted that as more data becomes available, memory based systems generally adapt better to changes in user interests. While in model based techniques, models must either be incremental or be rebuilt periodically to reflect the new trends.

From an algorithmic point of view recommender systems can be categorized into four general categories (Burke, 2002): knowledge-based systems, content-filtering systems, collaborative filtering systems and hybrid systems. Knowledge-based recommender systems make use of explicit domain knowledge about the items (such as their position in a concept hierarchy of the domain the items belong to) or the users (such as their demographic characteristics) to generate recommendations (Burke, 2000). Most of these recommenders employ some kind of knowledge-based decision rules for recommendation. This type of recommendation is heavily dependant on knowledge engineering by system designers to construct a rule base in accordance to the specific characteristics of the domain. While the user profiles are generally obtained through explicit interactions with users, there have also been some attempts at exploiting machine learning techniques for automatically deriving decision rules that can be used for personalization, e.g. (Pazzani, 1999).

In Content-based filtering systems, the user profile represents a content model of items in which that user has previously shown interest (Pazzani & Bilsus, 2007). These systems are rooted in information retrieval and information filtering research. The content model for an item is represented by a set of features or attributes characterizing that item. The recommendation generation is usually comprised of comparing extracted features from new items with content model in the user profile and recommending items that have adequate similarity to the user profile.

Collaborative techniques (Resnick & Varian, 1997; Herlocker *et al.*, 2000) are the most successful and the most widely used techniques in recommender systems, e.g. (Deshpande & Karypis, 2004; Konstan *et al.,* 1998; Wasfi, 1999). In the simplest from, in this class of systems, users are requested to rate the items they know and then the target user will be recommended the items that people with similar tastes had liked in the past. Recently, Web mining and especially web usage mining techniques have been used widely in web recommender systems (Cooley *et al.*, 1999; Fu *et al.*, 2000; Mobasher *et al.*, 2000a; Mobasher *et al.*, 2000b). Common approach in these systems is to extract navigational patterns from usage data by data mining techniques such as association rules and clustering, and making recommendations based on the extracted patterns. These approaches differ fundamentally from our method in which no static pattern is extracted from data.

More recently, systems that take advantage of a combination of content, usage and even structural information of the websites have been introduced and shown superior results in the web page recommendation problem (Li & Zaiane, 2004; Mobasher *et al.*, 2000b; Nakagawa & Mobasher, 2003). In (Nakagawa & Mobasher, 2003) the degree of connectivity based on the link structure of the website is used to choose from different usage based recommendation techniques, showing that sequential and non-sequential techniques could each achieve better results in web pages with different degrees of connectivity. A new method for generating navigation models is presented in (Li & Zaiane, 2004) which exploits the usage, content and structure data of the website. This method introduces the concept of user's *missions* to represent users' concurrent information needs. These missions are identified by finding content-coherent pages that the user has visited. Website structure is also used both for enhancing the content-based mission identification and also for ranking the pages in recommendation lists. In another approach (Eirinaki *et al.*, 2004, 2003) the content of web pages is used to augment usage profiles with semantics, using a domain-ontology and then performing data mining on the augmented profiles. Most recently, concept hierarchies were incorporated in a novel recommendation method based on web usage mining and optimal sequence alignment to find similarities between user sessions in (Bose *et al.*, 2007).

Markov Decision Process and Reinforcement Learning

Reinforcement learning (Sutton & Barto, 1998) is primarily known in machine learning research as a framework in which agents learn to choose the optimal action in each situation or state they are in. The agent is supposed to be in a specific state s, in each step it performs some action and transits to another state. After each transition the agent receives a reward. The goal of the agent is to learn which actions to perform in each state to receive the greatest accumulative reward, in its path to the goal states. The set of actions chosen in each state is called the agent's policy. One variation of this method is Q-Learning in which the agent does not compute explicit values for each state and instead computes a value function *Q(s,a)* which indicates value of performing action *a* in state *s* (Sutton & Barto, 1998; Mitchell, 1997). Formally the value of *Q(s,a)* is the discounted sum of future rewards that will be obtained by doing action *a* in s and subsequently choosing optimal actions. In order to solve the problem with Q-Learning we need to make appropriate definitions for our states and actions, consider a reward function suiting the problem and devise a procedure to train the system using web logs available to us.

The learning process of the agent can be formalized as a Markov Decision Process (MDP). The MDP model of the Problem includes:

1. **Set of states *S***, which represents the different 'situations' that the agent can observe. Basically, a state *s in S* must define what is important for the agent to know in order to take a good action. For a given situation, the complete set of states is called the *state space*.
2. **Set of possible actions *A*,** that the agent can perform in a given state *s* (*s* ∈ S) and that will produce a transition into a next state *s'* ∈ S. As we mentioned, the selection of the particular action depends on the *policy* of the agent. We formally define the policy as a function that indicates for each state *s*, the action *a* ∈ A taken by the agent in that state. In general, it is assumed that the environment, with which the agent interacts, is non-deterministic, i.e., after executing an action, the agent can transit into many alternative states.
3. **Reward function *rew*(*s, a*)** which assigns a scalar value, also known as the *immediate reward*, to the performance of each action *a* ∈ A taken in state *s* ∈ S. For instance, if the agent takes an action that is satisfactory for the user, then the agent should be rewarded with a positive immediate reward. On the other hand, if the action is unsatisfactory, the agent should be punished through a negative reward. However, the agent cannot know the reward function exactly, because the reward is assigned to it through the environment. This function can play a very important role in an MDP problem.
4. **Transition function *T*(*s, a, s'*)** which gives the probability of making a transition from state *s* to state *s'* when the agent performs the action *a*. This function completely describes the non-deterministic nature of the agent's environment. Explicit use of this function can be absent in some versions of Q-Learning.

Reinforcement Learning in Recommender Systems

Reinforcement Learning (RL) has been previously used for recommendations in several applications. Web Watcher (Joachims *et al.*, 1997), exploits Q-Learning to guide users to their desired pages. Pages correspond to states and hyperlinks to actions, rewards are computed based on the similarity of the page content and user profile keywords. There are fundamental differences between Web Watcher and our approach, two of the most significant are: (a) our approach requires no explicit user interest profile in any form, and (b) unlike our method, Web Watcher makes no use of previous usage based data. In most other systems, reinforcement learning is used to reflect user feedback and update current state of recommendations. A general framework is presented in (Golovin and Rahm, 2004), which consists of a database of recommendations generated by various models and a learning module that updates the weight of each recommendation by user feedback. In (Srivihok & Sukonmanee, 2005) a travel recommendation agent is introduced which considers various attributes for trips and customers, computes each trip's value with a linear function and updates function coefficients after receiving each user feedback. RL is used for information filtering in (Zhang & Seo, 2001) which maintains a profile for each user containing keywords of interests and updates each word's weight according to the implicit and explicit feedbacks received from the user. In (Shany *et al.*, 2005) the recommendation problem is modeled as an MDP. The system's states correspond to user's previous purchases, rewards are based on the profit achieved by selling the items and the recommendations are made using the theory of MDP and their novel state-transition function. In a more recent work (Mahmood & Ricci, 2007) RL is used in the context of a conversational travel recommender system in order to learn optimal interaction strategies. They model the problem with a finite state-space based on variables like the interaction stage, user action and the result size of a query. The set of actions represent what the system chooses to perform in each state e.g. executing a query, suggesting modification. Finally RL is used to learn an optimal strategy, based on a user behavior model. To the best of our knowledge our method differs from previous work, as none of them used reinforcement learning to train a system in making web site recommendations merely from web usage data.

REINFORCEMENT LEARNING FOR USAGE-BASED WEB PAGE RECOMMENDATION

The specific problem which our system is supposed to solve, can be summarized as follows: the system has, as input data, the log file of users' past visits to the website, these log files are assumed to be in any standard log format, containing records each with a user ID, the sequence of pages the user visited during a session and typically the time of each page request. A user session is defined as a sequence of temporally compact accesses by a user. Since web servers do not typically log usernames, sessions are considered as accesses from the same IP address such that they satisfy some constraints, e.g. the duration of time elapsed between any two consecutive accesses in the session is within a pre-specified threshold (Cooley et. al, 1999).

A user enters our website and begins requesting web pages, like a typical browser mostly by following the hyperlinks on web pages. Considering the pages this user has requested so far, the system has to predict in what other pages the user is probably interested and recommend them to her. Table 1 illustrates a sample scenario. Predictions are considered successful if the user chooses to visit those pages in the remaining of that session, e.g. page *c* recommended in the first step in Table 1. Obviously the goal of the system would be to make the most successful recommendations.

Modeling Recommendations as a Q-Learning Problem

Using the Analogy of a Game

In order to better represent our approach toward the problem we try to use the notion of a game. In a typical scenario a web user visits pages sequentially from a web site, let's say the sequence a user *u* requested is composed of pages *a, b, c* and *d*. Each page the user requests can be considered a step or move in our game. After each step the user takes, it will be the system's turn to make a move. The system's purpose is to predict user's next move(s) with the knowledge of his previous moves. Whenever the user makes a move (requests a page), if the system has previously predicted the move, it will receive positive points and otherwise it will receive none or negative points. For example predicting a visit of page *d* after viewing pages *a* and *b* by the user in the above example yields in positive points for the system. The ultimate goal of the system would be to gather as much points

Table 1. A sample user session and system recommendations

Visited Page	a	b	c	d	e	f
Navigation Trail	a	ab	abc	abcd	abcde	abcdef
System Prediction	c	d	e	s	f	h

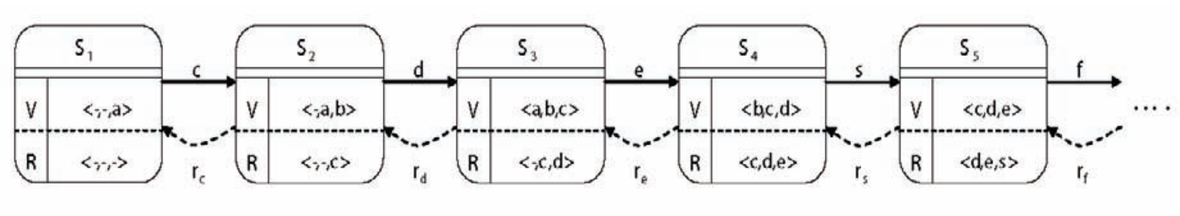
as possible during a game or actually during a user visit from the web site.

Some important issues can be inferred from this simple analogy: first of all, we can see the problem certainly has a stochastic nature and like most games, the next state cannot be computed deterministically from our current state and the action the system performs due to the fact that the user can choose from a great number of moves. This must be considered in our learning algorithm and our update rules for Q values; the second issue is what the system actions should be, as they are what we ultimately expect the system to perform. Actions will be prediction or recommendation of web pages by the system in each state. Regarding the information each state must contain, by considering our definition of actions, we can deduct that each state should at least show the history of pages visited by the user so far. This way we'll have the least information needed to make the recommendations. This analogy also determines the basics of rewarding function. In its simplest form it shall consider that an action should be rewarded positively if it recommends a page that will be visited in one of the consequent states, not necessarily the immediate next state. Of course, this would be an over simplification and in practice the reward would depend on various factors described in the coming sections. One last issue which is worth noting about the analogy is that this game cannot be categorized as a typical 2-player game in which opponents try to defeat each other, as in this game clearly the user has no intention to mislead the system and prevent the system from gathering points. It might be more suitable to consider the problem as a competition for different recommender systems to gather more points, than a 2-player game. Because of this intrinsic difference, we cannot use self-play, a typical technique used in training RL systems (Sutton & Barto, 1998) to train our system and we need the actual web usage data for training.

Modeling States and Actions

Considering the above observations we begin the definitions. We tend to keep our states as simple as possible, at least in order to keep their number manageable. Regarding the states, we can see keeping only the user trail can be insufficient. With that definition it won't be possible to reflect the effect of an action a performed in state s_i, in any consequent state s_{i+n} where $n>1$. This means the system would only learn actions that predict the immediate next page which is not the purpose of our system. Another issue we should take into account is the number of possible states: if we allow the states to contain any given sequence of page visits clearly we'll be potentially faced by an infinite number of states. What we chose to do was to limit the page visit sequences to a constant number. For this purpose we adopted the notion of N-Grams which is commonly applied in similar personalization systems based on web usage mining (Mobasher *et al.*, 2000a; Mobasher *et al.*, 2000b). In this model we put a sliding window of size w on user's page visits, resulting in states containing only the last w pages requested by the user. The assumption behind this model is that knowing only the last w page visits of the user, gives us enough information to predict his future page requests. The same problem rises when considering the recommended pages' sequence in

Figure 1. States and actions in the recommendation problem



the states, for which we take the same approach of considering *w'* last recommendations.

Regarding the actions, we chose simplicity. Each action is a single page recommendation in each state. Considering multiple page recommendations might have shown us the effect of the combination of recommended pages on the user, in the expense of making our state space and rewarding policy much more complicated.

Thus, we consider each state *s* at time *t* consisting of two sequences *V, R* indicating the sequence of visited and previously recommended pages respectively:

$$\begin{aligned} V_s &= < v_{t-w+1}, v_{t-w+2}, \ldots, v_t > \\ R_s &= < r_{t-w'+1}, r_{t-w'+2}, \ldots, r_t > \end{aligned} \quad (1)$$

Where v_{t-w+i} indicates the i^{th} visited page in the state and r_{t-w+i} indicates the i^{th} recommended page in the state *s*. The corresponding states and actions of the user session of Table 1 are presented in Figure 1, where straight arrows represent the actions performed in each state and the dashed arrows represent the reward received for performing each action.

Choosing a Reward Function

The basis of reinforcement learning lies in the rewards the agent receives, and how it updates state and action values. As with most stochastic environments, we should reward the actions performed in each state with respect to the consequent state resulted both from the agent's action and other factor's in the environment on which we might not have control. These consequent states are sometimes called the after-states (Sutton & Barto, 1998). Here this factor is the page the user actually chooses to visit. We certainly do not have a predetermined function *rew(s,a)* or even a state transition function $\delta(s, a)$ which gives us the next state according to current state s and performed action a.

It can be inferred that the rewards are dependent on the after state and more specifically on the intersection of previously recommended pages in each state and current page sequence of the state. Reward for each action would be a function of $V_{s'}$ and $R_{s'}$ where *s'* is our next state. One tricky issue worth considering is that though tempting, we should not base on rewards on $|V_s' \cap R_{s'}|$ since it will cause extra credit for a single correct move. Considering the above example a recommendation of page b in the first state shall be rewarded only in the transition to the second state where user goes to page b, while it will also be present in our recommendation list in the third state. To avoid this, we simply consider only the occurrence of the last visited page in state *s'*, in the recommended pages list to reward the action performed in the previous sate *s*. To complete our rewarding procedure we take into account common metrics used in web page recommender systems. One issue is considering when the page was predicted by the system and when the user actually visited the page. According to the goal of the system this might influence our rewarding. If we consider shortening user navigation as a sign of successful guidance of user to his required information, as is the most common case in recommender systems (Li & Zaiane, 2004; Mobasher *et al.,* 2000a) we should consider a greater reward for pages predicted sooner in the user's navigation path and vice versa. Another factor commonly considered in theses systems (Mobasher *et al.,* 2000a; Liu *et al.,* 2004; Fu *et al.,* 2000) is the time the user spends on a page, assuming the more time the user spends on a page the more interested he probably has been in that page. Taking this into account we should reward a successful page recommendation in accordance with the time the user spends on the page. The rewarding can be summarized as follows:

Algorithm 1. Usage Based Reward Function

1: **Assume** $\delta(s,a) = s'$

2: $K_{s'} = V_{s',w} \cap R_{s'} = v_{t+1} \cap R_{s'}$

3: **If** $K_{s'} \neq \emptyset$

4: **For each** *page k* **in** $K_{s'}$

5: *rew(s,a) += UBR(Dist(*$R_{s'}$*, k),Time(*v_{t+1}*))*

6: **End For**

7: **End If**

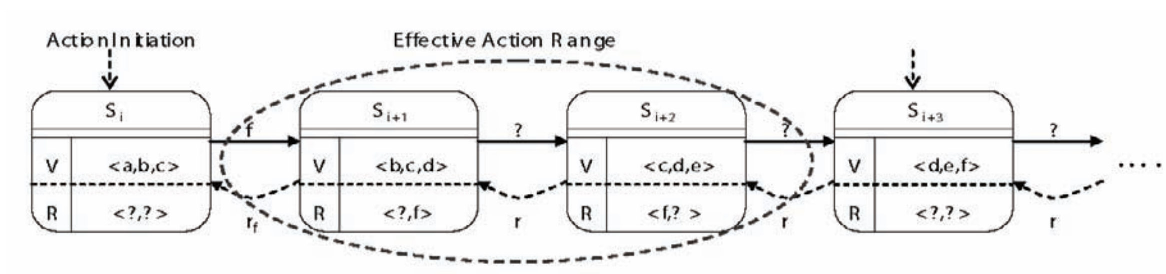
In line 1, $\delta(s,a) = s'$ shows that the transition of the system to the next state *s'* after performing *a* in state *s*. $K_{s'}$ represents the set of correct recommendations in each step and *rew(s,a)* is the reward of performing action *a* in state *s*. *Dist(*$R_{s'}$*,k)* is the distance of page *k* from the end of the recommended pages list in state *s'* and *Time(*v_{t+1}*)* indicates the time user has spent on the last page of the state. Here, *UBR* is the Usage-Based Reward function, combining these values to calculate the reward function *rew(s,a)*. We chose a simple linear combination of these values as Equation (2):

$$UBR(Dist, Time) = \alpha \times Dist + \beta \times Time \qquad (2)$$

Where $\alpha + \beta = 1$ and both α and β include a normalizing factor according to the maximum values dist and time can take.

The last modification we experimented was changing our reward function. We noticed as we put a sliding window on our sequence of previously recommended pages, practically we had limited the effect of each action to *w'* next states as can be seen in Figure 2. As can be seen in the example presented in this figure, a correct recommendation of page *f* in state s_i will not be rewarded in state s_{i+3} when using a window of size 2 on the *R* sequence (*w'=2)*. After training the system using this definition, the system was mostly successful in recommending pages visited around *w'* steps ahead. Although this might be quite acceptable while choosing an appropriate value for *w'*, it tends to limit system's prediction ability as large numbers of *w'* make our state space enormous. To overcome this problem, we devised a rather simple modification in our reward function: what we needed was to reward recommendation of a page if it is likely to be visited an unknown number of states ahead. Fortunately our definition of states and actions gives us just the information we need and this information is stored in *Q* values of each state. The basic idea is that when an action/recommendation is appropriate in state s_i, indicating the recommended page is likely to occur in the following states, it should also be considered appropriate in state s_{i-1} and the actions in that state that frequently lead to s_i. Following this recursive procedure we can propagate the value of performing a specific action beyond the limits imposed by *w'*. This change is easily reflected in our learning system by considering value of *Q(s',a)* in computation of *rew(s,a)* with a coefficient like γ. It should be taken into account that the effect of this modification in our reward

Figure 2. An example of limited action effectiveness due to the size of the recommendation window



function must certainly be limited as in its most extreme case where we only take this next *Q* value into account we're practically encouraging recommendation of pages that tend to occur mostly in the end of user sessions.

Having put all the pieces of the model together, we can get an initial idea why reinforcement learning might be a good candidate for the recommendation problem: it does not rely on any previous assumptions regarding the probability distribution of visiting a page after having visited a sequence of pages, which makes it general enough for diverse usage patterns as this distribution can take different shapes for different sequences. The nature of this problem matches perfectly with the notion of delayed reward or what is commonly known as temporal difference: the value of performing an action/recommendation might not be revealed to us in the immediate next state and sequence of actions might have led to a successful recommendation for which we must credit rewards. What the system learns is directly what it should perform, though it is possible to extract rules from the learned policy model, its decisions are not based on explicitly extracted rules or patterns from the data. One issue commonly faced in systems based on patterns extracted from training data is the need to periodically update these patterns in order to make sure they still reflect the trends residing in user behavior or the changes of the site structure or content. With reinforcement learning the system is intrinsically learning even when performing in real world, as the recommendations are the actions the system performs, and it is commonplace for the learning procedure to take place during the interaction of system with its environment.

Training the System

We chose Q-Learning as our learning algorithm. This method is primarily concerned with estimating an evaluation of performing specific actions in each state, known as Q-values. Each *Q(s,a)* indicates an estimate of the accumulative reward achievable, by performing action *a* in state *s* and performing the action *a'* with highest *Q(s',a')* in each future state *s'*. In this setting we are not concerned with evaluating each state in the sense of the accumulative rewards reachable from this state, which with respect to our system's goal can be useful only if we can estimate the probability of visiting the following states by performing each action. On the other hand Q-Learning provides us with a structure that can be used directly in the recommendation problem, as recommendations in fact are the actions and the value of each recommendation/action shows an estimation of how successful that prediction can be. Another decision is the update rule for Q values.

Because of the non-deterministic nature of this problem we use the following update rule (Sutton & Barto, 1998):

$$Q_n(s,a) = (1-\alpha_n)Q_{n-1}(s,a) + \alpha_n[r(s,a) + \gamma \max_{a'} Q_{n-1}(\delta(s,a),a')] \quad (3)$$

Figure 3. Algorithm 2: Training procedure

- Initial values of *Q(s,a)* for each pair *s,a* are set to zero
- Repeat until convergence
  - A random episode is chosen from the set of training episodes.
  - *s* is set to the first step/state of the episode.
  - For each step of the episode do
    - Choose an action *a* of this state using the ε-greedy policy.
    - Perform action a observe the next state and compute *rew(s,a)* as described before.
    - Update value of *Q(s,a)* with the above equation.
    - $s \leftarrow s'$.

With

$$\alpha_n = \frac{1}{1 + visits_n(s,a)} \tag{4}$$

Where $Q_n(s,a)$ is the Q-Value of performing *a* in state *s* after *n* iterations, and $visits_n(s,a)$ indicates the total number of times this state-action pair, i.e. *(s,a)*, has been visited up to and including the n^{th} iteration. This rule takes into account the fact that doing the same action can yield different rewards each time it is performed in the same state. The decreasing value of α_n causes these values to gradually converge and decreases the impact of changing reward values as the training continues.

What remains about the training phase is how we actually train the system using web usage logs available. As mentioned before these logs consist of previous user sessions in the web site. Considering the analogy of the game they can be considered as a set of opponent's previous games and the moves he tends to make. We are actually provided with a set of actual episodes occurred in the environment, of course with the difference that no recommendations were actually made during these episodes. The training process can be summarized as Figure 3. Algorithm 2:

One important issue in the training procedure is the method used for action selection. One obvious strategy would be for the agent in each state *s* to select the action *a* that maximizes *Q(s,a)* hereby exploiting its current approximation. However, with this greedy strategy there's the risk of over-committing to actions that are found during early training to have high Q values, while failing to explore other actions that might have even higher values (Mitchell, 1997). For this reason, it is common in Q learning to use a probabilistic approach to selecting actions. A simple alternative is to behave greedily most of the time, but with small probability ε, instead select an action at random. Methods using this near-greedy action selection rule are called ε-greedy methods (Sutton & Barto, 1998).

The choice of ε-greedy action selection is quite important for this specific problem as the exploration especially in the beginning phases of training, is vital. The Q values will converge if each episode, or more precisely each state-action pair is visited infinitely. In our implementation of the problem convergence was reached after a few thousand (between 3000 and 5000) visits of each episode. This definition of the learning algorithm completely follows a *TD(0)* off-policy learning procedure (Sutton & Barto, 1998), as we take an estimation of future reward accessible from each state after performing each action by considering the maximum *Q* value in the next state.

EXPERIMENTAL EVALUATION OF THE USAGE BASED APPROACH

We evaluated system performance in the different settings described above. We used simulated log files generated by a web traffic simulator to tune our rewarding functions. The log files were simulated for a website containing 700 web pages. We pruned user sessions with a length smaller than 5 and were provided with 16000 user sessions with average length of eight. As our evaluation data set we used the web logs of the Depaul University website, one of the few publicly available and widely used datasets, made available by the author of (Mobasher *et al.*, 2000a). This dataset is pre-processed and contains 13745 user sessions in their visits on 687 pages. These sessions have an average length around 6. The website structure is categorized as a dense one with high connectivity between web pages according to (Nakagawa & Mobasher, 2003). 70% of the data set was used as the training set and the remaining was used to test the system. For our evaluation we presented each user session to the system, and recorded the recommendations it made after seeing each page the user had visited. The system was allowed to make r recommendations in each step with $r<10$ and $r < \sqrt{O_v}$ where O_v is the number of outgoing links of the last page v visited by the user. This limitation on number of recommendations is adopted from (Li & Zaiane, 2004). The recommendation set in each state is composed by selecting the *top-r* actions of the sates with the highest Q-values, again by a variation of the ε-greedy action selection method.

Evaluation Metrics

To evaluate the recommendations we use the metrics presented in (Li & Zaiane, 2004) because of the similarity of the settings in both systems and the fact that we believe these co-dependent metrics can reveal the true performance of the system more clearly than simpler metrics. Recommendation Accuracy and Coverage are two metrics quite similar to the precision and recall metrics commonly used in information retrieval literature.

Recommendation accuracy measures the ratio of correct recommendations among all recommendations, where correct recommendations are the ones that appear in the remaining of the user session. If we have M sessions in our test log, for each visit session m after considering each page p, the system generates a set of recommendations *Rec(p)*. To compute the accuracy, *Rec(p)* is compared with the rest of the session *Tail(p)* as Equation (5). This way any correct recommendation is evaluated exactly once.

$$Accuracy = \frac{\sum_m \frac{\left|\bigcup_p (Tail(p) \cap \mathrm{Re}\,c(p))\right|}{\left|\bigcup_p \mathrm{Re}\,c(p)\right|}}{M} \tag{5}$$

Recommendation coverage on the other hand shows the ratio of the pages in the user session that the system is able to predict before the user visits them:

$$Coverage = \frac{\sum_m \frac{\left|\bigcup_p (Tail(p) \cap \mathrm{Re}\,c(p))\right|}{\left|\bigcup_p Tail(p)\right|}}{M} \tag{6}$$

As is the case with precision and recall, these metrics can be useful indicators of the system performance only when used in accordance to each other and lose their credibility when used individually. As an example, consider a system that recommends all the pages in each step, this system will gain 100% coverage, of course in the price of very low accuracy.

Another metric used for evaluation is called the shortcut gain which measures how many page-visits users can save if they follow the recommendations. The shortened session is derived by eliminating the intermediate pages in the session that the user could escape visiting, by following the recommendations. A visit time threshold is used on the page visits to decide which pages are auxiliary pages as proposed by Li and Zaiane (2004). If we call the shortened session *m'*, the shortcut gain for each session is measured as follows:

$$ShortcutGain = \frac{\frac{|m| - |m'|}{|m|}}{M} \tag{7}$$

Evaluation Results

In the first set of experiments we tested the effect of different decisions regarding state definition, rewarding function, and the learning algorithm on the system behavior. Afterwards we compared the system performance to the other common techniques used in recommendation systems.

Sensitivity to Active Window Size on User Navigation Trail

In our state definition, we used the notion of N-Grams by putting a sliding window on user navigation paths. The implication of using a sliding window of size w is that we base the prediction of user future visits on his w past visits. The choice of this sliding window size can affect the system in several ways. A large sliding window seems to provide the system a longer memory while on the other hand causing a larger state space with sequences that occur less frequently in the usage logs. We trained our system with different window sizes on user trail and evaluated its performance as seen in Figure 4. In these experiments we used a fixed window size of 3 on recommendation history.

As our experiments show the best results are achieved when using a window of size 3. It can be inferred form this diagram that a window of size 1 which considers only the user's last page visit does not hold enough information in memory to make the recommendation, the accuracy of recommendations improve with increasing the window size and the best results are achieved with a window size of 3. Using a window size larger than 3 results in weaker performance (only shown up to *w=4* in Figure 4 for the sake of readability), it seems to be due to the fact that, as mentioned above, in these models, states contain sequences of page visits that occur less frequently in web usage logs, causing the system to make decisions based on weaker evidence. In our evaluation of the short cut gain there was a slight difference when using different window sizes.

Sensitivity to Active Window Size on Recommendations

In the next step we performed similar experiments, this time using a constant sliding window of size 3 on user trail and changing size of active window on recommendations history. As this window size was increased, rather interesting results was achieved as shown in Figure 5.

In evaluating system accuracy, we observed improvement up to a window of size 3, after that increasing the window size caused no improvement while resulting in larger number of states. This increase in the number of states is more intense than when the window size on user trail was increased. This is manly due to the fact that the system is exploring and makes any combination of recommendations to learn the good ones. The model consisting of this great number of states is in no way efficient, as in our experiments on the test data only 25% of these states were actually visited. In the sense of shortcut gain the system achieved, it was observed that shortcut gain increased almost constantly with increase in window size, which seems a natural consequence as described in section "Reinforcement learning for usage-based web page recommendation".

Figure 4. System performance with various user visit windows sizes (w)

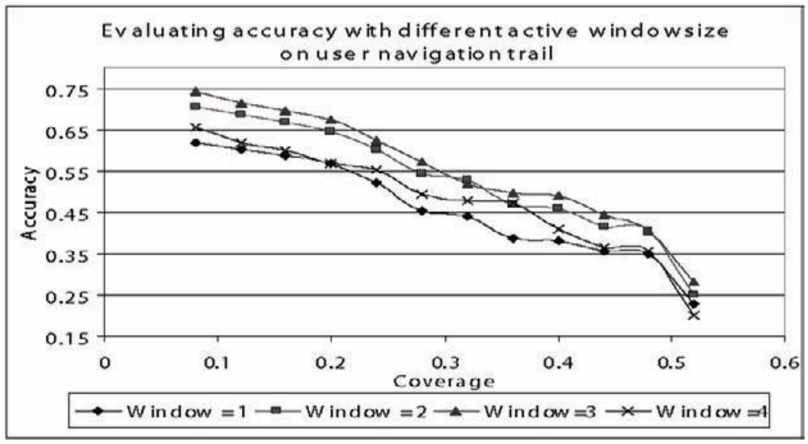
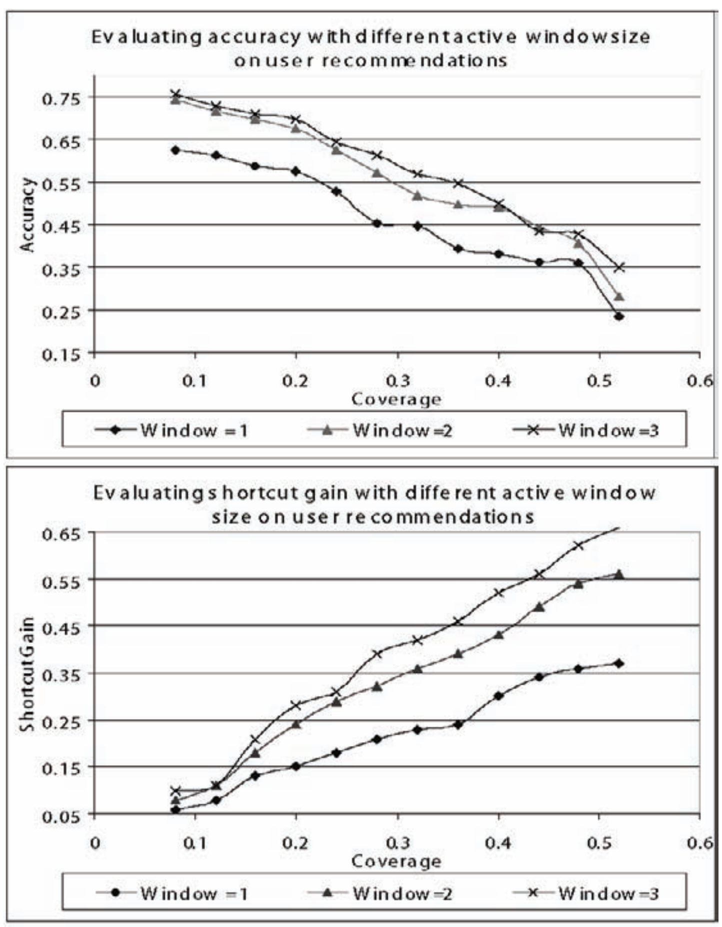


Figure 5. System performance with different active recommendation windows (w')



Evaluating Different Reward Functions

Next we changed the effect of parameters constituting our reward function. First we began by not considering the *Dist* parameter, described in section "Reinforcement learning for usage-based web page recommendation", in our rewards. We gradually increased it's coefficient in steps of 5% and recorded the results as shown in Table 2. These results show that increasing the impact of this parameter in our rewards up to 15% of total reward can result both in higher accuracy and higher shortcut gain. Using values greater than 15% has a slight negative effect on accuracy with a slight positive effect on shortcut gain and keeping it almost constant. This seems a natural consequence since although we're paying more attention to pages that tend to appear later in the user sessions, the system's vision into the future is bounded by the size of window on recommendations. This limited vision also explains why our accuracy is not decreasing as expected.

The next set of experiments tested system performance with the reward function that considers *next state Q-value* of each action in rewarding the action performed in the previous state, as described in section "Reinforcement learning for usage-based web page recommendation". We began by increasing the coefficient of this factor (γ) in the reward function the same way we did for the *Dist* parameter. In the beginning increasing this value, lead to higher accuracy and shortcut gains. After reaching an upper bound, the accuracy began to drop. In these settings, recommendations with higher values were those targeted toward the pages that occurred more frequently in the end of user sessions. These recommended pages, if recommended correctly, were only successful in predicting the last few pages in the user sessions. As expected, shortcut gain increased steadily with increase in this value up to a point where the recommendations became so inaccurate that rarely happened anywhere in the user sessions. More detailed evaluation results, which are not presented here due to space constraints, can be found in (Taghpour *et al.*, 2007).

A Comparison with other Methods

Finally we observed our system performance in comparison with two other methods: (a) association rules, as one approach based on of the usage-pattern and one of the most common approaches in web mining based recommender systems (Mobasher *et al.*, 2000a,b); and collaborative filtering which is commonly known as one of the most successful approaches for recommendations. We chose item-based collaborative filtering with

Table 2. System performance with varying α in the reward function. (AC=Accuracy, SG=Shortcut Gain)

Coverage	**Performance**									
	$\alpha = 0.1$		$\alpha = 0.15$		$\alpha = 0.20$		$\alpha = 0.25$		$\alpha = 0.30$	
	AC	SG	AC	SG	AC	SG	AC	SG	AC	SG
10	.75	.15	.78	.17	.76	.17	.73	.18	.69	.18
15	.71	.28	.73	.33	.72	.35	.69	.34	.65	.35
20	.69	.37	.68	.40	.67	.41	.67	.41	.61	.41
25	.65	.40	.66	.44	.65	.44	.61	.46	.58	.46
30	.55	.43	.57	.50	.54	.53	.52	.54	.49	.57
40	.48	.48	.50	.54	.45	.57	.40	.58	.36	.57
50	.36	.51	.39	.57	.33	.58	.29	.58	.27	.59

probabilistic similarity measure (Deshpande & Karypis, 2004), as a baseline for comparison because of the promising results it had shown. It should be noted that these techniques have already shown significantly superior results compared to common sense methods such as recommending most popular items (pages) of a collection. In Figure 6 the performance of these systems in the sense of their accuracy and shortcut gain in different coverage values can be seen. The statistical significance of any differences in performance between two methods was evaluated using two-tailed paired t-tests (Mitchell, 1997).

At lower coverage values we can see although our system still has superior results especially over association rules, accuracy and shortcut gain values are rather close. As the coverage increases, naturally accuracy decreases in all systems, but our system gains much better results than the other two systems. It can be seen the rate in which accuracy decreases in our system is lower than other two systems; at lower coverage values where the systems made their most promising recommendations (those with higher values), pages recommended were mostly the next immediate page and as can be seen had an acceptable accuracy. At lower coverage rates, where recommendations with lower values had to be made our system began recommending pages occurring in the session some steps ahead, while the other approaches also achieved greater shortcut gains, as the results show their lower valued recommendations were not as accurate and their performance declined more intensely. Regardless of the size of the difference at different coverage values, all the differences in Accuracy and Shortcut Gain between our proposed method and the baseline approaches are statistically significant (p<0.001 on the t-test).

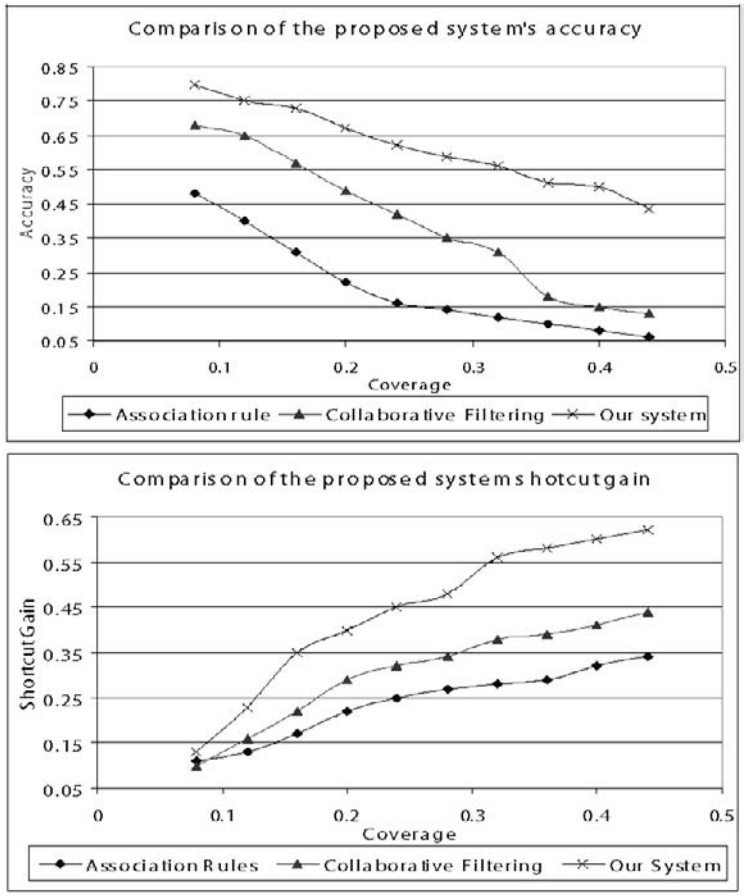
INCORPORATING CONTENT FOR HYBRID WEB RECOMMENDATIONS

In this section we exploit the idea of combining content and usage information to enhance the reinforcement learning solution, we had devised for web page recommendations based on web usage data. Although the mentioned technique showed promising results in comparison to common techniques like collaborative filtering and association rules, an analysis of the system's performance, reveals that this method still suffers from the problems commonly faced by other usage-based techniques. To address these problems, we made use of the conceptual relationships among web pages and derived a new model of the problem, enriched with semantic knowledge about the usage behavior. We used existing methods to derive a conceptual structure of the website. Then we came up with new definitions for our states, actions and rewarding functions which capture the semantic implications of users browsing behavior.

Observations on Performance of the Usage-Based Approach

In our evaluation of the system, we noticed that although we were faced with a rather large number of states, there were cases where the state resulted from the sequence of pages visited by the user had actually never occurred in the training phase. Although not the case here, this problem can be also due to the infamous "new item" problem commonly faced in collaborative filtering (Burke, 2002; Mobasher *et al.*, 2000b) when new pages are added to the website. In situations like these the system was unable to make any decisions regarding the pages to recommend to the users. Moreover, the overall coverage of the system on the website, i.e. percentage of the pages that

Figure 6. Comparing our system's performance with two other common methods



were recommended at least once, was rather low (55.06%). Another issue worth considering is the fact that the mere presence of a state in our state space cannot guarantee a high quality recommendation, to be more accurate it can be said that even a high Q-value cannot guarantee a high quality recommendation by itself. Simply put, when a pattern has few occurrences in the training data it cannot be a strong basis for decision making, a problem addressed in other methods by introducing metrics like support threshold in association rules (Mobasher *et al.*, 2000b). Similarly in our case a high Q-value, like a high confidence for an association rule, cannot be trusted unless it has strong supporting evidence in the data. In summary, there are cases when historical usage data provides no evidence or evidence that's not strong enough to make a rational decision about user's need or behavior.

This is a problem common in recommender systems that have usage data as their only source of information. Note that in the described setting, pages stored in the *V* sequence of each state *S* are treated as items for which the only information available is their id. The system relies solely on usage data and thus is unable to make any generalization. One common solution to this problem is to incorporate some semantic knowledge about the items being recommended, into the system. In the next section we describe our approach for adopting this idea.

Incorporating Concept Hierarchies in the Recommendation Model

One successful approach used to enhance web usage mining, is exploiting content information to transform the raw log files into more meaningful semantic logs (Bose *et al.*, 2006; Eirinaki *et al.*, 2004) and then applying data mining techniques on them. In a typical scenario pages are mapped to higher level concepts e.g. catalogue page, product page, etc and a user session consisting of sequential pages will be transformed to a sequence of concepts followed by the user. Consequently, generalized patterns are extracted from these semantically enhanced log files which can then be used for personalization.

We decided to exploit the same techniques in our system to improve our state and action model. In order to make our solution both general and applicable, we avoided using an ad-hoc concept hierarchy for this purpose. Instead we chose to exploit hierarchical and conceptual document clustering which can provide us with semantic relationships between pages without the need of a specifically devised ontology, concept hierarchy or manual assignment of concepts to pages. An important factor in our selection was the ability of the method to perform incremental document clustering, since we prefer to come up with a solution that is able to cope with the changes in the web site content and structure. In order to map pages to higher level concepts, we applied the DCC clustering algorithm (Godoy & Amandi, 2005) on the web pages. It is an incremental hierarchical clustering algorithm which is originally devised to infer user needs and falls in the category of conceptual clustering algorithms as it assigns labels to each cluster of documents. In this method each document would be assigned to a single class in the hierarchy. This method has shown promising results in the domain of user profiling based on the web pages visited by the user from web corpuses. We use this method to organize our documents similar to the manner in which they're assigned to nodes of a concept hierarchy. It should be noted that the output of other more sophisticated approaches like the one proposed in (Eirinaki *et al.*, 2004) for generating C-Logs could also be used for this purpose without affecting our general RL model.

Conceptual States and Actions

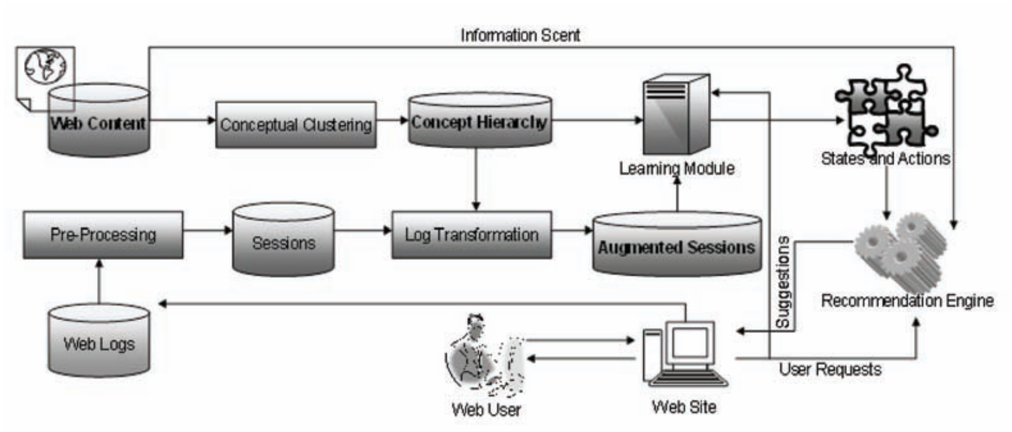
After clustering the web pages in the hierarchy, our state and action definition change as follows. Instead of keeping a sequence V of individual page visits by the user, each state would consist of a sequence of concepts visited by the user. Considering a mapping like $C : P \rightarrow H$ which transforms each page *p* in the set of pages *P* to the corresponding concept *c* in the concept hierarchy *H*, the states *s* in each time step *t* would now be defined as:

$$\begin{aligned} V_s &= < C(v_{t-w+1}), C(v_{t-w+2}), \ldots, C(v_t) > \\ R_s &= < C(r_{t-w'+1}), C(r_{t-w'+2}), \ldots, C(r_t) > \end{aligned} \qquad (8)$$

Also, the actions are now recommendation of pages that belong to a specific concept. In order to do so we need a module to find the node each page belongs to in the concept hierarchy and transform each usage log to a sequence of concepts in the training phase. The other aspects of the system like the reward function and the learning process would remain the same, e.g. an action a recommending a concept c is rewarded if the user visits a page belonging to concept c later in his browsing session.

This definition results in a much smaller state-action space as now the state space size is dependant on the number of distinct page clusters instead of the number of distinct web pages in the website. Consequently, the learning process will become more efficient and the system will have a more general model of users' browsing behavior on the site. With this generalized definition, the chance of confronting an unseen state will be much less

Figure 7. Architecture of the Hybrid Recommender System



and actually minimized as our evaluation results show. We'll no longer make decisions based on weak usage patterns as now the states represent a generalized view of many single visit sequences, and the average number of times a state is visited in user sessions is now 10.2 times the average visit of states in the usage-based setting. A general view of the system is depicted in Figure 7.

In the test phase, the user's raw session will be converted to a semantic session, the corresponding state will be found and the page cluster with the highest value is identified. When a concept is chosen as the action, the next step would be to recommend pages from the chosen cluster(s). Initially we chose to recommend the pages with a probability corresponding to their similarity to the cluster mean vector. This new definition of actions enables the system to cover a wider range of pages to be recommended as our evaluations show, and also the potential ability of avoiding the "new item" problem as any new page will be categorized in the appropriate cluster and have a fair chance of being recommended.

A Content-Based Reward Function

We can also make use of the content information of web pages and their relative positioning in the concept hierarchy in our reward function. The new rewarding function takes the content similarity of the recommended and visited pages into account. The basic idea behind this method is to reward recommendation of a concept c in s_i which might not be visited in s_{i+1} but is semantically similar to the visited page v, or more precisely, to the concept that v belongs to. The new reward function would be basically same as the one presented in Algorithm 1, the only difference is that instead of using *rew(s,a)* in step 5, now the reward would be computed by the new function *HybridRew(s,a)* shown in Equation 9.

$$HybridRew(s,a)= UBR(Dist(R_{s'}, a),Time(v_{t+1}))\times CBR(a, v_{t+1}) \quad (9)$$

Here *CBR* represents the content-based reward of an action and *UBR* is the usage based reward which is our previous reward function used in step 5 of Algorithm 1.

In order to compute the content based reward we use the method for computing similarity of nodes in a concept hierarchy proposed in (Bose *et al.*, 2006). In this method, first a probability $p(c)$ is assigned to each concept node c which is proportional to the frequency of pages belonging to this node and its descendants in user sessions. The information content of each node is then defined as:

$$I(c) = -\log p(c) \tag{10}$$

Then a set *LCA* is found which contain the Least Common Ancestors, those occurring at the deepest level, of the pair of concept nodes. And the similarity score between those are computed as:

$$Sim(c_1, c_2) = \max_{a \in LCA}\{I(a)\} \tag{11}$$

The *CBR* for each recommended page *a* will be equal to the similarity score:

$$CBR(a, v_{t+1}) = Sim(C(a), C(v_{t+1})) \tag{12}$$

This method seems specifically appropriate for the off-line training phase where recommendations are evaluated using the web usage logs. In this phase actions are predictions of the user's next visit and web pages are not recommended to the user in the on-line browsing sessions. As a result actual user reactions towards pages cannot be assessed and the assumption is made that users interest toward a recommendation can be estimated as function of conceptual similarity between the recommended and visited pages.

The situation is a bit different when the system provides on-line recommendations to the user. Here the usage-based reward is given more weight than the reward based on content similarity. This is based on the idea that the overall judgment of users can be trusted more than the content similarity of pages, since satisfying user information need is the ultimate goal of personalization.

Selection of Pages in a Concept

Based on the actions, we can decide which concept the user is interested in. In order to make recommendations, we should select a page belonging to that concept, which is not a trivial task especially when we're faced with large clusters of pages. Our initial solution was to rank pages with respect to their distance from cluster center. Our experiments show that this method does not yield in accurate recommendations. In order to enhance our method we exploited the content information of web pages and the hyperlinks that the users have followed in each state. The text around the chosen hyperlinks in each page has been used as an indicator of user information need in user modeling, based on the information scent model (Chi *et al.*, 2001). We also employ the information scent to compute a vector representing user information need in each state. The method we used is basically similar to (Chi *et al.*, 2001), using the text around the hyperlink, the title of the out going page etc., with the exception that we assign more weight to the hyperlinks followed later in each state. After computing this vector we use the cosine based similarity to find the most relevant pages in each selected page cluster for recommendation.

Overall, we experimented with three different methods for ranking pages for selection from a given concept *c'* (pages with lower ranks have higher probability of being selected):

1. Ranking based on the distance of a page from Cluster Mean (*HCM*): The basic idea here is that pages which are closer to the cluster mean vector are more relevant to the given concept and hence might be more relevant to a user interested in that concept. Considering $\boldsymbol{W}_{c'}$ as the mean content vector of concept *c'*, and the vector $\boldsymbol{W}_i$, representing each web page p_i ($p_i \in P$ and $C(p_i) = c'$), the selection rank of each p_i, shown by $SelRank_{CM}(p_i)$, would be computed according to (14). This rank is in reverse relation with the distance of $\boldsymbol{W}_i$ from $\boldsymbol{W}_{c'}$. In these experiments we computed the distance using the cosine of these two vectors.

$$SelRank_{CM}(p_i) \le SelRank_{CM}(p_j) \Leftrightarrow Dist(W_{C'}, W_i) \le Dist(W_{C'}, W_j) \tag{13}$$

2. Ranking based on the occurrence frequency of a page (*HFreq*): this method is primarily based on historical usage data. The rationale is that pages which are more frequently visited by users might more popular in the collection of pages related to a concepts and therefore more probable to be sought by the target user. Considering $Frq(p_i)$ as the occurrence frequency of each p_i ($p_i \in P$ and $C(p_i) = c'$), the selection rank of each p_i, shown by $SelRank_{Freq}(p_i)$, would be in reverse relation with the distance this frequency.

$$SelRank_{Freq}(p_i) \leq SelRank_{Freq}(p_j) \Leftrightarrow Frq(p_i) \geq Frq(p_j) \tag{14}$$

3. Ranking based on the Information Scent model (*HIS*): in this approach, based on the information foraging theory, it is assumed that the information need of the user van be estimated by the proximal cues that the user follows in his navigation on the web. Here, pages are ranked in accordance to their similarity to the vector derived by the information scent model from the sequence of pages visited in each state. Considering $\boldsymbol{W}_{IS}$ as the information scent vector, and the vector $\boldsymbol{W}_i$, representing each web page p_i ($p_i \in P$ and $C(p_i) = c'$), the selection rank of each p_i, shown by $SelRank_{IS}(p_i)$, would be computed according to (15). This rank is in accordance with the similarity of $\boldsymbol{W}_i$ to $\boldsymbol{W}_{IS}$. In our experiments, the similarity of two vectors was computed using the cosine-based similarity function commonly used in information retrieval.

$$SelRank_{IS}(p_i) \leq SelRank_{IS}(p_j) \Leftrightarrow Sim(W_{IS}, W_i) \geq Sim(W_{IS}, W_j) \tag{15}$$

EXPERIMENTAL EVALUATION OF THE HYBRID RECOMMENDATION METHOD

Evaluation Metrics

We pointed out the main weaknesses of the usage-based method in the previous section and proposed the hybrid approach as a solution to overcome these shortcomings. In order to assess the success of the proposed method in this regards, we need metrics that directly address these characteristics of the system. Thus, metrics beyond the ones used in evaluation of the usage-based method in the previous section should be used. We used the following metrics for this purpose, many of which were used by Bose *et al.* (2007). We also used some modifications of these metrics as needed. The metrics used are:

- **Recommendation Accuracy** *(RA)*: Percentage of correct recommendations between all the recommendations made by the system. A correct recommendation is, as before, a specific recommended web page that the user chooses to visit. These recommendations are generated in the hybrid approach by applying one of the page selection methods.
- **Predictive Ability** *(PA)*: Percentage of pages recommended at least once. Bose *et al.* (2007) mention this metric as one that measures how useful the recommendation algorithm is.
- **Prediction Strength** *(PS)*: Measures the average number of recommendation the system makes in each state (for each sequence of page visits). This metric aims at evaluating the ability of the recommender in generating recommendations for various scenarios of user behavior. It can specially reflect the performance of the system in the presence of the "new state" problem.

- **Shortcut Gain***(SG)*: average percentage of pages skipped because of recommendations. This is the same metric we used to evaluate the usage-based approach.
- **Recommendation Quality***(RQ):* average rank of a correct recommendation in the recommendation lists. This metric emphasizes the importance of ranking pages for recommendations (somehow similar to the manner in which ranking is valued in the results returned by a search engine).

Sensitivity to Visited Sequence Window Size

The first experiments were performed to evaluate system sensitivity to the size of visited concept sequence V in our states. To evaluate the choice of different window sizes, regardless of other parameters e.g. the page selection method, we used a new metric called *Concept recommendation Accuracy (CRA)* and *Concept Predictive Ability (CPA)* which are based on recommendation and visit of concepts instead of pages. For example, a recommendation of concept c_1 is considered successful if the user later visits any page p belonging to c_1, i.e. $C(p)=c_1$. Our evaluations indicate the best performances are achieved when using window sizes of 3 and 4 (Table 3). This is due to the fact that smaller values of w keep insufficient information about navigation history and larger values of w result in states that are numerous and less frequently visited, as the average session length in our data is 8.6. We choose w=3 in the rest of the experiments as it results in smaller number of states with a negligible decrease in accuracy.

Table 3. Comparison of different window sizes in the hybrid approach

Window Size	Metric		
	CRA	CPA	RQ
1	42	76	1.88
2	63	81	3.21
3	79.50	96	2.80
4	81.30	98	2.11
5	66.66	95	3.78

Comparison with Other Methods

We compared the proposed method with the previous usage-based approach *(UB-RL)* and a content-based approach that uses the info scent model to recommend pages from the whole website *(CIS)*. The latter method was used because of the promising results achieved in the system while using the page selection method based on information scent. Note that *UB-RL* has shown superior results than common usage-based methods, and is considered as the baseline usage-based method we aim to improve. We used three different methods for page selection in our hybrid approach: based on the distance from cluster mean *(HCM)*, using the frequency of occurrence in user sessions *(HFreq)* and the one based on Information Scent *(HIS)*. We also compared our method to a state of the art recommendation method proposed by Bose *et al.* (2007). This method makes use of concept hierarchies and sequence alignment methods in order to cluster user sessions and making recommendations based on the resulted clusters. It is abbreviated by *HSA* in the results. The results presented here are based different experiments of having 3, 5 and 10 as the maximum number of recommendations in each stage (length of the recommendation list).

An issue worth considering is that based on the experiments performed in the previous section (sensitivity to the *V* sequence), we have an upper bound estimation of the performance of our hybrid recommendation methods. For example, the *CRA* achieved by the system is the maximum *RA* the hybrid methods can achieve. Since now the methods have to select a specific page from a concept and we know the ability of the system in predicting the correct concept is limited by *CRA*. In fact, these results can be used to compare the performance of various page selection methods in the hybrid approaches.

As our evaluation shows (Table 4), *HIS* out performs the rest of the methods except with respect to *RA*, compared to *UB*-RL. Note that the *UB-RL* method shows a much lower *PA*, as it's a purely usage-based approach. An initial glance on the results can show the success of our hybrid methods in overcoming the shortcomings of the usage-based approach, especially in the sense of *PA* and *PS* metrics (both significant at $p<0.001$ on the *t*-test). Our hybrid approaches, especially *HIS* and *HFreq*, can also outperform the state of the art *HSA* recommendation method in almost every situation, although the better performance is marginal and less significant in *PS* measure, it is more significant on *PA* ($p<0.01$) and more emphasized and also statistically significant on RA, SG and RQ (all with $p<0.001$ on the *t*-test). The results achieved when using different lengths for recommendation lists almost show the same relative performance from different recommendation methods, while some features of the methods are more emphasized in higher or lower number of recommendations which we'll point out in the rest of this section. One important issue in analyzing the evaluation results is considering the logical dependencies that exist between various evaluation metrics, e.g. between *PS* and *RQ*. Considering dependencies, naturally there's not a single recommendation method that outperforms the rest with respect to *all* evaluation metrics. What should be noted is the importance of evaluating recommendation methods based their overall performance in all the evaluation metrics and also considering their relative performance in dependent evaluation metrics. As we will investigate further in the following subsections, we conclude from these results that our two hybrid approaches *HIS* and *HFreq* show an overall superior performance compared to the other methods and could be considered our suggestions for further development and implementation in real world applications, especially the *HIS* method which is the superior method in the majority of the metrics and the usually the second best in the rest. We will discuss the performance of various recommendation methods with respect to each metric in the following sub sections.

Predictive Ability

It can also be seen that all the hybrid approaches can achieve better predictive ability than the content based recommendation method *CIS* (significant at $p<0.001$ on the *t*-test). This issue is more emphasized when using shorter recommendation lists. This shows that semantic grouping of the web pages and then recommending a page from the *correct* can actually increase the chance of each page to be recommended appropriately. While, the *CIS* method which considers the whole set of pages as the search space is less successful in covering the web site.

Predictive Strength

Regarding the prediction strength metric, the *UB-RL* method is the weakest recommendation method, as expected. Various reasons for this phenomena such as the "new state" problem were mentioned in the previous section. On the other hand, the purely content-based *CIS* approach can achieve the perfect *PS* performance as there have always been some pages with some minimum similarity with the resulted content model. This can be an intrinsic characteristic of each content-based method, when not considering a lower bound on similarity. It should be noted that beside the number of recommendations shown by the *PS* value, the quality of the recommendation list is also of uttermost importance. In this regard, our hybrid approaches are able to achieve better results in almost every evaluation metric, while also achieving a *PS* very close to the optimal *CIS* approach. For example the *HIS* method achieves a 36% increase in compared to the baseline *UB-RL* method which is also statistically significant ($p<<0.001$). These results illustrate the strength

Table 4. Comparison of different recommendation methods

Method	Metric				
	RA	PA	PS	SG	RQ
Max. Number of Recommendation in Each Step=3					
UB-RL	**53.76**	51.06	2.73	10.26	**1.96**
CIS	35.09	67.12	2.99	7.01	2.54
HSA	45.11	93.60	2.97	21.07	2.26
HCM	38.09	91.01	**2.97**	11.33	2.39
HFreq	46.34	93.40	**2.97**	**24.11**	2.19
HIS	**51.66**	**94.10**	**2.97**	**22.31**	**2.15**
Max. Number of Recommendation in Each Step=5					
UB-RL	**49.81**	55.06	3.64	13.17	**2.21**
CIS	32.11	69.29	4.98	7.21	3.90
HSA	40.01	96.91	4.95	24.21	3.58
HCM	33.09	91.67	**4.96**	12.56	3.76
HFreq	42.12	95.91	**4.96**	**26.80**	3.11
HIS	**46.28**	**97.20**	**4.96**	**25.76**	**2.89**
Max. Number of Recommendation in Each Step=10					
UB-RL	**44.91**	58.15	5.69	14.17	**2.79**
CIS	29.82	73.12	9.96	8.14	8.11
HSA	34.31	96.91	9.79	25.18	6.27
HCM	30.09	92.23	9.82	14.62	6.97
HFreq	39.36	96.31	9.82	27.77	5.94
HIS	**42.17**	**97.20**	**9.84**	**27.95**	**5.33**

of the generalized models of user behavior, employed in the hybrid approaches, in capturing user behavior patterns and avoiding unseen navigation scenarios at a higher level of abstraction resulted from the generalized state and action model.

Recommendation Accuracy

While the *UB-RL* method receives the highest accuracy as expected, our proposed hybrid approaches *HIS* and *HFreq* are the second bests in almost every case with a rather small difference. This performance is especially important due to the fact that the hybrid approaches have lost the information at the detail level of page visits because of their generalized view of user behavior. Like any generalization this information loss is supposed to come inevitably with some loss in model accuracy. These results show the success of the page selection methods employed in *HFreq* and *HIS* and the importance of this selection. The rather low *RA* value achieved by *HCM* indicates the importance of page selection method in the process. It is also an indicator of the existing trade-off between generalized and detailed knowledge. As we can see this approach has a high *CRA* value (Table 3), but because of the information loss occurred at a higher level of abstraction and lack of an appropriate page selection method (at lower level of abstraction), it performs even worse than *CFreq* which is based on a rather simple metric, i.e. popularity of a page. The weaker performance of *CIS* (statistically significant at $p<0.001$) might be considered as further evidence in support of the importance of usage patterns in accurate inference of user information needs.

Shortcut Gain

Regarding the shortcut gain metric, the content-based *CIS* approach which makes no use of usage information receives the weakest results. The usage-based UB-RL method is able to achieve better shortcut gain in recommendations and *HIS* and *HFreq* hybrid recommendation methods achieve the best results in this regard (significant at $p<0.001$). The weaker performance of *HCM* in comparison to *UB-RL* is again due to the inappropriate page selection method in *HCM*, although it still manages to beat *CIS*, because of having a usage-based component. An interesting point is the ability of HIS and HFreq to achieve an increase of almost 100% in comparison to the usage-based approach. Of course, it should be mentioned that beside the higher accuracy and diversity of recommendations generated by these methods, the greater number of recommendation (*PS*) is also an effective factor in this regard.

Recommendation Quality

This metric shows the rank of correct recommendations in the recommendation lists. It can be seen that the *UB-RL* receives the best results in this regard, while our hybrid approaches are second bests and the content based approach is the weakest. The difference between the usage-based and the hybrid approaches is marginal in almost every case. One important issue is the logical dependency between the *RQ* and the *PS* metrics. Naturally, a recommender that makes fewer recommendations is more likely to achieve lower *RQ* values, e.g. a recommender that does not make more than 2 recommendations will definitely have $RQ \leq 2$. In fact, it is more appropriate to consider *RQ* in respect to the *PS* metric, e.g. the ratio *RQ/PS*. Considering this, we can see that the *HIS* method has better performance between all recommendation methods used in the experiments (significant at $p<0.001$ compared to all the baseline methods).

CONCLUSION AND FUTURE WORKS

In this chapter we presented novel web page recommendation methods based on reinforcement learning. First a usage-based method for web recommendation was proposed, which was based the reinforcement learning paradigm. This system learns to make recommendations from web usage data as the actions it performs in each situation rather than discovering explicit patterns from the data. We modeled web page recommendation as a Q-Learning problem and trained the system with common web usage logs. System performance was evaluated under different settings and in comparison with other methods. Our experiments showed promising results achieved by exploiting reinforcement learning in web recommendation based on web usage logs.

Afterwards we described a method to enhance our solution based on reinforcement learning, devised for web recommendations from web usage data. We showed the restrictions that a usage-based system inherently suffers from (e.g. low coverage on items, inability to generalize, etc.) and demonstrated how combining conceptual information regarding the web pages can improve the system. Our evaluation results show the flexibility of the proposed RL paradigm to incorporate different sources of information and to improve overall the quality of recommendations.

There are other alternatives that can potentially improve the system and constitute our future work. In the case of the reward function used, various implicit feedbacks from the user rather than just the fact that the user had visited the page can be used, such as those proposed in (Zhang & Seo, 2001). Another option is using a more complicated reward function rather than the linear combination of factors; a learning structure such as neural networks is an alternative. The hybrid method can also be extended in various ways. One is to find more sophisticated methods for organizing a website into a concept hierarchy. More accurate methods of assessing implicit feed-back can also be used to derive a more precise reward function. Integration of other sources of domain knowledge e.g. website topology or a domain-ontology into the model can also be another future work for this research. Finally, devising a model to infer higher level goals of user browsing, similar to the work done in categorizing search activities can be another future direction.

REFERENCES

Bose, A., Beemanapalli, K., Srivastava, J., & Sahar, S. (2006). Incorporating concept hierarchies into usage mining based recommendations. In O. Nasraoui, M. Spiliopoulou, J. Srivastava, B. Mobasher, B. M. Masand (Eds.), *Advances in Web Mining and Web Usage Analysis, 8th International Workshop on Knowledge Discovery on the Web*, *Lecture Notes in Computer Science 4811* (pp. 110-126). Berlin, Heidelberg, Germany: Springer.

Breese, J., Heckerman, S., & Kadie, C. (1998, July). Empirical analysis of predictive algorithms for collaborative filtering. In G. F. Cooper, S. Moral (Eds.), *UAI '98: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* (pp. 43-52). University of Wisconsin Business School, Madison, Wisconsin, USA: Morgan Kaufmann.

Burke, R. (2000). Knowledge-based recommender systems. In A. Kent (Ed.), *Encyclopedia of Library and Information Systems, 69*. New York: Marcel Dekker.

Burke, R. (2002). Hybrid recommender systems: survey and experiments. *User Modeling and User-Adapted Interaction*, *12*(4), 331–370. doi:10.1023/A:1021240730564

Chi, E. H., Pirolli, P., & Pitkow, J. (2001).Using information scent to model user information needs and actions on the web. *Proceedings of the ACM SIG-CHI on Human Factors in Computing Systems* (pp.490-497). Seattle, WA, USA: ACM Press.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining World Wide Web browsing patterns. *Knowledge and Information Systems*, *1*(1), 5–32.

Deshpande, M., & Karypis, G. (2004). Item-based top-N recommendation algorithms. *ACM Transactions on Information Systems*, *22*(1), 143–177. doi:10.1145/963770.963776

Eirinaki, M., Lampos, C., Paulakis, S., & Vazirgiannis, M. (2004). Web personalization integrating content semantics and navigational patterns. In A. H. Laender, D. Lee, M. Ronthaler (Eds.), *Proceeding of the Sixth ACM CIKM International Workshop on Web Information and Data Management* (pp.72-79), Washington, DC, USA: ACM Press.

Eirinaki, M., Vazirgiannis, M., & Varlamis, I. (2003). SEWeP: using site semantics and a taxonomy to enhance the web personalization process. In L. Getoor, T. E. Senator, P. Domingos, C. Faloutsos (Eds.), *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 99-108), Washington, DC, USA: ACM Press.

Fu, X., Budzik, J., & Hammond, K. J. (2000). Mining navigation history for recommendation. In *IUI 2000: Proceedings of the 5th International Conference on Intelligent User Interface* (pp. 106-112). New Orleans, LA, USA: ACM Press.

Godoy, D., & Amandi, A. (2005). Modeling user interests by conceptual clustering. *Information Systems*, *31*(4-5), 245–267.

Golovin, N., & Rahm, E. (2004). Reinforcement learning architecture for web recommendations. In *Proceeding of the International Conference on Information Technology: Coding and Computing, 1,* 398-403. Las Vegas, Nevada, USA: IEEE Computer Society.

Herlocker, J., Konstan, J., Brochers, A., & Riedel, J. (2000). An Algorithmic Framework for Performing Collaborative Filtering. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 230-237). Berkeley, CA, USA: ACM Press.

Joachims, T., Freitag, D., & Mitchell, T. M. (1997). Web Watcher: A tour guide for the World Wide Web. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence* (pp. 770-777). Nagoya, Japan: Morgan Kaufmann.

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L. R., & Riedl, J. (1997). GroupLens: applying collaborative filtering to Usenet news. *Communications of the ACM*, *40*(3), 77–87. doi:10.1145/245108.245126

Li, J., & Zaiane, O. R. (2004). Combining usage, content and structure data to improve web site recommendation. In K. Bauknecht, M. Bichler, B. Pröll (Eds.), *Proceeding of 5th International Conference E-Commerce and Web Technologies, Lecture Notes in Computer Science 3182* (pp. 305-315). Berlin, Heidelberg, Germany: Springer.

Mahmood, T., & Ricci, F. (2007, August). Learning and adaptivity in interactive recommender systems. In M. L. Gini, R. J. Kauffman, D. Sarppo, C. Dellarocas, & F. Dignum (Eds.), *Proceedings of the 9th International Conference on Electronic Commerce: The Wireless World of Electronic Commerce* (pp. 75-84). University of Minnesota, Minneapolis, MN, USA: ACM Press.

Mitchell, T. (1997). *Machine Learning*. New York, NY: McGraw-Hill.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on Web usage mining. *Communications of the ACM*, *43*(8), 142–151. doi:10.1145/345124.345169

Mobasher, B., Dai, H., Luo, T., Sun, Y., & Zhu, J. (2000). Integrating web usage and content mining for more effective personalization. In K. Bauknecht, S. K. Madria, G. Pernul (Eds.), *Proceeding of First International Conference E-Commerce and Web Technologies, Lecture Notes in Computer Science 1875* (pp. 165–176). Munich, Germany: Springer.

Nakagawa, M., & Mobasher, B. (2003). A hybrid web personalization model based on site connectivity. In R. Kohavi, B. Liu, B. Masnad, J. Srivastava, O. R. Zaiane (Eds.), *Web Mining as a Premise to Effective and Intelligent Web Applications, Proceedings of the Fifth International Workshop on Knowledge Discovery on the Web* (pp. 59-70). Washington DC, WA, USA: Quality Color Press.

Pazzani, M. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, *13*(5-6), 393–408. doi:10.1023/A:1006544522159

Pazzani, M., & Billsus, D. Content-based recommendation systems. In P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.), *The Adaptive Web: Methods and Strategies of Web Personalization, Lecture Notes in Computer Science 4321* (pp. 325-341). Berlin, Heidelberg, Germany: Springer-Verlag.

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, *40*(3), 56–58. doi:10.1145/245108.245121

Shany, G., Heckerman, D., & Barfman, R. (2005). An MDP-based recommender system. *Journal of Machine Learning Research*, *6*(9), 1265–1295.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explorations*, *1*(2), 12–23. doi:10.1145/846183.846188

Srivihok, A., & Sukonmanee, V. (2005). E-commerce intelligent agent: personalization travel support agent using Q-Learning. In Q. Li, & T. P. Liang (Eds.): *Proceedings of the 7th International Conference on Electronic Commerce* (pp. 287-292). Xi'an, China: ACM Press.

Sutton, R. S., & Barto, A. G. (1998) *Reinforcement Learning: An Introduction*, Cambridge, MA, USA: MIT Press.

Taghipour, N., & Kardan, A. (2007, September). Enhancing a recommender system based on Q-Learning. In A. Hinneburg (Ed.), *LWA 2007: Lernen - Wissen - Adaption, Workshop Proceedings, Knowledge Discovery, Data Mining and Machine Learning Tack,* (pp. 21-28). Halle, Germany: Martin-Luther-University Publications.

Taghipour, N., & Kardan, A. (2008, March). A hybrid web recommender system based on Q-Learning. In R. L. Wainwright, & H. Haddad (Eds.), *Proceedings of the 2008 ACM Symposium on Applied Computing* (pp. 1164-1168). Fortaleza, Brazil: ACM Press.

Taghipour, N., Kardan, A., & Shiry Ghidary, S. (2007, October). Usage-based web recommendations: a reinforcement learning approach. In J. A. Konstan, J. Riedl, & B. Smyth (Eds.), *Proceedings of the First ACM Conference on Recommender Systems* (pp. 113-120). Minneapolis, MN, USA: ACM Press.

Wasfi, A. M. (1999). Collecting User Access Patterns for Building User Profiles and Collaborative Filtering. In: *IUI '99: Proceedings of the 4th International Conference on Intelligent User Interfaces* (pp. 57-64).

Zhang, B., & Seo, Y. (2001). Personalized web-document filtering using reinforcement learning. [Los Angels, CA, USA: ACM Press.]. *Applied Artificial Intelligence*, *15*(7), 665–685. doi:10.1080/088395101750363993

This work was previously published in Collaborative and Social Information Retrieval and Access: Techniques for Improved User Modeling, edited by M. Chevalier; C. Julien; C. Soule-Dupuy, pp. 222-249, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 8.6
On the Use of Soft Computing Techniques for Web Personalization

G. Castellano
University of Bari, Italy

A. M. Fanelli
University of Bari, Italy

M. A. Torsello
University of Bari, Italy

ABSTRACT

Due to the growing variety and quantity of information available on the Web, there is urgent need for developing Web-based applications capable of adapting their services to the needs of the users. This is the main rationale behind the flourishing area of Web personalization that finds in soft computing (SC) techniques a valid tool to handle uncertainty in Web usage data and develop Web-based applications tailored to user preferences. The main reason for this success seems to be the synergy resulting from SC paradigms, such as fuzzy logic, neural networks, and genetic algorithms. Each of these computing paradigms provides complementary reasoning and searching methods that allow the use of domain knowledge and empirical data to solve complex problems. In this chapter, we emphasize the suitability of hybrid schemes combining different SC techniques for the development of effective Web personalization systems. In particular, we present a neuro-fuzzy approach for Web personalization that combines techniques from the fuzzy and the neural paradigms to derive knowledge from Web usage data and represent the knowledge in the comprehensible form of fuzzy rules. The derived knowledge is ultimately used to dynamically suggest interesting links to the user of a Web site.

DOI: 10.4018/978-1-60566-024-0.ch018

INTRODUCTION

The growing explosion in the amount of information and applications available on the World Wide Web has made more severe the need for effective methods of personalization for the Web information space. The abundance of information combined with the heterogeneous nature of the Web makes Web site exploration difficult for ordinary users, who often obtain erroneous or ambiguous replies to their requests. This has led to a considerable interest in Web personalization which has become an essential tool for most Web-based applications. Broadly speaking, Web personalization is defined as any action that adapts the information or services provided by a Web site to the needs of a particular user or a set of users, taking advantage of the knowledge gained from the users' navigational behavior and individual interests, in combination with the content and the structure of the Web site. In other words, the aim of a Web personalization system is to provide users with the information they want or need, without expecting them to ask for it explicitly (Nasraoui, 2005; Mulvenna, Anand, & Buchner, 2000).

The personalization process covers a fundamental role in an increasing number of application domains such as e-commerce, e-business, adaptive Web systems, information retrieval, and so forth. Depending on the application context, the nature of personalization may change. In e-commerce applications, for example, personalization is realized through recommendation systems which suggest products to clients or provide useful information in order to decide which products to purchase (Adomavicius & Thuzilin, 2005; Baraglia & Silvestri, 2004; Cho & Kim, 2004; Mobasher, 2007b; Schafer, Konstan, & Riedl, 2001). In e-business, Web personalization additionally provides mechanisms to learn more about customer needs, identify future trends, and eventually increase customer loyalty to the provided service (Abraham, 2003). In adaptive Web sites, personalization is intended to improve the organization and presentation of the Web site by tailoring information and services so as to match the unique and specific needs of users (Callan, Smeaton, Beaulieu, Borlund, Brusilovsky, Chalmers et al., 2001; Frias-Martinez, Magoulas, Chen, & Macredie, 2005). In practice, adaptive sites can make popular pages more accessible, highlight interesting links, connect related pages, and cluster similar documents together (Perkowitz & Etzioni, 1997). Finally, in information retrieval, personalization is regarded as a way to reflect the user preferences in the search process so that users can find out more appropriate results to their queries (Kim & Lee, 2001; Enembreck, Barthès, & Ávila, 2004).

The development of Web personalization systems gives rise to two main challenging problems: how to discover useful knowledge about the user's preferences from the uncertain Web data and how to make intelligent recommendations to Web users. A natural candidate to cope with such problems is **soft computing** (SC), a consortium of computing paradigms that work synergistically to exploit the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth in order to provide flexible information processing capabilities and obtain low-cost solutions and close resemblance to human-like decision making. Recently, the potentiality of SC techniques (i.e., neural networks, fuzzy systems, genetic algorithms, and combinations of these) in the realm of Web personalization has been explored by researchers (e.g., Jespersen, Thorhauge, & Pedersen, 2002; Pal, Talwar, & Mitra, 2002; Sankar, Varun, & Pabitra, 2002; Yao, 2005).

This chapter is intended to provide a brief survey of the stat-of-art SC approaches in the wide domain of Web personalization, with special focus on the use of hybrid techniques. As an example, we present a neuro-fuzzy Web personalization framework. In such a framework, a hybrid approach based on the combination of techniques taken from the fuzzy and the neural paradigms is employed in order to identify **user profiles** from Web usage data and to provide dynamical predictions about Web pages to be suggested to the current user, according to the user profiles previously identified.

The content of chapter is organized as follows. In Section 2 we deal in depth with the topic of Web personalization, focusing on the use of Web usage mining techniques for the development of Web applications endowed with personalization functions. Section 3 motivates the use of soft computing techniques for the development of Web personalization systems and overviews existing systems for Web personalization based on SC methods. In Section 4 we describe a neuro-fuzzy Web personalization framework and show its application to a Web site taken as case study. Section 5 closes the chapter by drawing conclusive remarks.

WEB PERSONALIZATION

Web personalization is intended as the process of adapting the content and/or the structure of a Web site in order to provide users with the information they are interested in (Eirinaki & Vazirgiannis 2003; Mulvenna et al., 2000; Nasraoui 2005). The personalization of services that a Web site may offer is an important step towards the solution of some problems inherent in Web information space, such as alleviating information overload and making the Web a friendlier environment for its individual user, and, hence, creating trustworthy relationships between the Web site and the visitor-customer. Mobasher, Cooley, and Srivastava (1999) simply define Web personalization as the task of making Web-based information systems adaptive to the needs and interests of individual users. Typically, a personalized Web site recognizes its users, collects information about their preferences, and adapts its services in order to match the users' needs. Web personalization improves the Web experience of a visitor by presenting the information that the visitor wants to see in the appropriate manner and at the appropriate time.

In literature, many different approaches have been proposed for the design and the development of systems endowed with personalization functionality (Kraft, Chen, Martin-Bautista, & Vila, 2002; Linden, Smith, & York, 2003; Mobasher, Dai, Luo, & Nakagawa, 2001;). In the majority of the existing commercial personalization systems, the personalization process involves substantial manual work and, most of the time, significant effort for the user. A better way to expand the personalization of the Web is to automate the adaptation of Web-based services to their users. Machine learning methods have a successful record of applications to similar tasks, that is, automating the construction and adaptation of information systems (Langley, 1999; Pohl, 1996; Webb, Pazzani, & Billsus, 2001). Furthermore, the integration of machine learning techniques in larger process models, such as that of **knowledge discovery** in data (KDD or data mining), can provide a complete solution to the adaptation task. Data mining has been used to analyze data collected on the Web and extract useful knowledge leading to the so-called Web mining (Eirinaki & Vazirgiannis, 2003; Etzioni, 1996; Kosala & Blockeel, 2000; Mobasher, 2007a; Pal et al., 2002). Web mining refers to a special case of data mining which deals with the extraction of interesting and useful knowledge from Web data. Three important subareas can be distinguished in Web mining:

- *Web content mining*: Extraction of knowledge from the content of Web pages (e.g., textual data included in a Web page such as words or also tags, pictures, downloadable files, etc.).
- *Web structure mining*: Extraction of knowledge from the structural information present into Web pages (e.g., links to other pages).
- *Web usage mining*: Extraction of knowledge from usage data generated by the visits of the users to a Web site. Generally,

usage data are collected into Web **log files** stored by the server whenever a user visits a Web site.

In this chapter, we focus mainly on the field of Web usage mining (WUM) that represents today a valuable source of ideas and solutions for the development of Web personalization systems. Overviews about the advances of research in this field are provided by several other authors (e.g., Abraham, 2003; Araya et al., 2004; Cho & Kim, 2004; Cooley, 2000; Facca and Lanzi, 2005; Mobasher, 2006, 2005; Mobasher, Nasraoui, Liu, & Masand, 2006; Pierrakos, Paliouras, Papatheodorou, & Spyropoulos, 2003). In general, regardless the application context, three main steps are performed during a WUM personalization process (Mobasher, Cooley, & Srivastava, 2000):

- **Preprocessing:** Web usage data are collected and preprocessed in order to identify **user sessions** representing the navigational activities of each user visiting a Web site.
- **Knowledge discovery**: The session data representing the users' navigational behaviour are analysed in order to discover useful knowledge about user preferences in the form of user categories of user profiles.
- **Recommendation**: The extracted knowledge is employed to customize the Web information space to the necessities of users, that is, to provide tailored recommendations to the users depending on their preferences.

While preprocessing and knowledge discovery are performed in an off-line mode, the employment of knowledge for recommendation is carried out in real time to mediate between the user and the Web site the user is visiting. In the following subsections, each step of the personalization process is more deeply examined.

Preprocessing

Access log files represent the most common source of Web usage data. All the information concerning the accesses made by the users to a Web site are stored in log files in chronological order. According to the common log format (www.w3.org/Daemon/User/Config/Loggin.htm#common-logfile-format) each log entry refers to a page request and includes information such as the user's IP address, the request's date and time, the request method, the URL of the accessed page, the data transmission protocol, the return code indicating the status of the request, and the size of the visited page in terms of number of bytes transmitted. By exploiting such information, models of typical user navigational behavior can be derived and used as input to the next step of knowledge discovery. The derivation of navigational patterns from log data is achieved through a preprocessing activity that filters out redundant and irrelevant data, and selects only log entries related to explicit requests made by users. Cooley (2000) extensively discusses the methods adopted to execute data preparation and preprocessing activity. Typically Web data preprocessing includes two main tasks, namely, data cleaning and user session identification.

The aim of data cleaning is to remove from log files all records that do not represent the effective browser activity of the connected user, such as those corresponding to requests for multimedia objects embedded in the Web page accessed by the user. Elimination of these items can be reasonably accomplished by checking the suffix of the URL name (all log entries with filename suffixes such as gif, jpeg, GIF, JPEG, jpg, JPG and map are removed). Also, records corresponding to failed user requests and accesses generated by Web robots are identified and eliminated from log data. Web robots (also known as Web crawlers or Web spiders) are programs which traverse the Web in a methodical and automated manner, downloading complete Web sites in order to update the index of a search engine. This task is performed by maintaining a list of known spiders and through heuristic identification of Web robots. Tan and Kumar (2002) propose a robust technique which is able to detect, with a high accuracy, Web robots by using a set of relevant features extracted from access logs (e.g., percentage of media files requested, percentage of requests made by HTTP methods, average time between requests, etc.).

The next task of Web log preprocessing is the identification of user sessions. Based on the definitions found in different works of scientific literature, a user session can be defined as a finite set of URLs corresponding to the pages visited by a user from the moment the user enters a Web site to the moment the same user leaves it (Suryavanshi, Shiri, & Mudur, 2005). The process of segmenting the activity of each user into sessions, called sessionization, relies on heuristic methods. Spiliopoulou (1999) divides the sessionization heuristics into two basic categories: time-oriented and structure-oriented. Time-oriented heuristics establish a timeout to distinguish between consecutive sessions. The usual solution is to set a minimum timeout and assume that consecutive accesses within it belong to the same session, or set a maximum timeout, where two consecutive accesses that exceed it belong to different sessions. On the other hand, structure-oriented heuristics consider the static site structure or they refer to the definition of conceptual units of work to identify the different user sessions. More recently, Spiliopoulou, Mobasher, Berendt, and Nakagawa (2003) have proposed a framework to measure the effectiveness of such heuristics and the impact of different heuristics on various Web usage mining tasks.

Knowledge Discovery

After preprocessing, the next step of a Web personalization process consists in discovering knowledge from data in the form of user models or profiles embedding the navigational behavior by expressing the common interests of Web visitors. Statistical and data mining techniques have been widely applied to derive models of user navigational behavior starting from Web usage data (Facca & Lanzi 2005; Mobasher, 2005; Pierrakos et al., 2003). In particular, analysis techniques of Web usage data can be grouped into three main paradigms: association rules, sequential patterns, and **clustering** (Han and Kamber (2001) detail an exhaustive review).

Association rules are used to capture relationships among Web pages which frequently appear in user sessions, without considering their access ordering. Typically, an association rule is expressed in the form:“A.html, B.html C.html”which states that if a user has visited page A.html and page B.html, it is very likely that in the same session the same user also visits page C.html. This kind of approach has been used in Joshi, Joshi, and Yesha (2003), and Nanopoulus, Katsaros, and Manolopoulos (2002), while some measures of interest to evaluate association rules mined from Web usage data have been proposed by Huang, Cercone, and An (2002a), and Huang, Ng, Ching, Ng, and Cheung (200a). Fuzzy association rules, obtained by the combination of association rules and fuzzy logic, have been extracted by Wong and Pal (2001).

Sequential patterns in Web usage data detect the set of Web pages that are frequently accessed by users in their visits, considering the order that they are visited. To extract sequential patterns, two main classes of algorithms are employed: methods based on association rule mining and methods based on the use of tree structures and Markov chains. Some well-known algorithms for mining association rules have been modified to obtain sequential patterns. For example, the Apriori algorithm has been properly extended to derive two new algorithms: the AprioriAll and GSP proposed by Huang et al. (2002a) and Mortazavi-Asl (2001). An alternative algorithm based on the use of a tree structure has been presented by Pei, Han, Mortazavi-asl, and Zhu (2000). Tree structures have been also used by Menasalvas, Millan, Pena, Hadjimichael, and Marban (2002).

Clustering is the most widely employed technique to discover knowledge in Web usage data. An exhaustive overview of Web data clustering methods is provided by Vakali, Pokorný, and Dalamagas (2004). Two forms of clustering can be performed on usage data: user-based clustering and item-based clustering.

User-based clustering groups similar users on the basis of their ratings for items (Banerjee & Ghosh, 2001; Heer & Chi, 2002; Huang et al., 2001). Each cluster center is an n-dimensional vector (being n the number of items) where the *i*-th component is the average rating expressed by users in that cluster for the *i*-th item. The recommendation engine computes the similarity of an active user session with each of the discovered user categories represented by cluster centroids to produce a set of recommended items.

Item-based clustering identifies groups of items (e.g., pages, documents, products) on the basis of similarity of ratings by all users (O'Connor & Herlocker, 1999). In this case a cluster center is represented by a m-dimensional vector (being m the number of users) where the *j*-th component is the average rating given by the *j*-th user for items within the clusters. Recommendations for users are computed by finding items that are similar to other items the user has liked.

Various clustering algorithms have been used for user- and item-based clustering, such as K-means (Ungar & Foster, 1998) and divisive hierarchical clustering (Kohrs & Merialdo, 1999). User-based and item-based clustering are typically used as alternative approaches in Web personalization. Nevertheless, they can also be integrated and used in combination, as demonstrated by Mobasher, Dai, Nakagawa, and Luo (2002).

In the context of Web personalization, an important constraint to be considered in the choice of a clustering method is the possibility to derive overlapping clusters. The same user may have different goals and interests at different times. It is inappropriate to capture such overlapping interests of the users in crisp clusters. This makes fuzzy clustering algorithms more suitable for usage mining. In fuzzy clustering, objects which are similar to each other are identified by having high memberships in the same cluster. "Hard" clustering algorithms assign each object to a single cluster that is using the two distinct membership values of 0 and 1. In Web usage profiling, this "all or none" or "black or white" membership restriction is not realistic. Very often there may not be sharp boundaries between clusters and many objects may have characteristics of different classes with varying degrees. Furthermore, a desired clustering technique should be immune to noise, which is inherently present in Web usage data. The browsing behavior of users on the Web is highly uncertain and fuzzy in nature. Each time the user accesses the site, the use may have different browsing goals. The main advantage of fuzzy clustering over hard clustering is that it can capture the inherent vagueness, imprecision, and uncertainty in Web usage data. Fuzzy clustering has been largely used in the context of user profiling for Web personalization (Joshi & Joshi, 2000; Suryavanshi et al., 2005). Castellano, Mesto, Minunno, and Torsello (2007e) prove the applicability of the well-known fuzzy C-means algorithm to extract user profiles. Nasraoui, Krishnapuram, and Joshi (1999) propose a relational fuzzy clustering algorithm named relational fuzzy clustering–maximal density estimator (RFC-MDE). Nasraoui and Frigui (2000) propose a competitive agglomeration relational data (CARD) algorithm to cluster user sessions. A hierarchical fuzzy clustering algorithm has been proposed by Dong and Zhuang (2004) to discover the user access patterns in an effective manner.

Recommendation

Once user preferences are understood by analyzing the derived user profiles, personalized services can be provided to each user, such as sending targeted advertisement to the connected users, adapting the content/structure of the Web site to the user needs, providing a guide to the user navigation, and so forth. Personalization functions can be accomplished in a manual or in an automatic and transparent manner for the user. In the first case, the discovered knowledge has to be expressed in a comprehensible manner for humans, so that knowledge can be analyzed to support human experts in making decisions. To accomplish this task, different approaches have been introduced in order to provide useful information for personalization. An effective method for presenting comprehensive information to humans is the use of visualization tools such as WebViz (Pitkow & Bharat, 1994) that represents navigational patterns as graphs. Reports are also a good method to synthesize and to visualize useful statistical information previously generated. Personalization systems as WUM (Spiliopoulou & Faulstich, 1998) and WebMiner (Cooley, Tan, & Srivastava, 1999) use SQL-like query mechanisms for the extraction of rules from navigation patterns.

Nevertheless, decisions made by the user may create delay and loss of information. A more interesting approach consists of the employment of Web usage mining for personalization. In particular, the knowledge extracted from Web data is automatically exploited to adapt the Web-based system by means of one or more of the personalization functions.

Various approaches can be used for generating a personalized experience for users. These are commonly distinguished in rule-based filtering, content-based filtering, and collaborative or social filtering (Mobasher et al., 2000). In rule-based filtering, static user models are generated through the registration procedure of the users. To generate personalized recommendations, a set of rules is specified, related to the content which is provided to the users with different models. Among the several products which adopt the rule-based filtering approach, Yahoo (Manber, Patel, & Obison, 2000) and Websphere Personalization (IBM) constitute two valid examples. Content-based filtering systems generate recommendations on the basis of the items previously rated by a user. The user profile is obtained by considering the content description of the items and it is exploited to predict a rating for previously unseen items. Examples of systems which adopt this personalization approach are represented by Personal WebWatcher (Mladenic, 1996), NewsWeeder (Lang, 1994), and Letizia (Liebermann & Letizia, 1995). Collaborative filtering systems are based on the assumption that users preferring similar items have the same interests. Personalization is obtained by searching for common features in the preferences of different users which are usually expressed explicitly in the form of item ratings or also in a dynamical manner through the navigational patterns extracted from usage data. Currently, collaborative filtering is the most employed approach of personalization. Amazon.com (Linden et al., 2003) and Recommendation Engine represent two major examples of collaborative filtering systems.

SOFT COMPUTING TECHNIQUES FOR WEB PERSONALIZATION

The term soft computing (SC) indicates a collection of methodologies that work synergistically to find approximate solutions for real-world problems which contain various kinds of inaccuracies and uncertainties. The guiding principle is to devise methods of computation that lead to an acceptable solution at low cost by seeking for an approximate solution to an imprecisely/precisely formulated problem. Computing paradigms underlying SC are:

- Neural computing that supplies the machinery for learning and modeling complex functions;
- Fuzzy logic computing that gives mechanisms for dealing with imprecision and uncertainty underlying real-life problems; and

- Evolutionary computing that provides algorithms for optimization and searching.

Systems based on such paradigms are neural networks (NN), fuzzy systems (FS), and genetic/evolutionary algorithms (GA/EA). Rather than a collection of different paradigms, SC is better regarded as a partnership in which each of the partners provides a methodology for addressing problems in a different manner. From this perspective, the key-points and the shortcomings of SC paradigms appear to be complementary rather than competitive. Therefore, it is a natural practice to build up integrated strategies combining the concepts of different SC paradigms to overcome limitations and exploit advantages of each single paradigm (Hildebrand, 2005; Tsakonas, Dounias, Vlahavas, & Spyropoulos 2002). This relationship enables the creation of hybrid computing schemes which use neural networks, fuzzy systems, and evolutionary algorithms in combination. An inspection of the multitude of hybridization strategies proposed in literature which involve NN, FS, and GA/EA would be somewhat impractical. It is however straightforward to indicate **neuro-fuzzy** (NF) **systems** as the most prominent representatives of hybridizations in terms of the number of practical implementations in several application areas (Lin & Lee, 1996; Nauck, Klawonn, & Kruse, 1997). NF systems use NN to learn and fine tune rules and/or membership functions from input-output data to be used in a FS (Mitra & Pal, 1995). With this approach, the main drawbacks of NN and FS are the black box behavior of NN and the lack of learning mechanism in FS are avoided. NF systems automate the process of transferring expert or domain knowledge into fuzzy rules, hence, they are basically FS with an automatic learning process provided by NN, or NN provided with explicit form of knowledge representation.

In the last few years, the relevance of SC methodologies to Web personalization tasks has drawn the attention of researchers, as indicated in a recent review (Frias-Martinez et al., 2005). Indeed, SC can improve the behavior of Web-based applications, as both imprecision and uncertainty are inherently present in the Web activity. Web data, being unlabeled, imprecise/incomplete, heterogeneous, and dynamic, appear to be good candidates to be mined in the SC framework. Besides, SC seems to be the most appropriate paradigm in Web usage mining where, being human interaction its key component, issues such as approximate queries, deduction, personalization, and learning have to be faced. SC methodologies, being complementary rather than competitive, can be successfully employed in combination to develop intelligent Web personalization systems. In this context, NN with self organization abilities are typically used for pattern discovery and rule generation. FS are used for handling issues related to incomplete/imprecise Web data mining, understandability of patterns, and explicit representation of Web recommendation rules. EA are mainly used for efficient search and retrieval. Finally, various examples of combination between SC techniques can be found in the literature concerning Web personalization, ranging from very simple combination schemas to more complicated ones. An example of simple combination is by Lampinen and Koivisto (2002), where user profiles are derived by a clustering process that combines a fuzzy clustering (the fuzzy C-means clustering) and a neural clustering (using a self-organising map). A Kuo and Chen (2004) discuss a more complex form of hybridization using all the three SC paradigms together, and also design a recommendation system for electronic commerce using fuzzy rules obtained by a combination of fuzzy neural networks and genetic algorithms. Here, fuzzy logic has also been used to provide a soft filtering process based on the degree of concordance between user preferences and the elements being filtered.

NF techniques are especially suited for Web personalization tasks where knowledge interpretability is desired. One of these tasks is the extraction of association rules for recommendation. Gyenesei (2000) explores how fuzzy association rules understandable to humans are learnt from a database containing both quantitative and categorical attributes by using a neuro-fuzzy approach like the one proposed by Nauck (1999). Lee (2001) uses a NF system for recommendation in an e-commerce site. Stathacopoulou, Grigoriadou, and Magoulas (2003) and Magoulas, Papanikolau, and Grigoriadou (2001) use a NF system to implement a classification/recommendation system with the purpose of adapting the contents of a Web course according to the model of the student. Recently Castellano, Fanelli, and Torsello (2007d) have proposed a Web personalization approach that uses fuzzy clustering to derive user profiles and a neural-fuzzy system to learn fuzzy rules for dynamic link recommendation. The next section is devoted to outlining the main features of our approach, in order to give an example of how different SC techniques can be used synergistically to perform Web personalization.

A NEURO-FUZZY WEB PERSONALIZATION SYSTEM

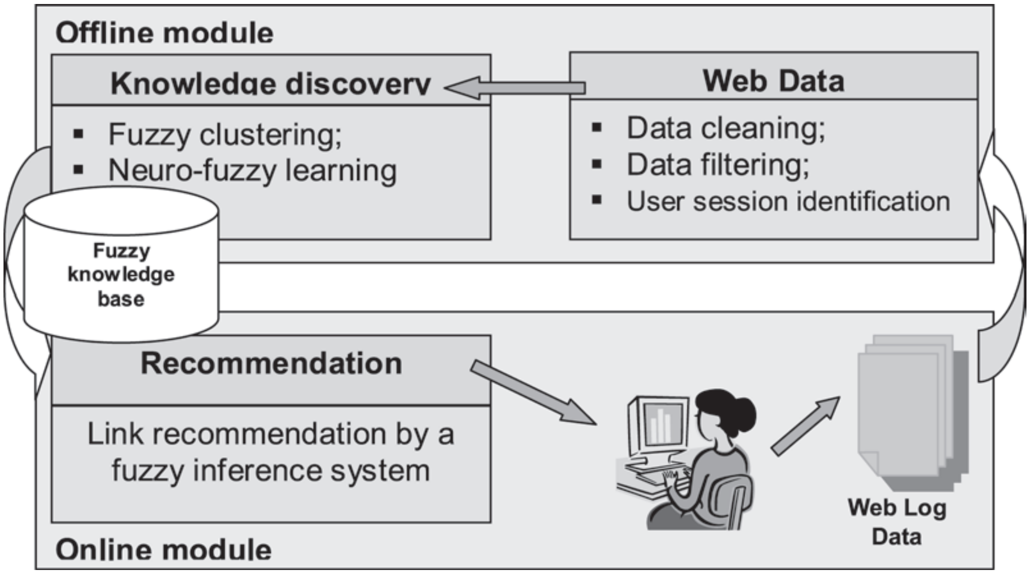
In this section, we describe a WUM personalization system for dynamic link suggestion based on a neuro-fuzzy approach. A fuzzy clustering algorithm is applied to determine user profiles by grouping preprocessed Web usage data into session categories. Then, a hybrid approach based on the combination of the fuzzy reasoning with a neural network is employed in order to derive fuzzy rules useful to provide dynamical predictions about Web pages to be suggested to the active user, according to user profiles previously identified.

According to the general scheme of a WUM personalization process described in section 3, three different phases can be distinguished in our approach:

- **Preprocessing** of Web log files in order to extract useful data about URLs visited during user sessions.
- **Knowledge discovery** in order to derive user profiles and to discover associations between user profiles and URLs to be recommended.
- **Recommendation** in order to exploit the knowledge extracted through the previous phases to dynamically recommend interesting URLs to the active user.

As illustrated in Figure 1, two major modules can be distinguished in the system: an off-line module that performs log data preprocessing and knowledge discovery, and an online module that recommends interesting Web pages to the current user on the basis of the discovered knowledge. In particular, during the preprocessing task, user sessions are extracted from the log files which are stored by the Web server. Each user session is represented by one record which registers the accesses exhibited by the user in that session. Next, a fuzzy clustering algorithm is executed on these records to group similar sessions into session categories representing user profiles. Finally, starting from the extracted user profiles and the available data about user sessions, a knowledge base expressed in the form of fuzzy rules is extracted via a neuro-fuzzy learning strategy. Such a knowledge base is exploited during the recommendation phase (performed by the online module) to dynamically suggest links to Web pages judged interesting for the current user. Specifically, when a user requests a new page, the online module matches the user's current partial session with the session categories identified by the off-line module and derives the degrees of relevance for URLs by means of a fuzzy inference process. In the following, we describe in more detail all the tasks involved in the Web personalization process.

Figure 1. The scheme of the proposed Web personalization system



Preprocessing

The aim of the preprocessing step is to identify user sessions starting from the information contained in a Web log file. Preprocessing of access log files is performed by means of log data preprocessor (LODAP) (Castellano, Fanelli, & Torsello, 2007a), a software tool that analyzes usage data stored in log files to produce statistics about the browsing behavior of the users visiting a Web site and to create user sessions by identifying the sequence of pages accessed by each visitor. LODAP preprocesses log data into three steps: data cleaning, data structuration, and data filtering. During data cleaning, Web log data are cleaned from the useless information in order to retain only records corresponding to the explicit requests of the users (i.e. requests with an access method different from "GET," failed and corrupt requests, requests for multimedia objects, and visits made by Web robots are removed). Next, significant log entries are structured into user sessions. In LODAP, a user session is defined as the finite set of URLs accessed by a user within a predefined time period (in our work, 25 minutes). Since the information about the user login is not available, user sessions are identified by grouping the requests originating from the same IP address during the established time period. The set of all users (IP) is defined by $U = \left\{u_1, u_2, \ldots, u_{n_U}\right\}$ and a user session is defined as the set of accesses originating from the same user (IP) within a predefined time period. Formally, a user session is represented as a triple $\mathbf{s}_i = \left\langle u_i, t_i, \mathbf{p}_i \right\rangle$ where $u_i \in U$ represents the user identifier, t_i is the total time access of the *i*-th session, and $\mathbf{p}_i$ is the set of all pages requested during the *i*-th session. More in detail,

$$\mathbf{p}_i = \left\langle \left(p_{i1}, t_{i1}, N_{i1}\right), \left(p_{i2}, t_{i2}, N_{i2}\right), \ldots, \left(p_{in_i}, t_{in_i}, N_{in_i}\right) \right\rangle$$

where p_{ik} is the *k*-th URL visited during the i-th session, *t*ik is the total access time to page p_{ik}, and N_{ik} represents the number of accesses to page *p*ik during the *i*-th session. Summarizing, after data structuration, a collection $S = \left\langle \mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{n_s} \right\rangle$ of n_s sessions is identified from the log data. Finally, LODAP applies a data filtering process in order to remove requests for very low support URLs, that is, requests to pages which do not appear in a sufficient number of sessions, and requests for very high support URLs, that is, requests to pages which appear in nearly all sessions. Also, all sessions that include a very low number of visited URLs are removed. Hence, after data filtering, only m page requests (with $m \leq n_P$) and only n sessions (with $n \leq n_S$) are retained.

Once user sessions have been identified by LODAP, we create a visitor behavior model by defining a measure expressing the interest degree of the users for each visited page during a session. In our approach, we measure the interest degree for a page as the average access time on that page. Precisely, the interest degree for the *j*-th page in the *i*-th user session is defined as $ID_{ij} = t_{ij} / N_{ij}$ where t_{ij} is the overall time spent by the user on the *j*-th page and N_{ij} is the number of accesses to that page during the *i*-th session. Hence, we model the visitor behavior of each user through a pattern of interest degrees for all pages visited by that user. Since the number of pages visited by different users may vary, visitor behavior patterns may have different dimensions. To obtain a homogeneous behavior model for all users, we translate behavior patterns into vectors having the same dimension equal to the number *m* of pages retained by LODAP after page filtering. In particular, the behavior of the *i*-th user $i = 1, \ldots, n$) is modeled as a vector $\mathbf{b}_i = \left(b_{i1}, b_{i2}, \ldots, b_{im}\right)$ where

$$b_{ij} = \begin{cases} ID_{ij} & \text{if page } p_j \text{ is visited during session } \mathbf{s}_i \\ 0 & \text{otherwise} \end{cases}$$

Summarizing, we model the visitor behaviors by a $n \times m$ matrix $\mathbf{B} = \left[b_{ij}\right]$ where each entry represents the interest degree of the *i*-th user for the *j*-th page. Based on this matrix, visitors with similar preferences can be successively clustered together to create user profiles, as described in the following subsection.

Knowledge Discovery

In our approach, the knowledge discovery phase involves the creation of user profiles and the derivation of recommendation rules. This is performed by rule extraction for Web recommendation (REXWERE) (Castellano, Fanelli, & Torsello, 2007b), a software tool designed to extract knowledge from user sessions identified by LODAP. REXWERE employs a hybrid approach based on the combination of fuzzy reasoning and neural learning to extract knowledge in two successive phases: user profiling and fuzzy rule extraction. In user profiling, similar user sessions are grouped into clusters (user profiles) by means of a fuzzy clustering algorithm. Then, a neuro-fuzzy approach is applied to learn fuzzy rules which capture the association between user profiles and Web pages to be recommended. These recommendation rules are intended to be exploited by the online component of a WR system that dynamically suggests links to interesting pages for a visitor of a Web site, according to the profiles the user belongs to. A key feature of REXWERE is the wizard-based interface that guides the execution of the different steps involved in the extraction of knowledge

Figure 2. The start-up panel of REXWERE

for recommendation. Figure 2 shows the start-up panel of REXWERE.

Starting from the behavior data derived from user sessions, REXWERE extracts recommendation rules in two main phases:

1. User profiling, that is, the extraction of user profiles through clustering of behavior data.
2. Fuzzy rule extraction, that is, the derivation of a set of rules that capture the association between the extracted user profiles and Web pages to be recommended. This task is carried out through three modules:
    - The *dataset creation* module which creates the training set and the test set needed for the learning of fuzzy rules;
    - The *rule extraction* module that derives an initial fuzzy rule base by means of an unsupervised learning; and
    - The *rule refinement* module that improves the accuracy of the fuzzy rule base by means of a supervised learning.

As result, REXWERE provides in output a set of fuzzy recommendation rules to be used as knowledge base in an online activity of dynamic link suggestion.

Discovery of User Profiles

The first task of REXWERE is the extraction of user profiles that categorize user sessions on the basis of similar navigational behaviors. This is accomplished by means of the *profile extraction* module that is based on a clustering approach. Clustering algorithms are widely used in the context of user profiling since they have the capacity to examine large quantity of data in a fairly reasonable amount of time. In particular, fuzzy clustering techniques seem to be particularly suited in this context because they can partition data into overlapping clusters (user profiles). Due to this peculiar characteristic, a user may belong to more than one profile with a certain membership degree. Two fuzzy clustering algorithms are implemented in REXWERE to extract user profiles:

- The well-known fuzzy C-means (FCM) algorithm (Castellano et al., 2007d), that belongs to the category of clustering algorithms working on object data expressed in the form of feature vectors.
- The CARD+ algorithm (Castellano, Fanelli, & Torsello, 2007c), a modified version of the competitive agglomeration relational data algorithm (Nasraoui & Frigui, 2000), which works on relational data representing the pairwise similarities (dissimilarities) between objects to be clustered.

These two algorithms differ in some features. While the FCM directly works on the behavior matrix **B** containing the interest degrees of each user for each page, CARD+ works on a relation matrix containing the dissimilarity values between all pairs of behavior vectors (rows of matrix **B**). Moreover, one key feature of CARD+ is the ability to automatically determine the final number of clusters starting from an initial random number. On the contrary, the FCM requires the number of clusters to be fixed in advance. In this case, the proper number of profiles is established by calculating the Xie-Beni index (Halkidi, Batistakis, & Vazirgiannis, 2002) for different partitions corresponding to different number of clusters; the partition with the smallest value of the Xie-Beni index corresponds to the optimal number of clusters for the available input data.

Both the FCM and the CARD+ provide the following results:

- *C* cluster centers (user profiles) represented as vectors $\mathbf{v}_c = \left(v_{c1}, v_{c2}, \ldots, v_{cm}\right)$ with

$c = 1,\dots,C$

- A fuzzy partition matrix $\mathbf{U} = \left[u_{ic}\right]_{i=1,\dots,n}^{c=1,\dots,C}$ where each component u_{ic} represents the membership degree of the *i*-th user to the *c*-th profile.

These results are used in the successive knowledge discovery task performed by REXWERE.

Discovery of Recommendation Rules

Once profiles have been extracted, REXWERE enters in the second knowledge extraction phase, that is, the extraction of fuzzy rules for recommendation. Such rules represent the knowledge base to be used in the ultimate online process of link recommendation. Each recommendation rule expresses a fuzzy relation between a behavior vector $\mathbf{b} = \left(b_1, b_2, \dots, b_m\right)$ and relevance of URLs in the following form:

IF (*b*1**is***A*1k) AND … AND (*b*m is *A*mk) THEN (relevance of URL_1 is *y*1k) AND … AND (relevance of URL_m is *y*1k)

for $k = 1,..,K$ where *K* is the number of rules, *A*jk (*j*=1,…, *m*) are fuzzy sets with Gaussian membership functions defined over the input variables *b*j, and y_{jk} are fuzzy singletons expressing the relevance degree of the *j*th URL.

The main advantage of using a fuzzy knowledge base for recommendation is the readability of the extracted knowledge. Actually, fuzzy rules can be easily understood by human users since they can be expressed in a linguistic fashion by labelling fuzzy sets with linguistic terms such as LOW, MEDIUM, and HIGH. Hence, a fuzzy rule for recommendation can assume the following linguistic form:

IF (the degree of interest for URL_1 is LOW) AND … AND (the degree of interest for URL_m is HIGH) THEN (recommend URL_1 with relevance 0.3) AND … AND (recommend URL_m with relevance 0.8)

Such fuzzy rules are derived through a hybrid strategy based on the combination of fuzzy reasoning with a specific neural network that encodes in its structure the discovered knowledge in form of fuzzy rules. The network is trained on a set of input-output samples describing the association between user sessions and preferred URLs. Precisely, the training set is a collection of *n* input-output vectors: $\mathbf{T} = \left\langle\left(\mathbf{b}_i, \mathbf{r}_i\right)\right\rangle_{i=1,\dots,n}$ where the input vector **b**i represents the behavior vector of the *i*-th user and the desired output vector $\mathbf{r}_i$ expresses the relevance degrees associated to the *m* URLs for the *i*-th visitor. To compute such relevance degrees, we exploit information embedded in the profiles extracted through fuzzy clustering. Precisely, for each behavior vector **b**i we consider its membership values $\left\{u_{ic}\right\}_{c=1,\dots,C}$ in the fuzzy partition matrix **U**. Then, we identify the two top matching profiles $c_1, c_2 \in \left\{1,..,C\right\}$ as those with the highest membership values. The relevance degrees in the output vector $\mathbf{r}_i = \left(r_i^1, r_i^2, \dots, r_i^m\right)$ are hence calculated as follows: $r_i^j = u_{ic}v_{ic_1} + u_{ic}v_{ic_2}$ for $j = 1,\dots,m$ and $i = 1,\dots,n$. Once the training set has been constructed, the neural network can enter the learning phase to extract the knowledge embedded into training set and represent it as a collection of fuzzy rules. The learning is articulated in two steps. The first step is based on an unsupervised learning, based on a rival penalized mechanism, which provides a clustering of the behavior vectors and the definition of an initial fuzzy rule base. In this step, the structure and the parameters of fuzzy rules are identified. Successively, the obtained knowledge base is refined by a supervised learning process. Here, fuzzy rule parameters are tuned via supervised learning to improve the accuracy of the derived knowledge. Major details on the algorithms underlying the learning strategy can be retrieved in the work of Castellano, Castiello, Fanelli, and Mencar (2005).

Recommendation

The ultimate task of our Web personalization approach is the online recommendation of links to Web pages judged interesting for the current user of the Web site. Specifically, when a new user accesses the Web site, an online module matches the user's current partial session against the fuzzy rules currently available in the knowledge base and derives a vector of relevance degrees by means of a fuzzy inference process.

Formally, when a new user has access to the Web site, an active user's current session is created in the form of a vector $\mathbf{b}^0$. Each time the user requests a new page, the vector is updated. To maintain the active session, a sliding window is used to capture the most recent user's behavior. Thus, the partial active session of the current user is represented as a vector $\mathbf{b}^0 = \left(b_1^0, \ldots, b_m^0\right)$ where some values are equal to zero, corresponding to unexplored pages.

Based on the set of *K* rules generated through the neural learning described above, the recommendation module provides URL relevance degrees by means of the following fuzzy reasoning procedure:

(1) Calculate the matching degree of current behavior vector $\mathbf{b}^0$ to the *k*-th rule, for $k = 1,..,K$ by means of product operator:

$$\mu_k\left(\mathbf{b}^0\right) = \prod_{j=1}^{n} \mu_{jk}\left(b_j^0\right)$$

(2) Calculate the relevance degree r_j^0 for the *j*-th URL as:

$$r_j^0 = \frac{\sum_{k=1}^{K} r_{jk}\mu_k\left(\mathbf{b}^0\right)}{\sum_{k=1}^{K} \mu_k\left(\mathbf{b}^0\right)}, j = 1 \ldots m$$

This inference process provides the relevance degree for all the considered *m* pages, independently on the actual navigation of the current user. In order to perform dynamic link suggestion, the recommendation module first identifies URLs that have been not visited by the current user, that is, all pages such that $b_j^0 = 0$. Then, among unexplored pages, only those having a relevance degree r_j^0 greater than a properly defined threshold α are recommended to the user. In practice, a list of links is dynamically included in the page currently visited by the user.

A Case Study

The proposed Web personalization approach was applied on a Web site targeted to young users (average age 12 years old), that is, the Italian Web site of the Japanese movie Dragon Ball (www.dragonballgt.it). This site was chosen because of its high daily number of accesses (thousands of visits each day).

The LODAP system was used to identify user sessions from the log data collected during a period of 24 hours. After data cleaning, the number of requests was reduced from 43,250 to 37,740 that were structured into 14,788 sessions. The total number of distinct URLs accessed in these sessions was 2,268. Support-based data filtering was used to eliminate requests for URLs having a number of accesses less than 10% of the maximum number of accesses, leading to only 76 distinct URLs and 8,040 sessions. Also, URLs appearing in more than 80% of sessions (including the site entry page) were filtered out, leaving 70 final URLs and 6,600 sessions. In a further filtering step, LODAP eliminated short sessions, leaving only sessions with at least three distinct requests. We obtained a final number of 2,422 sessions. The 70 pages in the Web site were labeled with a number (see Table 1) to facilitate the analysis of results. Once user sessions were identified and visitor behavior models were derived by calculating the interest degrees of each user for each page, leading to a 2422x70 behavior matrix.

Table 1. Description of the pages in the Web site

Pages	Content
1	Home page
2	Comments by users
3,…,12	Pictures related to the movie
13,…,18	Pictures of characters
19, 26, 27	Matches
20, 21, 36, 47, 48	Services (registration, login,…)
22, 23, 25, 28, ..., 31, 50, 51	General information about the movie
32, ..., 35, 55	Entertainment (games, videos,...)
37, ..., 46, 49, 52, ..., 54, 56	Description of characters
57, ..., 70	Galleries

Next, the two fuzzy clustering algorithms implemented in REXWERE were applied to the behavior matrix in order to obtain clusters of users with similar navigational behavior. Several runs of FCM were carried out with different number of clusters (C=30, 20, 15, 10). For each trial, we analyzed the obtained cluster center vectors and we observed that many of them were identical. Hence, an actual number of three clusters were found in each run. Also, a single run of the CARD+ was carried out by setting a maximum number of clusters equal to C=15. As a result, this clustering algorithm provided three clusters, confirming the results obtained by the FCM algorithm. This demonstrated that three clusters were enough to model the behavior of all the considered users. Table 2 summarizes the three clusters obtained by CARD+ that are very similar to those obtained after different trials of FCM. For each cluster, the cardinality and the first eight (most interesting) pages are displayed. It can be noted that some pages (e.g., Pages 12, 22, and 28) appear in more than one cluster, thus showing the importance of producing overlapping clusters. In particular, Page 28 (i.e., the page that lists the episodes of the movie) appears in all the three clusters with the highest degree of interest.

An interpretation of the three clusters revealed the following profiles:

- Profile 1. Visitors in this profile are mainly interested in pictures and descriptions of characters.
- Profile 2. These visitors prefer pages that link to entertainment objects (games and video)
- Profile 3. These visitors are mostly interested in matches among characters.

A qualitative analysis of these profiles made by designer of the considered Web site confirmed that they correspond to real user categories reflecting the interests of the typical site users.

The next step was the creation of recommendation rules starting from the extracted user profiles. A neural network with 70 inputs (corresponding to the components of the behavior vector) and 70 outputs (corresponding to the relevance values of the Web pages) was considered. The network was trained on a training set of 1,400 input-output samples derived from the available 2,000 behavior patterns and from the three user profiles, as described in Section 5.2.2. The remaining 600 samples were used for testing. The training of the network was stopped when the error on the training set dropped below 0.01, corresponding to a testing error of 0.03.

The derived fuzzy rule base was integrated into the online recommendation module to infer the relevance degree of each URL for the active user. These relevance degrees were ultimately used to suggest a list of links to unexplored pages retained interesting to the current user. To perform link recommendation, the navigational behavior of the active user was observed during a temporal window of 3 minutes in order to derive the behavior pattern corresponding to the user's partial

Table 2. Clusters of visitor behaviour

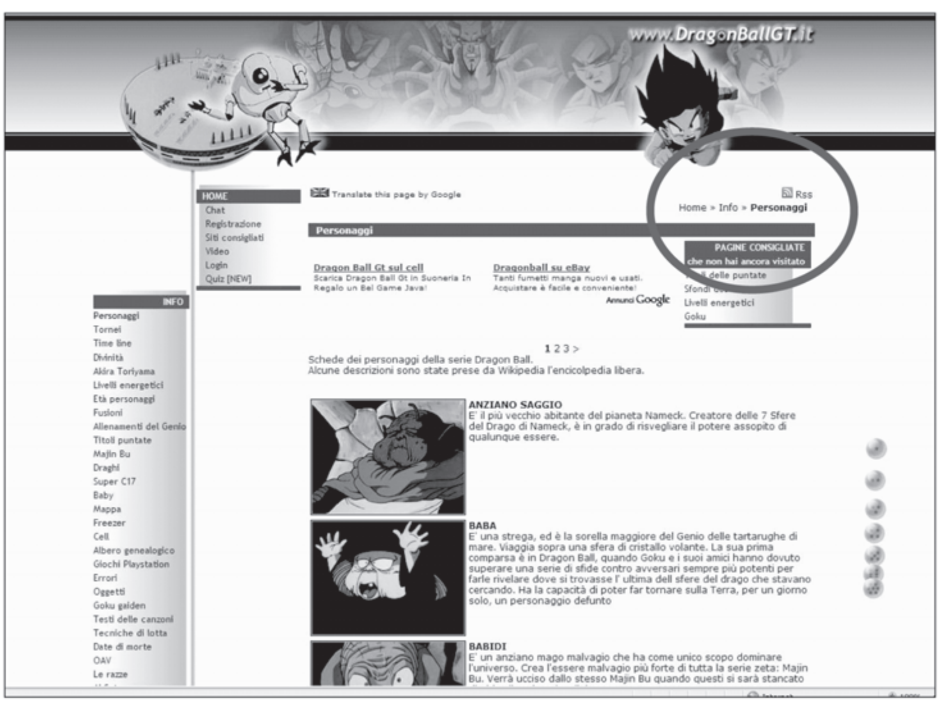
Cluster	Cardinality	Visited pages	Interest degree
1	906	(28, 12, 15, 43, 17, 22, 13, 50)	(11.1, 7.3, 6.9, 6.6, 6.59, 5.14, 4.5, 4.4)
2	599	(28, 26, 22, 55, 12, 33, 49, 32)	(80.8, 43.3, 30.1, 25.9, 24.5, 19.2, 18.1, 14.3)
3	917	(28, 12, 26, 13, 22, 9, 19, 16)	(5.3, 4.0, 3.4, 3.1, 2.6, 2.60, 2.3, 2.2)

visit. Such behavior pattern was used as input to the fuzzy rule inference process that computes the relevance degrees for all the considered 10 pages. Then, among the unexplored pages, only those having a relevance degree greater than $\alpha = 0.7$ were included in the list of links to be suggested. As an example, Figure 3 shows a page of the considered Web site after the online recommendation module has dynamically included the list of suggested links.

CONCLUSION AND FUTURE TRENDS

The rapid development of the World Wide Web as a medium for information dissemination has generated a growing interest in the domain of Web personalization that may offer a variety of functionalities in several context, such as customization, task performance support, personalized guidance, and so forth. Specifically, in personal-

Figure 3. A personalized page of the Web site. The recommended links are displayed in the up-right corner (inside the red circle)© 2008 Fabrizio Mesto. Used with permission.



ized guidance, the knowledge acquired from the analysis of users' navigational behavior (usage data) can be conveniently exploited in order to customize the Web information space to the necessities of users. As a consequence, there is growing interest in tools for automatic identification of user profiles by modeling the preferences of different user categories. Once user preferences are understood by analyzing the discovered user profiles, personalized services can be provided to each user.

In the Web personalization context, soft computing techniques emerge as valid tools to handle the ambiguity and uncertainty inherent in Web usage data. A brief survey of recent approaches to Web personalization that employ SC techniques has been presented. The survey emphasizes how most of Web personalization applications developed so far are based on combinations of SC techniques. As an example, a Web personalization system for dynamic link recommendation joining techniques from the neural and the fuzzy paradigms has been described. This neuro-fuzzy personalization system extracts knowledge from Web usage data in a twofold form: a set of fuzzy user profiles that capture preferences of similar users and a collection of fuzzy rules that describe associations between user profiles and links to be recommended.

In addition to those discussed in this chapter, there are some other aspects of Web personalization where SC is likely to play a key role. For example, user profiles generated by Web mining techniques are typically represented in a simplistic manner, by means a vector of ratings. More expressive models based on fuzzy logic could be explored in order to represent the vague and heterogeneous information characterizing user preferences. Another important facet is the ability to identify the continuous changes in interests of users and dynamically adapt user profiles according to these changes. Neural network learning algorithms based on online schemas could be investigated to cope with this issue. On the whole, hybrid approaches that synergistically combine SC methods show great potential for Web personalization, opening new research directions within the area of Web intelligence.

REFERENCES

Abraham, A. (2003). Business intelligence from Web usage mining. *Journal of Information & Knowledge Management*, *2*(4), 375–390. doi:10.1142/S0219649203000565

Adomavicius, G., & Thuzilin, A. (2005). Towards the next generation of recommender systems: A survey of the state of the art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 734–749. doi:10.1109/TKDE.2005.99

Banerjee, A., & Ghosh, J. (2001). Clickstream clustering using weighted longest common subsequences. In *Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining*.

Baraglia, R., & Silvestri, F. (2004). *An online recommender system for large Web sites.* Paper presented at the IEEE/WIC/ACM International Conference on Web Intelligence, Beijing, China.

Callan, J., Smeaton, A., Beaulieu, M., Borlund, P., Brusilovsky, P., Chalmers, M., et al. (2001). Personalization and recommender systems in digital libraries. In *Proceedings of the 2nd DELOS Workshop on Personalization and Recommender Systems in Digital Libraries*.

Castellano, G., Castiello, C., Fanelli, A. M., & Mencar, C. (2005). Knowledge discovering by a neuro-fuzzy modelling framework. *Fuzzy Sets and Systems*, *149*, 187–207. doi:10.1016/j.fss.2004.07.015

Castellano, G., Fanelli, A. M., & Torsello, A. M. (2007c, July 15-22). Web user profiling using relational fuzzy clustering. In *Proceedings of 12th International Conference on Fuzzy Theory and Technology (FTT 2007),* Salt Lake City.

Castellano, G., Fanelli, A. M., & Torsello, M. A. (2007a, February 16-19). LODAP: A log data preprocessor for mining Web browsing patterns. In *Proceedings of the 6th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Base (AIKED 2007),* Corfu, Greece (pp. 12-17).

Castellano, G., Fanelli, A. M., & Torsello, M. A. (2007b). REXWERE: A tool for fuzzy rule extraction in Web recommendation. In *Proceedings of NAFIPS 2007*, San Diego.

Castellano, G., Fanelli, A. M., & Torsello, M. A. (2007d) A neuro-fuzzy collaborative filtering approach for Web recommendation. *International Journal of Computational Science, 1*(1), 27-39. ISSN: 1992-6669 (Print)/ISSN: 1992-6677.

Castellano, G., Mesto, F., Minunno, M., & Torsello, M. A. (2007e, July 7-10). Web user profiling using fuzzy clustering. In *Proceedings of 7th International Workshop on Fuzzy Logic and Applications (WILF 2007)*, Portofino Vetta (Genova), Italy.

Cho, Y. H., & Kim, J. K. (2004). Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce. *Expert Systems with Applications, 26*, 233–246. doi:10.1016/S0957-4174(03)00138-6

Cooley, R. (2000). *Web usage mining: Discovery and application of interesting patterns from Web data.* Unpublished doctoral thesis, University of Minnesota.

Cooley, R., Tan, P. N., & Srivastava, J. (1999). *Discovering of interesting usage patterns from Web data* (Tech. Rep. No. 99-022). University of Minnesota.

Dong, Y., & Zhuang, Y. (2004, June 15-19). Fuzzy hierarchical clustering algorithm facing large database. In *Proceedings of the 5th World Congress on Intelligent Control and Automation*, Hangzhou, P.R. China.

Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for Web personalization. *ACM TOIT*, *3*(1), 2–27.

Enembreck, F., Barthès, J.-P., & Ávila, B. C. (2004). *Personalizing information retrieval with multi-agent systems* (LNCS 3191/2004, pp. 77-91). Springer, Berlin-Heidelberg.

Etzioni, O. (1996). The world wide Web: Quagmire or gold mine. *Communications of the ACM*, *39*(11), 65–68. doi:10.1145/240455.240473

Facca, F. M., & Lanzi, P. L. (2005). Mining interesting knowledge from Weblogs: A survey. *Data & Knowledge Engineering*, *53*, 225–241. doi:10.1016/j.datak.2004.08.001

Frias-Martinez, E., Magoulas, G., Chen, S., & Macredie, R. (2005). Modeling human behavior in user-adaptive systems: recent advances using soft computing techniques. *Expert Systems with Applications*, *29*(2), 320–329. doi:10.1016/j.eswa.2005.04.005

Gyenesei, A. (2000). *A fuzzy approach for mining quantitative association rule* (TUCS Tech. Rep. 336). Univ. Turku, Dept. Comput. Sci., Lemminkisenkatu 14, Finland.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). *Cluster validity methods: Part II*. in SIGMOD Record.

Han, J., & Kamber, M. (2001). *Data mining concepts and techniques.* Morgan Kaufmann.

Heer, J., & Chi, E. H. (2002). Mining the structure of user activity using cluster stability. In *Proceedings of the Workshop on Web Analytics/Second SIAM Conference on Data Mining.* ACM Press.

Hildebrand, L. (2005). Hybrid computational intelligence systems for real world applications. *Series studies in fuzziness and soft computing* (Vol. 179, pp. 165-195). Springer Berlin/Heidelberg.

Huang, J. Z., Ng, M., Ching, W.-K., Ng, J., & Cheung, D. (2001, August 26). A cube model and cluster analysis for Web access sessions. In R. Kohavi, B. Masand, M. Spiliopoulou, & J. Srivastava (Eds.), *WEBKDD 2001—Mining Web Log Data Across All Customers Touch Points, Third International Workshop*, San Francisco (LNCS 2356, pp. 48-67). Springer.

Huang, X., Cercone, N., & An, A. (2002a). Comparison of interestingness functions for learning Web usage patterns. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management* (pp. 617-620). ACM Press.

Jespersen, S. E., Thorhauge, J., & Pedersen, T. B. (2002). A hybrid approach to Web usage mining. In *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery* (pp. 73-82). Springer-Verlag.

Joshi, A., & Joshi, K. (2000). *On mining Web access logs*. Paper presented at the ACM SIGMOID Workshop on Research Issues in Data Mining and Knowledge Discovery (Vol. 63, No. 69).

Joshi, K. P., Joshi, A., & Yesha, Y. (2003). On using a warehouse to analyze Web logs. *Distributed and Parallel Databases*, *13*(2), 161–180. doi:10.1023/A:1021515408295

Kim, D.-W., & Lee, K. H. (2001). A new fuzzy information retrieval system based on user preference model. *FUZZY-IEEE*, 127-130.

Kohrs, A., & Merialdo, B. (1999). Clustering for collaborative filtering applications. *Computational intelligence for modelling, control and automation*. IOS Press.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining*, *2*(1), 1-15. ACM.

Kraft, D. H., Chen, J., Martin-Bautista, M. J., & Vila, M. A. (2002). Textual information retrieval with user profiles using fuzzy clustering and inferencing. In P. S. Szczepaniak, J. Segovia, J. Kacprzyk, & L. A. Zadeh (Eds.), *Intelligent exploration of the Web*. Heidelberg, Germany: Physica-Verlag.

Kuo, R. J., & Chen, J. A. (2004). A decision support system for order selection in electronic commerce based on fuzzy neural network supported by real-coded genetic algorithm. *Expert Systems with Applications*, *26*, 141–154. doi:10.1016/S0957-4174(03)00115-5

Lampinen, T., & Koivisto, H. (2002). Profiling network applications with fuzzy c-means clustering and self-organising map. In *Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery: Computational Intelligence for the E-Age* (pp. 300-304).

Lang, K. (1995). Newsweeder: Learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning*.

Langley, P. (1999). User modeling in adaptive interfaces. In *Proceedings of the Seventh International Conference on User Modeling*, Canada (pp. 357-370).

Lee, R. S. T. (2001). iJADE IWShopper: A new age of intelligent Web shopping system based on fuzzy-neuro agent technology. In N. Zhong, Y. Yao, J. Liu, & S. Ohsuga *Web intelligence: Research and development* (LNAI 198, pp. 403-412).

Lieberman, H., & Letizia (1995). An agent that assists Web browsing. In *Proceedings of the 14th International Joint Conference in Artificial Intelligence* (pp. 924-929).

Lin, C. T., & Lee, C. S. (1996). *Neural fuzzy systems: A neuro-fuzzy synergism to intelligent systems*. Englewood Cliffs, NJ: Prentice-Hall.

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76–80. doi:10.1109/MIC.2003.1167344

Magoulas, G. D., Papanikolau, K. A., & Grigoriadou, M. (2001). Neurofuzzy synergism for planning the content in a Web-based course. *Informatica*, *25*(1), 39–48.

Manber, U., Patel, A., & Obison, J. (2000). Experience with personalization on Yahoo. *Communications of the ACM*, *43*(8), 35–39. doi:10.1145/345124.345136

Menasalvas, E., Millan, S., Pena, J., Hadjimichael, M., & Marban, O. (2002). Subsessions: A granular approach to click path analysis. In *Proceedings of the FUZZ-IEEE Fuzzy Sets and Systems Conference, at the World Congress on Computational Intelligence*, Honolulu (pp. 12-17).

Mitra, S., & Pal, S. K. (1995). Fuzzy multi-layer perceptron, inferencing and rule generation. *IEEE Transactions on Neural Networks*, *6*, 51–63. doi:10.1109/72.363450

Mladenic, D. (1996). *Personal Web watcher: Implementation and design* (Tech. Rep. IJS-DP-7472). Department of Intelligent Systems, J. Stefan Institute, Slovenia.

Mobasher, B. (2005). Web usage mining and personalization. In M. P. Singh (Ed.), *Practical handbook of internet computing*. CRC Press.

Mobasher, B. (2006). Web usage mining. In B. Liu (Ed.), *Web data mining: Exploring hyperlinks, contents and usage data*. Berlin-Heidelberg: Springer.

Mobasher, B. (2007a). Data mining for personalization. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive Web: Methods and strategies of Web personalization* (LNCS 4321, pp. 90-135). Berlin-Heidelberg: Springer.

Mobasher, B. (2007b). Recommender systems. *Kunstliche Intelligenz, Special Issue on Web Mining*, (3), 41-43. Bremen, Germany: BottcherIT Verlag.

Mobasher, B., Cooley, R., & Srivastava, J. (1999). *Automatic personalization based on Web usage mining* (Tech Rep. 99010). DePaul University, Department of Computer Science.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on Web usage mining. *Communications of the ACM*, *43*(8), 142–151. doi:10.1145/345124.345169

Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2001). *Effective personalization based on association rule discovery from Web usage data.* Paper presented at the ACM Workshop on Web Information and Data Management, Atlanta.

Mobasher, B., Dai, H., Nakagawa, M., & Luo, T. (2002). Discovery and evaluation of aggregate usage profiles for Web personalization. *Data Mining and Knowledge Discovery*, *6*, 61–82. doi:10.1023/A:1013232803866

Mobasher, B., Nasraoui, O., Liu, B., & Masand, B. (2006). *Advances in Web mining and Web usage analysis* (2004 WebKDD Volume) (LNAI 3932). Berlin-Heidelberg: Springer.

Mortazavi-Asl, B. (2001). *Discovering and mining user Web-page traversal patterns*. Unpublished master's thesis, Simon Fraser University.

Mulvenna, M., Anand, S., & Buchner, A. (2000). Personalization on the net using Web mining. *CACM*, *43*(8), 123–125.

Nanopoulos, A., Katsaros, D., & Manolopoulos, Y. (2001, August 26). Exploiting Web log mining for Web cache enhancement. In R. Kohavi, B. Masand, M. Spiliopoulou, & J. Srivastava (Eds.), *WEBKDD 2001—Mining Web Log Data Across All Customers Touch Points, Third International Workshop*, San Francisco (LNCS 2356, pp. 68-87). Springer.

Nasraoui, O. (2005). World Wide Web personalization. In J. Wang (Ed.), *Encyclopedia of data mining and data warehousing*. Hershey, PA: IGI Global, Inc.

Nasraoui, O., & Frigui, H. (2000). Extracting Web user profiles using relational competitive fuzzy clustering. *International Journal of Artificial Intelligence Tools*, *9*(4), 509–526. doi:10.1142/S021821300000032X

Nasraoui, O., Krishnapuram, R., & Joshi, A. (1999). Relational clustering based on a new robust estimator with applications to Web mining. In *Proceedings of the International Conf. North American Fuzzy Info. Proc. Society (NAFIPS 99)*, New York (pp. 705-709).

Nauck, D. (1999). Using symbolic data in neuro-fuzy classification. In *Proceedings of NAFIPS'99*, New York (pp. 536-540).

Nauck, D., Klawonn, F., & Kruse, R. (1997). *Foundations of neuro–fuzzy systems*. Chichester, UK: Wiley.

O'Connor, M., & Herlocker, J. (1999). Clustering items for collaborative filtering. In *Proceedings of ACM SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation*.

Pal, S., Talwar, V., & Mitra, P. (2002). Web mining in soft computing framework: Relevance, state of the art and future directions. *IEEE Transactions on Neural Networks*, *13*(5), 1163–1177. doi:10.1109/TNN.2002.1031947

Pei, J., Han, J., Mortazavi-asl, B., & Zhu, H. (2000). Mining access patterns efficiently from Web logs. Paper presented at the *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 396-407).

Perkowitz, M., & Etzioni, O. (1997). *Adaptive sites: Automatically learning from user access patterns* (Tech. Rep. UW-CSE-97-03-01). University of Washington.

Pierrakos, D., Paliouras, G., Papatheodorou, C., & Spyropoulos, C. D. (2003). Web usage mining as a tool for personalization: A survey. *User Modeling and User-Adapted Interaction*, *13*(4), 311–372. doi:10.1023/A:1026238916441

Pitkow, J., & Bharat, K. (1994). WEBVIZ: A tool for World-Wide Web access log visualization. In *Proceedings of the 1st International World-Wide Web Conference*. Geneva, Switzerland (pp. 271-277).

Pohl, W. (1996). Learning about the user modeling and machine learning. In V. Moustakis & J. Herrmann (Eds.), *International Conference on Machine Learning Workshop Machine Learning Meets Human-Computer Interaction* (pp. 29-40).

Sankar, K. P., Varun, T., & Pabitra, M. (2002). Web mining in soft computing framework: Relevance, state of the art and future directions. *IEEE Transactions on Neural Networks*, *13*(5), 1163–1177. doi:10.1109/TNN.2002.1031947

Schafer, B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. [Kluwer Academic Publishers.]. *Data Mining and Knowledge Discovery*, *5*(12), 115–152. doi:10.1023/A:1009804230409

Spiliopoulou, M. (1999). *Tutorial: Data mining for the Web*. Paper presented at the PKDD'99. Prague, Czech Republic.

Spiliopoulou, M., Mobasher, B., Berendt, B., & Nakagawa, M. (2003). A framework for the evaluation of session reconstruction heuristics in Web usage analysis. *INFORMS Journal of Computing - Special Issue on Mining Web-Based Data for E-Business Applications, 15*(2), 171-190.

Spiliopoulou, N., & Faulstich, L. C. (1998). *WUM: A Web utilization miner.* Paper presented at the International Workshop on the Web and Databases, Valencia, Spain (LNCS 1590, pp. 109-115). Springer.

Stathacopoulou, R., Grigoriadou, M., & Magoulas, G. D. (2003). A neurofuzzy approach in student modeling. In *Proceedings of the 9th International Conference on User Modeling, UM2003* (LNAI 2702, pp. 337-342).

Suryavanshi, B. S., Shiri, N., & Mudur, S. P. (2005). An efficient technique for mining usage profiles using relational fuzzy subtractive clustering. In *Proceedings of the 2005 International Workshop on Challenges in Web Information Retrieval and Integration (WIRI'05)* (pp. 23-29).

Tan, P. N., & Kumar, V. (2002). Discovery of Web robot sessions based on their navigational patterns. *Data Mining and Knowledge Discovery*, *6*(1), 9–35. doi:10.1023/A:1013228602957

Tsakonas, A., Dounias, G., Vlahavas, I. P., & Spyropoulos, C. D. (2002). *Hybrid computational intelligence schemes in complex domains: An extended review* (. *LNCS*, *2308*, 494–511.

Vakali, A., Pokorný, J., & Dalamagas, T. (2004). *An overview of Web data clustering practices.* Paper presented at the EDBT Workshops (pp. 597-606).

Webb, G. I., Pazzani, M. J., & Billsus, D. (2001). Machine learning for user modeling. [Kluwer.]. *User Modeling and User-Adapted Interaction*, *11*, 19–29. doi:10.1023/A:1011117102175

Wong, S. S. C., & Pal, S. (2001). *Mining fuzzy association rules for Web access case adaptation.* Paper presented at the Workshop on Soft Computing in Case-Based Reasoning/International Conference on Case-Based Reasoning (ICCBR_Ol).

Yao, Y. Y. (2005, May 19-21). Web intelligence: New frontiers of exploration. In H. Tarumi, Y. Liand, & T. Yoshida (Eds.), *Proceedings of the 2005 International Conference on Active Media Technology (AMT 2005)*, Takamatsu, Kagawa, Japan (pp. 3-8).

Chapter 8.7
Enhancing the Testability of Web Services

Daniel Brenner
University of Mannheim, Germany

Barbara Paech
University of Heidelberg, Germany

Matthias Merdes
Heidelberg Mobil International GmbH, Germany

Rainer Malaka
University of Bremen, Germany

ABSTRACT

For the foreseeable future, testing will remain the mainstay of software quality assurance and measurement in all areas of software development, including Web services and service-oriented systems. In general, however, testing Web services is much more challenging than testing normal software applications, not because they are inherently more complex, but because of the limited control and access that users of Web services have over their development and deployment. Whereas the developers of normal applications, by definition, have full control over their application until release time, and thus, can subject them to all kinds of tests in various combinations (e.g., integration testing, system testing, regression testing, acceptance testing, etc.), users of Web services can often only test them at run-time after they have already been deployed and put into service. Moreover, users of Web services often have to share access to them with other concurrent users. In order to effectively test Web services under these conditions special measures and approaches need to be taken to enhance their testability. Right from the early phases of development, the testability of services needs to be taken into account and "designed into" services. In this chapter we consider these issues and with the aid of a case study we present a methodology that can be used to enhance the testability of Web services.

DOI: 10.4018/978-1-60566-042-4.ch010

INTRODUCTION

Service-oriented development is based on the idea of building new software applications using software "services," often built and deployed by third party organizations. It therefore assumes a fundamental separation of concerns between *service developers*, who create and offer services with no knowledge of specific applications to which they may be put, and *service users*, who assemble new applications which use the services available on the Internet or in a company's Intranet.

Since they are themselves software applications, services are typically developed and tested using the same basic practices and techniques used to develop normal software applications, and as a result, they can be expected to exhibit the same levels and variations in quality found in the general population of software applications. However, in service-oriented development, the quality of a service-based application is not just based on the inherent quality of the services it uses, it is also dependent on whether they are used or assembled in the correct way. A system assembled from perfectly correct services may still function incorrectly if it uses the services in a different way to that intended; in other words, if the system's understanding of its contract with a service is different to the service provider's.

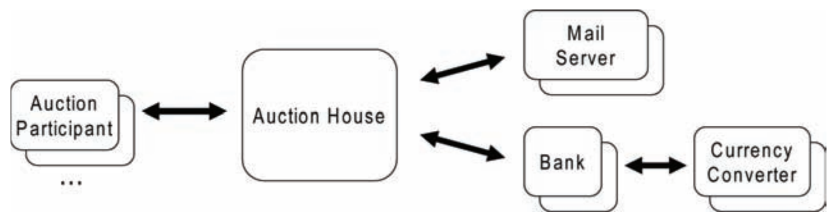
This problem exists in all component-based approaches whenever a new application is created from prefabricated parts. However, the situation is more acute in service-based development because service users have much less access to, and control over, their components than in regular component-oriented development approaches using such technologies as EJB, SPRING, or .NET. If the developers are creating all of their own components in-house, they can test larger assemblies of components while the development process is underway (integration testing). And even when some of the components are purchased from a third party, once the first version of the system has been completed they can still often test the full system under controlled conditions in the development/test environment (system testing). Finally, once a traditional component-based system has been delivered to the customer, it can be further tested in the customer's target environment to determine whether it fulfils the customer's needs (acceptance testing) before being made accessible to end users.

In general, none of these traditional testing activities can be carried out in the normal way when developing a service-oriented system, however. First, the final collection of "components" that make up an application is typically not known until deployment-time, making full integration and system testing in the traditional form impossible. Second, many of the "components" may already provide services to other users, and cannot be shut down even temporarily to participate in traditional testing activities in a controlled environment. A new application will often have to share its components with other applications and cannot assume that these will cooperate while it is in testing mode.

Although subtle, these differences have a fundamental impact on the role and goals of testing in the system development process. Because services are in effect independent, multiuser systems in their own right, new applications have to test them at run-time in a way that combines acceptance testing with integration and system testing. In other words, tests of "components" by applications can no longer be regarded as a pure verification exercise, as has traditionally been the case, but must also include an element of "validation" (Boehm, 1984). An application that is connected to a new service at run-time needs to determine whether the service does the "right thing," just as much as it needs to determine whether it does that thing "right."

Building services so that they can be tested in this way requires changes to the way they are traditionally designed and the way that tests are carried out. The purpose of this chapter is to discuss these changes and to present a methodology

Figure 1. Overview of the case study



intended to improve the testability of services. Testability characterizes how easy it is to test a system based on the information provided by it (Voas & Miller, 1995). In the next section we discuss some of the main issues involved in testing services, and describe the basic principles behind the approach. In Section 2 we explain how the basic functionality of services can be modelled in a practical way using UML, and then in Section 0 we explain how services can be extended to enhance their testability. In the following three sections we then consider how to use these enhanced services to design and apply run-time tests. Finally, in Section 0 we round up with some concluding remarks.

To present the ideas in a coherent way, we use a single case study throughout the chapter. This is a so-called "auction house" system whose job is to enable attendees of an auction to interact and participate electronically using standard mobile devices. Unlike fully electronic auction applications like E-bay, the users of this system need to be actually present at a physical auction. The system supports the auctioneer by allowing users to offer and bid for items and conduct payment transactions electronically.

In Figure 1, the auction house service is represented by the big rectangle in the middle. The figure indicates that the auction house uses two other types of services to help it deliver its functionality to auction participants' bank and mail server. The bank, in turn, uses currency converter services to deliver its functionality. It is possible that some of the internal components of the auction house might also be implemented as components, in which case the same principles can be applied to their interfaces as well. However, here we are not concerned with the internal implementation of the auction house service.

SERVICE TESTING

In the following we discuss two important issues in service testing: testing phases and test isolation.

Testing Phases

Web services are essentially software components "designed to support interoperable machine-to-machine interaction over a network" (Austin, Babir, Ferris, & Garg, 2004) (unless explicitly stated otherwise we regard the terms "service" and "Web service" as being synonymous for the purposes of this chapter). As such, they can be developed and tested according to the principle of *contract-driven design* (Meyer, 1992), which holds that the interface between a service and its users should be documented in terms of the rights and obligations of each party. Like parties in a legal contract, the idea is that each party involved in an interaction should know what it may expect and what it must provide in order for an interaction to be completed successfully. The contract therefore defines the criteria by which an interaction can be judged as having succeeded or failed.

With traditional development approaches, two basic forms of testing are used to check that a systems assembled from separate modules does what it is supposed to: "integration testing" and "acceptance testing." Integration testing is a verification technique that focuses on the testing of successively larger groupings of modules, leading up to the system as a whole (for the purpose of this discussion we regard "system testing" as a special case of integration testing). It is performed in the context of the development environment and involves the actual components that will be delivered in the final product, rather than "stubs" or "mocks" for them. According to the terminology of Boehm (1984), integration testing aims to verify that "we are building the system right" according to some well defined description of what the system should do. Acceptance testing, in contrast, is a validation technique which revolves around the testing of a deployed instance of the system in the target execution environment before it is put into service. In Boehm's terminology acceptance testing aims to validate that "we are building the right system" based on the expectations of the customer or users.

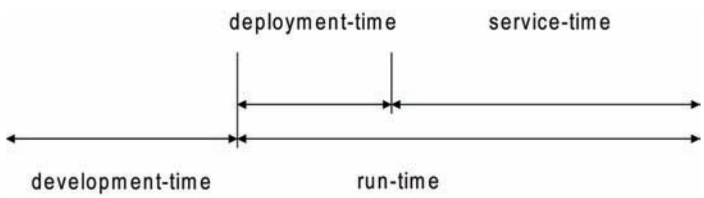
In service-oriented development, when systems are assembled from third party services, many of the assumptions which underpin these traditional testing approaches are no longer valid. Integration testing is no longer applicable in its traditional form because the precise composition of a system is not known at development-time when integration testing is traditionally performed. The notion of testing "the system" as an integrated whole at development-time no longer applies in the traditional sense therefore. Testing at development-time is still important, but its role is to test the implementation of a service's provided interface with representative implementations of its required interfaces. A "representative implementation" of a required service can either be a full working version of the service or a stub/mock which mimics the service for a few chosen test cases. Since these tests are performed at development-time and are exclusively focused on verification against a specification, we simply use the term *development-time testing* for this activity.

In the context of service-oriented development neither the notion of integration testing nor the notion of acceptance testing is fully appropriate in its traditional form. The former is not appropriate because integration testing can no longer be fully performed at development-time as has hitherto been the case. The latter is not appropriate because the testing that is performed at deployment-time should no longer focus just on validation as has traditionally been the case. Instead, the testing activities that are performed at deployment-time also need to include tests to verify the assembly of services against the system's specification. It therefore makes sense to combine the notions of integration and acceptance testing into a single activity known as *deployment-time testing*. Such a deployment-time, service testing activity serves the dual roles of validation and verification of the assembled system in its run-time environment.

For a system whose structure remains constant after initial deployment there is clearly no need to revalidate the system once it is up and running because any tests that have already been executed will not be able to uncover new problems. However, many service-oriented systems do not have a constant structure. On the contrary, an important benefit of service-oriented development is that it allows the structure of a system to be changed while it is running. If a change is made, then clearly tests performed at deployment-time may no longer be valid.

The notions of development-time and deployment-time testing are therefore not sufficient to cover the full spectrum of testing scenarios in dynamically reconfigurable service-based systems. We need to add the notion of *service-time testing* as well. Service-time tests are carried out once a system has entered service and is delivering value to users (i.e., is being used to fulfil its purpose). Deployment-time and service-time testing both take place at "run-time" in the sense that they are applied to a "running" system in its final execution environment.

Figure 2. Life-cycle phases of a system



The relationship and role of these different phases in the life-cycle of a service-oriented system are summarized in Figure 2. At the highest level of abstraction, two different phases exist, that is, the development phase, in which the system is developed and tested using representative servers in the development environment, and the run-time phase, in which an instance of the system is connected to actual servers and is running in its final execution environment. The run-time phase is divided into two subphases: the *deployment phase* and the *service phase*. In the deployment phase, the system is set up in its initial configuration and starts to run in its execution environment, but it is not yet delivering service to users. This is important because it allows testing activities to be performed under controlled conditions with known assumptions. In the service phase the system has been put into service and is delivering value to users. During service-time the assumptions that held during deployment-time may no longer be valid.

Test Isolation

As with other development abstractions, an important distinction in service-oriented development is the distinction between types and instances. Service types are classifiers in the UML sense. In other words, service types are templates that can be instantiated to create service instances with specified properties. WSDL files actually contain a mix of both concepts. The first part defines the abstract type exposed by the Web services, while the second part specifies the location and identity of an individual instance.

A common misconception about Web services is that they are stateless because they are often used as facades or wrappers to databases or legacy content providers and are not responsible for maintaining state themselves. However, Web services can actually be stateful objects whose behaviour is dependent on the history of interactions. In fact, from a programming point of view (e.g., assuming Java), Web services are like classes which expose a set of methods (operations) and can have their own internal attributes. A single Java class can even be deployed as a Web services.

From the perspective of testing, these properties of Web services create two significant complications:

1. The effects of an operation may not be the same each time it is executed (with the same arguments) due to the state of the service, and
2. The operations exposed by the service may be used by other users while a single user is trying to test it. Thus, the state of the service may not always be known before a service is invoked.

The basic problem is how run-time tests can be performed without having unintended side effects on data or state. In other words, how can the execution of tests and the execution of regular business logic for real users be separated? The testing of a service by one user should not have any effect on the delivery of services to other simultaneous users. To do this it is necessary to intercept each invocation of a service operation and check whether the invocation is part of a test or a regular service request. If the operation is executed from a test, then the test should be *isolated* from the service's response to regular requests.

Web services live in a run-time environment which frees the developer from the burden of developing the functionality which is necessary to interact with the outside world over a network. The Web service is deployed and the functionality offered by the run-time environment is immediately available. This must also include run-time testing functionality built into the run-time environment to support test execution and test isolation. An example for such an environment which the authors have developed is presented by Suliman, Paech, Borner, Atkinson, Brenner, Merdes, et al. (2006).

Several different so-called "test isolation" mechanisms can be identified. One strategy is to prioritize the execution of business logic over the execution of a test. The execution with the lower priority then has to wait. Once the high priority request has been completed, the test is then executed. This strategy can be made more rigorous by ensuring that whenever a regular request is received during the execution of a test, the execution of the test is cancelled and rescheduled.

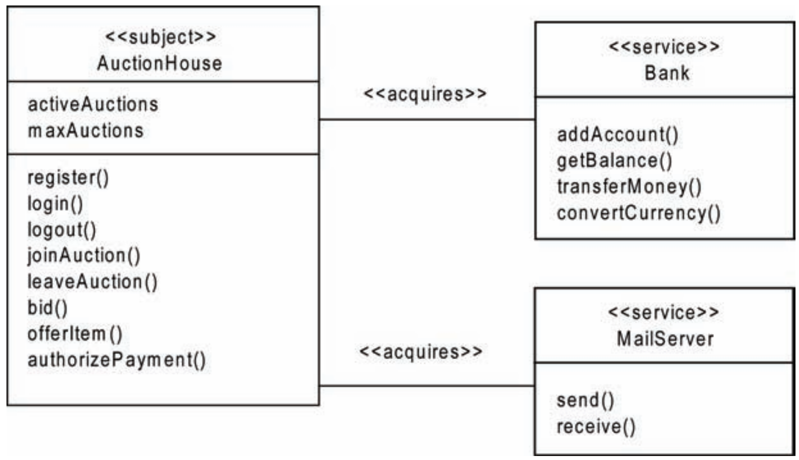
A second possible test isolation strategy is to just clone the service. A client that wishes to test a service can then work on a copy of the service rather than on the real live version that is servicing user requests. This assumes that it is possible to create a copy of the service easily (and automatically). Through a standardized interface it might even be possible to give the service developer the chance to program the cloning themselves. This would be like a service-driven cloning facility.

FUNCTIONAL INTERFACE SPECIFICATION

The first task to be performed in the development of a service is the creation of a high-level specification. The purpose of this specification is to describe the basic functionality offered by the service (the provided interface) and the basic functionality that it needs from other services (the required interface). It also describes the non-functional properties that the service offers and requires. Since a service's specification defines everything that is visible to its clients at run-time, it effectively defines the contract between instances of the service and its run-time clients and servers. In general, a service can support multiple interfaces, but for simplicity we regard these as a single composite interface. Thus, without loss of generality we regard services as having just one provided interface. The same also holds for the required interfaces. Without loss of generality we regard services as having just one required interface.

Ideally a service specification is captured in a platform independent language such as the UML, but any suitable language is acceptable. The approach described in this chapter applies the UML according to the principles of the KobrA method (Atkinson, Bayer, Bunse, Kamsties, Laitenberger, Laqua, et al., 2001), since this provides a clean separation of concerns and documents the externally visible properties of a service in a complete but easy-to-read way. A KobrA service specification involves the creation of three distinct views of a service: the structural view, which describes all structural information that a user of the service needs to be aware of, the functional view, which describes the effects of the operations exported by the service in terms of pre- and post-conditions, and the behavioural view, which describes the allowable sequences of operation invocations in terms of externally visible states and state transitions.

Figure 3. Auction house specification class diagram



Structural Specification

The structural view describes the information that potential clients of the service need to be aware of when interacting with the service. Primarily, these are the types of the parameters and return values of the service's operations, but it can also include other information such as the service's position in one or more taxonomies, or any important associations that the service maintains.

In KobrA, the structural view of a service is represented as a class diagram. The structural class diagram for the specification of the auction house service is illustrated in Figure 3.

This diagram shows the auction house service, represented as the class with stereotype <<subject>>, together with its required interfaces, represented as the classes with stereotype <<service>>. The AuctionHouse class lists the operations provided by the auction house service.

Functional Specification

The functional view of the service describes the effects of its operations. In general, one operation specification is created for each operation of the service which describes its behaviour in terms of OCL pre- and post-conditions. Figure 4 shows the operation specification of the bid() operation of the AuctionHouse.

Behavioural Specification

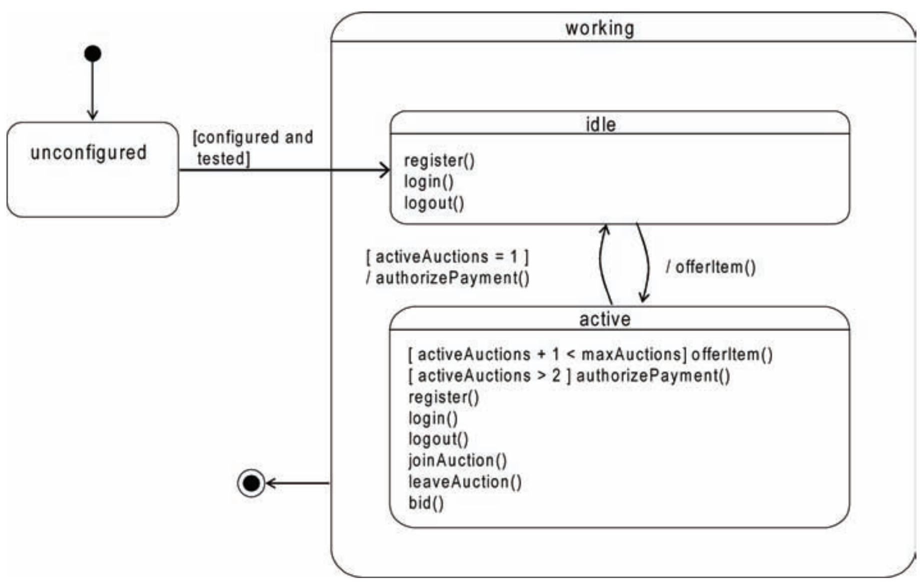
The behavioural view provides a description of any externally visible states that the service exhibits. Figure 5 shows the behavioural model for our auction house example.

The AuctionHouse has two top level states: unconfigured and working. A service enters the unconfigured state when it is first created. In this state it has not been configured and has not been connected to any of its required servers. Only when the configuration operations have been executed, the appropriate dependencies have been set, and the service has been tested the system moves into the working state. This is the state in which the system is servicing user requests. The working state has two substates, that is, idle and active. In the idle state the system is able to respond to a register(), login(), and logout() request, but since there are no active auctions it is unable to respond to requests that relate to active auctions. Only when an offerItem() operation has been invoked and an auction has been started is an AuctionHouse instance is in the active mode so that it can respond to all actions.

Figure 4. Specification of the auction house's bid() operation

Name	**bid()**
Description	
Constraints	the executing bidder is registered at the AuctionHouse
Receives	sessionId : String; bid : Money;
Returns	boolean
Sends	anActivityLogger.logBid(username, bid);
Reads	
Changes	
Rules	
Assumes	bidder is logged in; bidder is not the initiator of the current auction;
Results	if the sent bid is higher than the current highest bid, return true; otherwise return false;

Figure 5. Auction house behavioural model



Extra-Functional Requirements Definition

A service specification will usually have numerous extra-functional requirements in addition to its functional requirements. In general, these can simply be stated in the requirements document alongside the functional requirements discussed in the preceding sections. Probably one of the most important extra-functional requirements is the reliability or equivalent to the failure rate. One way to state such a requirement is to give the maximum allowed failure rate for each operation, as illustrated in Figure 6.

Alternatively, it is possible to define the probability of failure on demand (POFOD) (Sommerville, 2004). For each operation, this value gives the probability that any particular invocation

Figure 6. Reliability requirements

operation	failure rate
AuctionHouse	
register()	0.02
login()	0.03
logout()	0.10
joinAuction()	0.05
leaveAuction()	0.02
bid()	0.002
offerItem()	0.01
authorizePayment()	0.002

of the operation will fail. If the POFOD and/or the likely failure rate has been defined and the relative execution frequencies have been determined, corresponding POFOD and failure rates can be calculated and specified for the service as a whole.

TESTABLE INTERFACE DEFINITION

We refer to the properties of the service that have been defined up to this point as the *functional interface* of the service. This describes the core characteristics of the service as seen by potential users, that is, the contract that it offers. However, not all of the functionality of the service are represented in the operations of the service. Some are defined in the behavioural or structural views and, thus, are not automatically accessible to users at run-time (e.g., states). The next step in the specification process is thus to make sure that *all* behaviour and properties defined in the specification are accessible as operations, not just those operations explicitly identified and specified in the functional view. This is necessary to give users of the service a way of detecting and possibly setting the (logical) states and attributes of a service without having direct access to is its internal implementation (and thus breaking it encapsulation). Providing full access to the contract offered by a service makes it fully testable at run-time.

Information in the behavioural and structural views which is not testable via the basic *functional interface* has to be "functionalized" in the form of additional operations. This gives rise to the so-called *testable interface*. The functional interface and the testable interface together form the *extended interface*.

Interface Extension

The purpose of the testable interface is to make sure all of the semantic properties defined in the service specification are accessible as operations so that they are amenable to testing at run-time. In general this can be done with information that comes from three sources:

1. Logical attributes of the service
2. Logical states of the service
3. Extra-functional requirements

In the case of (1), operations for setting and getting the value of each logical attribute should be added. In the case of (2), operations for setting and getting (or confirming) the logical states of services should be added, and in the case of (3), operations for getting the value of each extra-functional quality-of-service (QoS) property should be added. In addition, an operation should be defined for any other semantic information of any kind in the service specification that affects the run-time behaviour of the service and is in principle measurable.

Furthermore, the testable interface could offer operations which support test isolation. These operations indicate whether the service is sensitive to testing (that means tests and business functionality cannot be performed at the same time because of the risks of destroying the service state) or whether it supports the execution of tests in parallel with the business functionality, for example, by offering a specific clone operation.

In this example it would make sense to define the following additional operations in the testable interface of the AuctionHouse:

- **Logical Attributes**
  - setActiveAuctions()
  - getActiveAuctions()
  - getMaxAuctions()
- **Logical State**
  - setIdle()
  - isIdle()
  - setActive()
  - isActive()
  - isWorking()
- **Extra-Functional Requirements**
  - getAllocatedMemory()
- **Test Isolation**
  - cloneAuctionHouse()

These operations are first class citizens of the service interface, and, thus, need to be added to the specification views developed in the previous sections. Each operation needs an operation specification, and should appear in the behavioural and structural views of the service as well. The specification of the extended interface is known as the extended specification of the service. For example, Figure 7 shows the specification of the setActive() operation.

As a result, the extended structural view is shown in Figure 8, and the extended behavioural view is shown in Figure 9.

Usage Profile Definition

Another part of the specification of a service from the perspective of testability is its so-called usage profile (Juhlin, 1992). This is composed of two parts. One part is a specification of the relative execution frequencies of the operations offered by the service. The other part is a specification of the expected distribution frequencies of the parameter values of the operations.

The simplest way to represent the usage profile is to extend the behavioural view and operation specification identified before. The behavioural view can be enhanced to show execution frequencies by showing the relative probability of each of the exit transitions from each state. The sum of the probabilities of all exit transitions from a state must equal 1. This is shown in Figure 10 below.

The assumed parameter value distribution is best described by adding additional information to the “receives” fields of the operation specifications, as illustrated in Figure 11. In this figure, the bracketed information after each parameter type defines the assumed distribution. Thus, sessionId is a String parameter with Strings values distributed uniformly and bid is parameter of type Money with values distributed according to a Poisson distribution.

RUN-TIME TESTING

Once the extended interface of the service has been fully specified it can be turned into an executable

Figure 7. Operation specification for setActive()

Name	**setActive()**
Description	sets the service to the “active” state
Results	the service is in the “active” state

Figure 8. Extended structural view

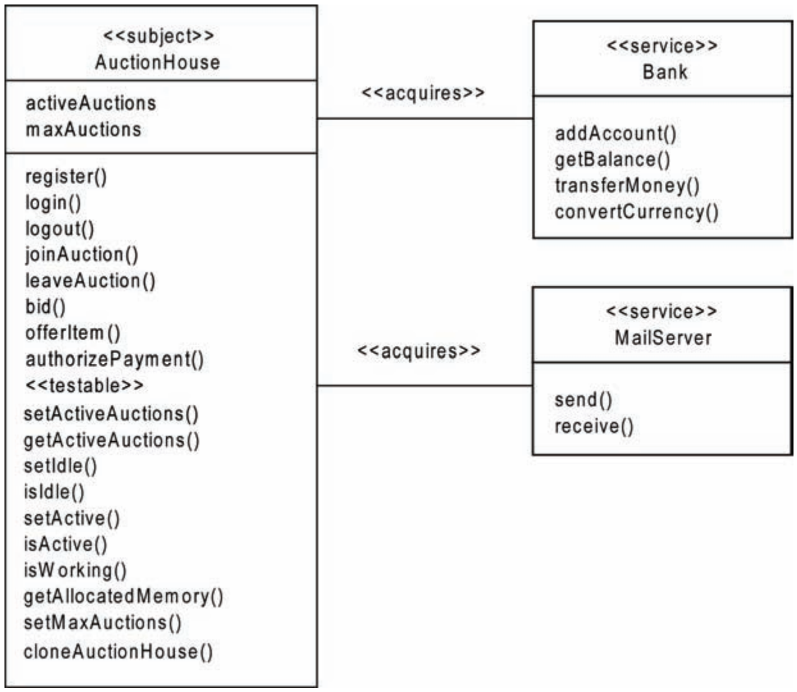


Figure 9. Extended behavioural view

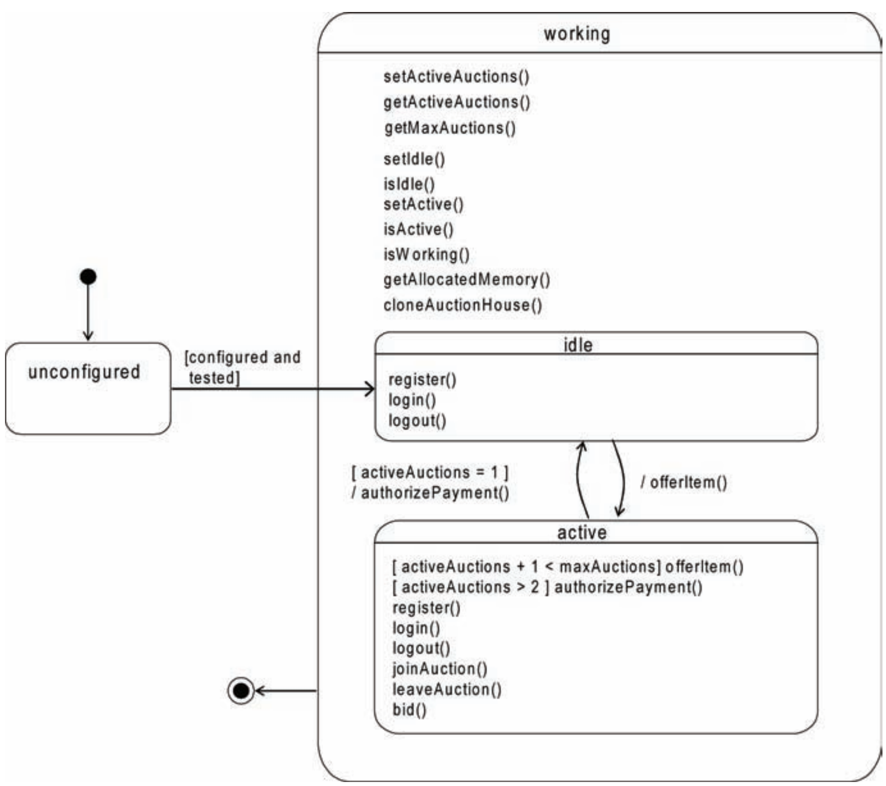


Figure 10. Usage profile – state machine form

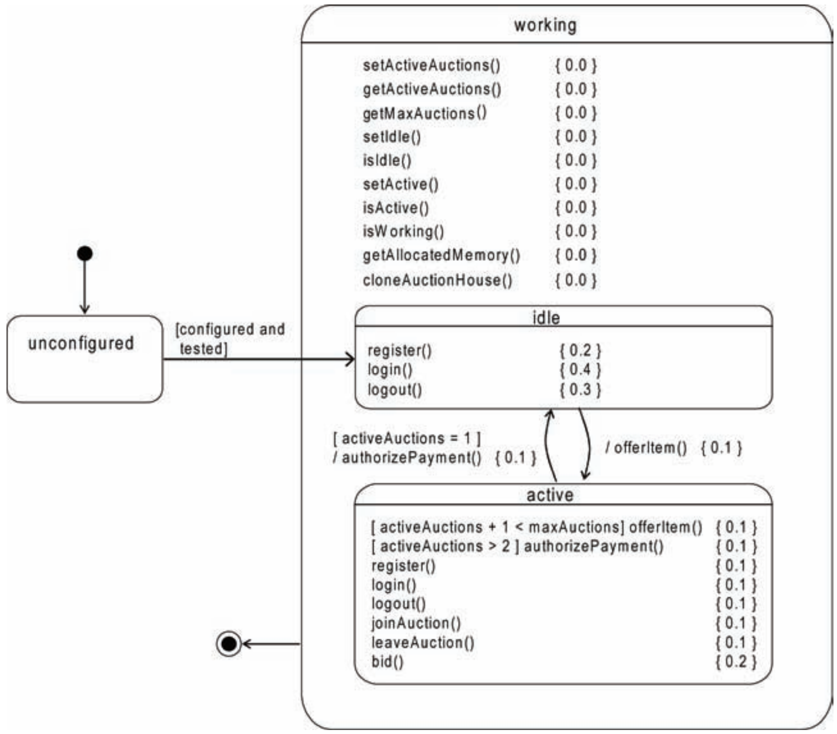


Figure 11. Extended specification of the AuctionHouse's bid() operation

Name	**bid()**
Description	
Constraints	the executing bidder is registered at the AuctionHouse
Receives	sessionId : String (Uniform); bid : Money (Poisson);
…	…

system using the normal, widely available tools for implementing services. While under development it should also be tested using the typical defect testing techniques used to test software systems with a well defined interface. Any errors discovered will typically be fixed until a fully functional service is available, which is correct w.r.t. to the usual range and number of test cases. If the service is made available as a Web service, which is usually the case, the implementation process will also result in a WSDL file of the form sketched in Figure 12, which describes the extended interface to the service in a form that can be invoked via SOAP messages.

Developers of new applications can now use this service in their applications. However, in order to have confidence that the service will fulfil their needs, it is advisable for them to test the service using the features offered by the full extended interface. These tests can either be performed during the development of the application using a small, dedicated test harness, or they can be "built-in" to the client application and tested at run-time.

Figure 12. WSDL file of the auction house service

```xml
- <wsdl:definitions targetNamespace="http://auctionhouse.interfaces.application.morabit.org">
  - <wsdl:types>
      <!-- ... -->
    </wsdl:types>
    <!-- ... -->
  - <wsdl:message name="bidResponse">
      <wsdl:part name="bidReturn" type="xsd:boolean"/>
    </wsdl:message>
  - <wsdl:message name="bidRequest">
      <wsdl:part name="in0" type="soapenc:string"/>
      <wsdl:part name="in1" type="tns5:Auction"/>
      <wsdl:part name="in2" type="tns2:Money"/>
    </wsdl:message>
  - <wsdl:portType name="AuctionHouse">
      <!-- ... -->
    - <wsdl:operation name="bid" parameterOrder="in0 in1 in2">
        <wsdl:input message="impl:bidRequest" name="bidRequest"/>
        <wsdl:output message="impl:bidResponse" name="bidResponse"/>
        <wsdl:fault message="impl:CurrencyNotSupportedException" name="CurrencyNotSupportedException"/>
      </wsdl:operation>
    </wsdl:portType>
  - <wsdl:binding name="localhostSoapBinding" type="impl:AuctionHouse">
      <wsdlsoap:binding style="rpc" transport="http://schemas.xmlsoap.org/soap/http"/>
      <!-- ... -->
    - <wsdl:operation name="bid">
        <wsdlsoap:operation soapAction=""/>
      - <wsdl:input name="bidRequest">
          <wsdlsoap:body encodingStyle="http://schemas.xmlsoap.org/soap/encoding/" namespace="http://auctionhouse.interfaces.application.morabit.org" use="..."/>
        </wsdl:input>
      - <wsdl:output name="bidResponse">
          <wsdlsoap:body encodingStyle="http://schemas.xmlsoap.org/soap/encoding/" namespace="http://auctionhouse.interfaces.application.morabit.org" use="..."/>
        </wsdl:output>
      - <wsdl:fault name="CurrencyNotSupportedException">
          <wsdlsoap:fault encodingStyle="http://schemas.xmlsoap.org/soap/encoding/" name="CurrencyNotSupportedException" namespace="http://..." use="..."/>
        </wsdl:fault>
      </wsdl:operation>
    </wsdl:binding>
  - <wsdl:service name="AuctionHouseService">
    - <wsdl:port binding="impl:localhostSoapBinding" name="localhost">
        <wsdlsoap:address location="http://localhost"/>
      </wsdl:port>
    </wsdl:service>
</wsdl:definitions>
```

In this section we discuss the various issues and strategies involved in designing such built-in, run-time tests. To put these tests into use, mechanisms must exist which define how these tests can be associated with a service, when the tests should be executed, and how a user can react to a failed test (Atkinson, Brenner, Paech, Malaka, Borner, Merdes, et al., 2006).

Analysis and Ranking

The failure of a server to fulfil a contract, as perceived by the application client, can have various consequences ranging from nothing of particular concern to the complete inability of the service to fulfil its specification. The first step in the design of the tests is therefore to analyse the likelihood and consequences of service failure to the application under development (which may be another service). The identified risks are balanced against the cost of detecting the risk and reacting to it at run-time. Based on a trade-off analysis, tests are designed for the failures with the highest risk-cost ratio. Initially, a risk and cost analysis is performed to analyze the nature of an application's dependency on a service. This is driven by the following questions:

1. What are the types of contract failures that could occur and what is the likelihood of each type occurring?
2. For each failure type, what is the consequence of that failure occurring in terms of the service's ability to provide its own

services and in terms of possible effects on the overall system in which an instance of the service might be deployed?
3. What are the estimated costs of detecting this failure in terms of the ability of the application to deliver its service?
4. What countermeasures are feasible?

For example:

A. Risk: The auction house does not meet its functional contract, for example, does not send confirmation e-mail to the right address.
   1. Low likelihood because this is standard functionality
   2. Failure can be serious for high priority e-mails
   3. Estimated cost of detection is low; could be checked when the service sends an e-mail to itself via the external e-mail service
   4. Stop using the service and choose a different service provider
B. Risk: The auction house does not meet its reliability contract.
   1. Fairly high likelihood in case of high load
   2. Failure can be serious
   3. High estimated cost as this requires quantitative testing
   4. Implement checking algorithms

The risks are ranked according to their likelihood of occurrence and their potential impact, taking into account the cost of detection and reaction. Thus, in the example the test of e-mail functionality would be ranked high since its impact is high, and the cost manageable. Also, despite the high cost, the testing of the auction house's reliability is ranked high because of the high impact.

Contract Test Case Definition

Once the goal and relative priority of each risk have been identified, the next step is to define contract test cases (CTCs) to uncover the most highly ranked contract risks. The aim of these CTCs is to uncover at run-time a failure in the service's ability to fulfil the part of the contract that leads to the risk in question.

There are two basic criteria for defining CTCs depending on whether quantitative or qualitative information is required from the execution of run-time tests. Qualitative tests return binary (pass/fail) information depending on whether or not a service fulfils a contract as understood by the client application. This form of run-time test is therefore the most basic and is the most widely applicable. Quantitative tests return a numeric measurement of the level of reliability to which a service implements the contract as understood by client application. This kind of run-time test is therefore only needed if the contract has an associated reliability requirement.

Both forms of testing rely on the ability to determine whether a particular invocation of a service's operation succeeds or fails from the perspective of the client. There are three basic ways in which such an invocation can be judged to have failed:

1. The operation completes, but returns a result that was not the expected one;
2. The operation does not complete and returns some indications to the caller that it was unable to do so (e.g., an error message); and
3. The operation does not complete within a required period of time.

In principle, all three forms can be used in both quantitative as well as qualitative testing. However, since forms (2) and (3) do not require an expected result to be determined, they are particularly suited to quantitative testing. The creation of expected results for the first form of failure has traditionally been one of the biggest stumbling blocks to quantitative testing because it is difficult if not impossible to do automatically.

Qualitative Contract Test Case Definition

The goal of qualitative CTCs is to uncover mismatches between the service provider's understanding (i.e., interpretation) of its contracts and those of the application using it. The CTCs are therefore designed to maximize the chances of uncovering contract understanding mismatches at run-time. Qualitative test cases are similar to the test cases defined in the functional and structural test case design for defect testing. However, since it is assumed that the service has already been tested at development-time to detect coding defects, the focus is on test cases which reveal problems in the service's behaviour due to dependencies on its environment or due to implicit assumptions. In general, the types of "misunderstanding" of a contract that can lead to perceived failure of a service fall into the following two categories, that is, syntactical misunderstandings and semantic misunderstandings.

Syntactical misunderstandings can arise in several parts of a provided interface. In statically typed languages such as Java, the compiler will return an error message when, for example, either a method name does not exist or parameter types of the called method do not match. It is possible to invoke an operation of a Web service without taking the nature of its interface inaccount, but in most cases, invocations are at least checked against, if not created from, the signature information in the WSDL file (such as Figure 12).

Thus, in the AuctionHouse case study, if the AuctionHouse service attempted to invoke the transferMoney() operation of the Bank service with the following type profile

transferMoney(int, int, double)

whereas the operation defined by the Bank service had the following signature

transferMoney(String, String, double).

the Java compiler will return an error. Such a misunderstanding can quite easily occur when using Web services. Since this depends on the programming language used to implement the application or using service we do not further discuss this type of misunderstanding here.

Semantic misunderstandings come from an operation's input parameters. When two or more parameters have the same type the correct order cannot be inferred unambiguously from the WDSL file. This means that the client of a service might assume a different order of input parameters than the called service, probably leading to a nunexpected result. So if, for example, the Bank has a transferMoney() operations with the following intended meaning of its parameters

transferMoney(String fromAccount, String toAccount, double amount)

while the client expects the opposite

transferMoney(String toAccount, String fromAccount, double amount)

the invocation of the operation fails. Even though this kind of semantic misunderstanding can only happen with parameters of the same type, it happens quite often. This is something that can easily be checked via a run-time test. Test cases should be defined for all operations where such parameter swapping is likely. All possible permutations of the parameters need to be checked.

When a third party service is used, it often happens that the client does not know what the valid input values are, and, thus, invokes the operation with *invalid input parameter values*. It is also unknown to the client how the used service will react to the invalid values. Services differ in the way they handle such cases. Some services use default values, others return an error message. This leads to a *different understandings of exception handling*. Therefore, the client should create tests that invoke operations with invalid (boundary) values and check the service's reaction.

If no error occurs during the execution of an operation, the called service will return an output value. The returned output and expected result can differ in various ways. An important distinction in this regard is between functional and extra-functional differences. A functional difference is, for example, the *accuracy of the output*. This kind of difference can occur when the bank uses an additional (third-party) service for converting a certain amount of money between two currencies. The output might be perfectly correct, but whereas the client expected a precision of three decimal places the bank returned a precision of only two. In contrast, extra-functional differences relate to *quality properties* such as response time. When the response time of the AuctionHouse for the bidding operation is too long, it becomes unusable. However, the quality is dependent on the whole system (environment) so that no general conclusions can be drawn from one single test case. Therefore, multiple test cases need to be run, the quality property measured (here the response time), and the average value determined. If this average value lies within a certain range, it will be classified as acceptable.

It can also happen that the service called by the client functions correctly, but still delivers a wrong output. This is the case when the called service depends on other servers in order to fulfil its provided functionality. Unfortunately, such a server usually cannot be detected directly. When the AuctionHouse uses a Bank service to access certain functionality and the bank itself uses a CurrencyConverter service then an error in the interaction between the Bank and the CurrencyConverter cannot be detected by the AuctionHouse. Constructing test cases that invoke a chain of operations is a possibility. The problem is that there might not be enough information available to construct such dependence chains.

Quantitative Contract Test Case Definition

Quantitative contract test cases are used to determine, to a given level of confidence, whether a service delivers its functionality with a level of reliability. A prerequisite for quantitative tests of a service is an extended specification which includes usage profile information of the kind outlined in Section 0. We therefore assume that all provided services have been specified according to the approach.

To attain a reasonable level of confidence in the reliability bound, quantitative tests usually require many more test cases than qualitative tests. The minimum number of test cases that must be executed is a function of the desired reliability threshold and the desired level of confidence. Following Brenner, Atkinson, Malaka, Merdes, Paech, and Suliman (2007) the number of required test cases can be calculated based on the given values for failure rate and confidence. The failure rate f for each operation was already specified earlier in the extra-functional requirements (Section 0). For the bid() operation the failure rate is $f = 0.002$.

In order to calculate the minimum number of test cases, the confidence in the test result is needed. Clearly, only one test case could be run and if no error is detected with this test case, the failure rate is said to be 0 and, thus, better than required. But the confidence in this test result cannot be high. On the other hand, it is obvious that to obtain higher confidence, the more test cases need to be run. In our example we assume a target confidence (c) of 95%.

When the number of test cases, n, is greater than 100 the calculation can be simplified. Then the Poisson distribution can be assumed and the term $\ln(1 - c)$ can be looked up in a table containing the values for the parameter a of the Poisson distribution. For $c = 0.95$, this is approximately $a = 3$, implying that during the test case execution no errors occur. So,

$$n = a / f = 3 / 0.002 = 1500$$

This means that 1500 test cases need to run without an error before we can be 95% sure that the failure rate of the bid() operation is smaller or equal to 0.002. If we accept one error during the test case execution, the minimum number of test cases required to hold the assumptions f = 0.002 and c = 0.95 raises to 2,400 (= 4.8 / 0.002). This provides an estimate for the required numbers of test cases.

RELATED WORK

The idea of testing services at run-time to determine if they are acceptable is related to the notion of quality of service. Several approaches for building Web services focus on QoS, such as availability, security, and throughput (Menasce, 2002). Typically, a global model of the interactions is created which is then evaluated so that adaptation strategies can be defined (Porter & Katz, 2006). However, the creation of such a global model takes a lot of effort and in dynamic contexts, such as Web services, it will change very quickly. Menasce, Ruan, and Gomaa (2004) propose an architecture where the server monitors its own quality and adapts accordingly. Our proposed approach differs from these approaches in that it focuses on functional qualities, that is, on whether the functionality provided by the service fits to the functionality required by the service consumer. This question is often considered as part of service matching. However, as always with software, additional testing is needed to uncover defects.

Numerous component-based development approaches have also developed techniques for enhancing systems testability using metadata provided by the component supplier, such as component behaviour specification, test cases together with coverage data, quality of service information, as well as specific information on code such as dependencies between variables (Orso, Harrold, & Rosenblum, 2000). This information is provided, for example, in the form of testable beans which comprise a testing interface and a traceability interface (Gao, Tsao, & Wu, 2003). The former enables the test to be set up, executed, and evaluated. The latter enables access to the history of test results. The wrapper approach separates the code for the metadata clearly from original code (Edwards, 2001). The notion of built-in tests enhances the testing interface with test cases (Gross, 2005). The test cases can be applied to the provided interface of the component (self-test) (Wang, King, Patel, Patel, & Dorling, 1999), or to the interface of its server components (contract test). This idea has been adapted to Web services where the service provider provides test cases for the service which can be applied by the service consumer to the service (Bruno, Canfora, DiPenta, Esposito, & Mazza, 2005).

Today, the main challenge of testing is still to define a test strategy which minimizes the relationship between test costs and defect costs. Such a strategy comprises a test focus (what should be minimized through the test), the test intensity (how much test effort should be spent on each risk), the test plan (who tests what), the test case definition method, the test case order, the test end criteria (when to stop testing), and the reaction to the test result (Borner, Illes, & Paech, 2007).

In component- and service-oriented systems, the main risk is a misunderstanding between the service and its clients. These risks are integration test risks, and can be classified according to typical defects. Wu, Pan, and Chen (2001) use such defects classification together with a global component interaction model to derive the test cases. Similarly, architecture-based approaches like that of Bertolino, Corradini, Inverardi, and Muccin (2000) and Muccini, Bertolino, and Inverardi (2004) use a global interaction model. As mentioned above, such global models are costly and, thus, it is important to identify typical defects of component interactions from the viewpoint of individual components.

The test cases can either be defined by hand, generated from models, or generated from the test history. Smythe (2006) discusses generation from models at development time. Run-time generation would save space, but requires an execution. In our method we do not give specific support for run-time generation. This is still a question for further research.

Another issue is *when* test cases are executed at run-time. Web service quality assurance approaches so far concentrate on run-time tests triggered and evaluated by humans, for example, after a new service is released. However, this will not be sufficient for a true service-oriented architecture where client-service relationships change very often. Merdes, Malaka, Suliman, Paech, Brenner, and Atkinson (2006) show how tests can be triggered and evaluated by the components themselves. This requires specific testing times and test reactions.

Obviously, the run-time test strategy will always be based on heuristics. Therefore, it is important to monitor and evaluate the test execution. This topic is still in its infancy, even for development-time testing. For run-time tests, evaluation strategies such as mutation test (Delamaro, Maldonado, & Mathur, 2001) and benchmarks (Zhu, Gorton, Liu, & Bui, 2006) need to be adopted. It is necessary to distinguish the quality of individual services as well as of the overall system (where the test strategies of several services interact). So far, our method does not provide specific support for this evaluation.

SUMMARY AND CONCLUSION

We have a presented a method for maximizing the testability of services by systematically ensuring that all information in their model-based specifications are included in their interfaces. As explained, there are numerous other approaches for designing tests of Web services based on their published interfaces, but these all "start" *after* the published interface has been fixed and encoded in a WSDL document or something similar. The novelty of the approach outlined in this chapter is that it addresses the problem of testability *before* the precise interface of the component has been fixed. In fact, it works by extending the basic functional interface beyond what it would normally be in normal methods.

The methodology exploits the fact that complex services are systems which are (or should be) designed using a systematic process of specification, realization, and implementation. More specifically, it exploits the fact that not all the information in the specification of a service (and thus in the contract) takes the form of operations that are accessible at run-time by clients. Semantic information such as externally visible states and attributes are often encapsulated within the body of service implementation, where they are accessible to development-time tests but not third-party client tests at run-time.

Since the main contribution of the approach is how to construct the concrete (i.e., WSDL) interface of a service rather than how to use them, it is compatible with most other approaches to service testing. In fact, we believe it naturally complements them. As part of the MORABIT project (2007), we have developed a prototype infrastructure to directly implement the kind of run-time tests supported in the approach and are currently in the process of building a tool to support the view-based specification of services upon which it is based.

REFERENCES

Atkinson, C., Bayer, J., Bunse, C., Kamsties, E., Laitenberger, O., Laqua, R., et al. (Eds.). (2001). *Service-based product line engineering with UML*. Addison Wesley.

Atkinson, C., & Brenner, D. (Eds.). Paech, B., Malaka, M., Borner, L., Merdes, M., et al. (2006). *The MORABIT development method* (*MORABIT Deliverable M3*). Mannheim, Germany: University of Mannheim.

Austin, D., Babir, A., Ferris, C., & Garg, S. (2004). Web services architecture requirements. *W3C working group*. Retrieved October 27, 2007, from http://www.w3.org/TR/wsa-reqs/

Bertolino, A., Corradini, F., Inverardi, P., & Muccini, H. (2000). *Deriving test plans from architectural descriptions.* Paper presented at International Conference on Software Engineering, Limerick, Ireland.

Boehm, B. (1984). Verifying and validating software requirements and design specifications. *IEEE Software*, ▪▪▪, 205–218.

Borner, L., Illes, T., & Paech, B. (2007). *The testing process: A decision based approach.* Paper presented at the Second International Conference on Software Engineering Advances (ICSEA 2007), Cap Esterel, France.

Brenner, D., Atkinson, C., Malaka, R., Merdes, M., Paech, B., & Suliman, D. (2007). Reducing verification effort in component-based software engineering through built-in testing. *Information Systems Frontiers*, *9*(2-3), 151–162. doi:10.1007/s10796-007-9029-4

Bruno, M., Canfora, G., DiPenta, M., Esposito, G., & Mazza, V. (2005). *Using test cases as contract to ensure service compliance across releases.* Paper presented at the International Conference on Service Oriented Computing (LNCS 3826).

Delamaro, M. C., Maldonado, J. C., & Mathur, A. P. (2001). Interface mutation: An approach for integration testing. *IEEE Transactions on Software Engineering*, *27*(3), 228–247. doi:10.1109/32.910859

Edwards, S. H. (2001). A framework for practical, automated black-box testing of component-based software. *Software Testing . Verification and Reliability*, *11*, 97–111. doi:10.1002/stvr.224

Gao, J. Z., Tsao, H. S. J., & Wu, Y. (Eds.). (2003). *Testing and quality assurance for component-based software*. Artech House Computing Library.

Gross, H. G. (Ed.). (2005). *Component-based software testing with UML*. Springer.

Juhlin, B. D. (1992). *Implementing operational profiles to measure system reliability*. Paper presented at Third International Symposium on Software Reliability Engineering, Research Triangle Park, NC.

Menasce, D. A. (2002). QoS issues in Web services. *IEEE Internet Computing*, ▪▪▪, 72–75. doi:10.1109/MIC.2002.1067740

Menasce, D. A., Ruan, H., & Gomaa, H. (2004). *A framework for QoS-aware software components*. Paper presented at Workshop on Software and Performance, San Francisco.

Merdes, M., Malaka, R., Suliman, D., Paech, B., Brenner, D., & Atkinson, C. (2006). *Ubiquitous RATs: How resource-aware run-time tests can improve ubiquitous software systems*. Paper presented at 6th International Workshop on Software Engineering and Middleware, Portland.

Meyer, B. (1992). Applying design by contract. *IEEE Computer*, *25*(10), 40–51.

MORABIT project. (2008). Retrieved October 27, 2007, from http://www.morabit.org/

Muccini, H., Bertolino, A., & Inverardi, P. (2004). Using software architecture for code testing. *IEEE Transactions on Software Engineering*, *30*(3), 160–171. doi:10.1109/TSE.2004.1271170

Orso, A., Harrold, M. J., & Rosenblum, D. S. (2000). *Component metadata for software engineering tasks.* Paper presented at 2nd International Workshop on Engineering Distributed Objects (EDO), Davis.

Porter, G., & Katz, R. H. (2006). Effective Web service load balancing through statistical monitoring. *Communications of the ACM*, *49*(3), 49–54. doi:10.1145/1118178.1118201

Smythe, C. (2006). *Initial investigations into interoperability testing of Web services from their specification using the unified modelling language*. Paper presented at the International Workshop on Web Services Modeling and Testing (WS-MaTe), Palermo, Italy.

Sommerville, I. (Ed.). (2004). *Software engineering*. Addison Wesley.

Suliman, D., Paech, B., Borner, B., Atkinson, C., Brenner, D., Merdes, M., et al. (2006). *The MORABIT approach to run-time component testing*. Paper presented at Second International Workshop on Testing and Quality Assurance for Component-Based Systems (TQACBS 2006), Chicago.

Voas, J. M., & Miller, K. W. (1995). Software testability: The new verification. *IEEE Software*, 17–28. doi:10.1109/52.382180

Wang, Y., King, G., Patel, D., Patel, S., & Dorling, A. (1999). On coping with real-time software dynamic inconsistency by built-in tests. *Annals of Software Engineering*, *7*, 283–296. doi:10.1023/A:1018990322378

Wu, Y., Pan, D., & Chen, M. H. (2001). *Techniques for testing component based systems.* Paper presented at 7th IEEE International Conference on Engineering of Complex Computer Systems (ICECCS'01), Skovde, Sweden.

Zhu, L., Gorton, I., Liu, Y., & Bui, N. B. (2006). *Model driven benchmark generation for Web services*. Paper presented at ICSE-Workshop Service-Oriented Software Engineering, Shanghai, China.

This work was previously published in Managing Web Service Quality: Measuring Outcomes and Effectiveness, edited by K. Khan, pp. 223-244, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 8.8
Making the Web Accessible to the Visually Impaired

Simone Bacellar Leal Ferreira
Universidade Federal do Estado do Rio de Janeiro, Brazil

Denis Silva da Silveira
Programa de Engenharia de Produção - COPPE/UFRJ, Brazil

Marcos Gurgel do Amaral Leal Ferreira
Holden Comunicação Ltda, Brazil

Ricardo Rodrigues Nunes
Universidade Federal do Estado do Rio de Janeiro, Brazil

ABSTRACT

Accessibility is the possibility of any person to make use of all the benefits of society, including the Internet. As the interfaces are typically graphic, sites can be an obstacle for visually impaired persons to access. For a site to be accessible to blind persons it's necessary the information contained in the visual resources be reproduced by means of an "equivalent" textual description, capable of transmitting the same information as the visual resources. This study is aimed at identifying and defining usability guidance compliant with accessibility W3C directives that can facilitate the interaction between visually impaired and Internet and still guarantee sites with understandable navigation content. Towards this end an exploratory study was conducted, comprised of a field study and interviews with visually disabled people from Instituto Benjamin Constant, reference center in Brazil for the education of visually impaired persons, in order to get to know these users better.

DOI: 10.4018/978-1-60566-386-9.ch014

INTRODUCTION

Accessibility is the term used to indicate the possibility of any person to make use of all the benefits of society, among which, the use of the Internet. (Nicholl, 2001). Digital accessibility is more specific and refers only to access to computer resources; accessibility to the Internet is the right to use the resources in the worldwide computer Web and accessibility to the Web, or *e-accessibility*, referring specifically to the Web component (Sales, 2003).

The Web component plays a fundamental role in the innovation that the Internet represents in the daily lives of persons with special needs; it facilitates the lives of these people as it allows them to create new ways of relating to others and performing activities previously unattainable (Takagi, 2004) and (Petrie, 2006). But getting digital accessibility is no simple matter; it requires organizations to adapt their resources in order to make the use of the computer accessible to any person (http_1).

In order to be accessed by visually impaired users, the graphic interface of computer systems should be designed with an "equivalent" textual description. These "equivalent" interfaces should be built in such a way that when accessed by support technology, they continue to provide "friendly" interaction, i.e., an interaction focused on usability. Hence, the present study is aimed at identifying and defining usability guidance compliant with accessibility laws, which may facilitate the interaction between those visually impaired and the Web, guaranteeing sites with understandable navigation content. This research is focused on Brazil`s necessities. To achieve this end, a Field work was conducted at the Instituto Benjamin Constant (IBC), an agency of the Ministry of Education of Brazil, and a center of excellence and national reference in matters related to studies of visual impairment (http_6

ACCESSIBILITY TO THE WEB OR E-ACCESSIBILITY

Digital accessibility refers to access to any Information Technology resource, whereas the term accessibility to the Internet is used, widely speaking, to define universal access to all components of the worldwide computer Web, such as chats, e-mail, and so on. The term Web accessibility, or *e-accessibility*, specifically refers to the Web component, which is a set of pages written in HTML language and interconnected by links to the hypertext (Sales, 2003), (Modelo, 2005) and (Nevile, 2005).

Aimed at making the Web accessible to all, W3C (the World Wide Web Consortium), an international committee that regulates matters linked to the Internet, created, in 1999, the WAI (Web Accessibility Initiative), made up of work groups intent on producing guidance to guarantee Web content accessibility to people with disabilities and to people accessing the Web under special conditions related to environment, equipment, navigator and other Web tools (Nevile, 2005), (http_5) and (Enap, 2007).

The members of W3C/WAI put together "W3C Accessibility Guidelines" (WCAG 1.0); this document is the first version for Accessibility to Web Content, released in May 1999, and has been the main reference to Web accessibility until today (http_5). In Brazil, accessibility began to be a part of public policy in the year 2000, when Federal Laws no. 10,048 dated November 8 2000, prioritizing services rendered to people with special needs, and no. 10,098 dated December 19 2000, establishing norms and criteria to guarantee accessibility were promulgated (Enap, 2007). In December 2004 these laws were regulated by decree no. 5,296 that initially established a 12-month deadline for all public administration or public interest sites to undergo an accessibility process; this deadline was subject to prorogation (Queiroz, 2007).

In order to define accessibility guidance at all levels, from physical to virtual spaces, ABNT's CB-40 Committee was put in charge of comparing accessibility norms in various countries and analyzing the guidelines proposed by W3C. As a result, a Brazilian Accessibility Model was designed (e-MAG) so as to generate a set of recommendations that could standardize and harmonize the accessibility process for Brazilian Government sites, enabling easy installation, thereby coherent with Brazilian needs and in conformity with international standards (Model, 2005) and (http_1).

Importance of Internet and Web Accessibility for those with Visual Impairment

The Web plays a fundamental role in the innovation that the Internet represents in the daily lives of those with visual impairment, making their lives easier; it allows them to establish new relationships, find job opportunities and forms of entertainment (Petrie, 2006) and (Queiroz, 2007).

Upon accessing a site, a user with normal eyesight uses a Browser. However, a blind or partially sighted person accessing the Internet would require support technology connected to the Browser, consisting of software called "screen readers", associated to other programs called "voice synthesizers".

Though important, digital accessibility is no simple matter. People with disabilities have sensorial and motor limitations which must be compensated for, one way or another, so as to enable their access to computer resources. With this in mind, organizations need to adapt their systems so that one single computer can be used by any person whatsoever (Harrison, 2005). The problem is that this adaptation requires technical expertise and specialized help, and this is why organizations very often do not make the needed effort to introduce accessibility procedures.

Levels of Accessibility

W3C Accessibility Guidelines (WCAG 1.0) proposed a set of fourteen directives for the Accessibility of Web Content. These directives deal with two generic themes: assure that the sites are accessible in a harmonious way and produce sites with understandable navigational content (http_5). WAI defined verification points for directives; each verification point was attributed a priority level, based on accessibility impact. Three levels of accessibility were defined; the Brazilian model also adopted these same priority levels (Model, 2005).

- **Priority level 1:** Norms and requirements related to verification points that Web designers must comply with so as not to make access unattainable to any group of users.
- **Priority level 2:** Norms and requirements related to verification points that Web designers must comply with so as not to hinder the access of any group of users.
- **Priority level 3:** Norms and recommendations that Web designers can comply with so as not to hinder access to saved files.

Programs for Accessibility Evaluation

Based on W3C/WAI recommendations, programs were designed to evaluate the level of accessibility to the site. These programs detect the HTML code and analyze content, verifying whether it is in compliance with the established set of rules or not; finally, they write reports listing the problems that need to be addressed for the site to be considered accessible (Spelta, 2003).

Some of this software is worth highlighting: *Bobby* (designed by the "Watchfire Corporation"), and *Lift* (designed by "Usablenet"). In Brazil, the *daSilva* program was designed to evaluate sites according to the rules established by WCAG and by e-MAG (http_4). This program was designed by "*Acessibilidade Brasil*", a "Public Interest Corporation" (OSCIP), whose mission is to develop studies for the social and economic insertion of people with impairments (http_1).

METHODOLOGY

The study, exploratory in nature, was carried out in three stages: (*a*) selection of the category of users; (*b*) bibliographic and documental research; (*c*) field work. These stages were accomplished concurrently.

The research work aimed at identifying and defining usability directives that are aligned with accessibility laws and which of these might facilitate the interaction between those visually impaired and the Internet, as well as guaranteeing sites with understandable navigation content.

Stages

a. **Selecting the category of users:** Users with visual impairment were chosen as the object of study of the present work; this decision was made based on the fact that the Internet has done much to contribute to improving the quality of life of those visually impaired, allowing them not only access to information that was previously only attainable with the help of another person, but also providing them with other facilities (Harrison, 2005).

b. **Biographical and documental research**: initially, we sought to understand the principles of accessibility and its implications for Internet sites. During this stage, some institutions provided different software destined for visually impaired users. This software was used to navigate in "common" sites, such as newspapers, and make a deeper observation and analysis of the various aspects brought up in the literature.

c. **Field work**: Field work was conducted at the Instituto Benjamin Constant (IBC), an agency of the Ministry of Education, founded in Rio de Janeiro in 1854 under the name of *Imperial Instituto dos Meninos Cegos*. IBC has become a center of excellence and national reference in matters related to studies of visual impairment. Its main aim is to promote the education and integration of visually impaired persons within a greater framework (http_6). During the field work, which took three months, different sectors of the institute were observed. In addition, several informal interviews and six in-depth interviews were conducted with employees, students and former students at the institution, most of whom are visually impaired and work there nowadays.

MODELS

Mental models are representations existing in the minds of people, which are used to explain, simulate, predict or control objects in the world. These representations are externalized through conceptual models. The elaboration of a user's conceptual model depends on the previous knowledge and experience of each person and is based on the expectations, aims and understanding of the user with regard to the system. Users create models based on "objects" they already know from their daily activities; they try to relate the computer elements to these familiar "objects", in an attempt to understand the machine better (Pressman, 2004).

As the perception of the system is influenced by the experiences of a person, each user creates his/her own conceptual model; since it is highly unlikely that people without special needs undergo similar experiences when surfing the Web as those with deficiencies, the models for disabled people tend to be distinct from the models for non-disabled people (Takagi, 2004). For example, according to Prof. Hercen, who was born blind, the window metaphor (Windows), which indicates the visualization of a work area, has no such meaning for a blind person (Hildebrandt, 2005).

When accessing a system, disabled users make use of a very different environment from non-disabled people. They relate the computer elements to "objects" from their day to day lives, developed to supply their needs. In addition, people with disabilities, such as blind persons, develop special skills, *e.g.*, excellent hearing; they hardly ever sit passively waiting to hear a spoken exit; they move around Web pages using complex combinations of keys. By means of this process, they create their own models and attempt to surf Web pages in a logical way. As these facts increase the level of difficulty when interacting with sites (Hanson, 2004), this ends up influencing their conceptual models (Takagi, 2004).

In systems geared to usability, the perception the user has of the system should be the closest possible to the system *per se*. This is why the designer should know the final users well enough to understand how they perceive the system, *i.e.*, their conceptual models. Thus when dealing with impaired users, it becomes essential to identify what types of impositions and limits they are subject to, in order to understand better their needs and special abilities (Takagi, 2004); an attempt should be made to understand all the hurdles users need to overcome to access information. If these hurdles are understood, it becomes possible to design easy-to-use interfaces for people with special needs as well (Harrison, 2005). The field work for the present study was conducted with this goal in mind.

VISUALLY IMPAIRED-MACHINE INTERACTION

The interface, graphic or otherwise, is the part of the software that users use to communicate with the system in order to perform their tasks; it should be designed to meet the users' expectations, allowing them to direct their attention to the objects they work with, which in turn reflect the real world (Pressman, 2004).

The interface should allow for user-friendly interaction; its design should be aimed at usability, the characteristic that determines whether the handling of a *product* is easy and quickly learned, difficult to forget, does not provoke operational errors, satisfies its users and efficiently resolves the tasks for which it was designed (Ferreira, 2003) and (Nielsen, 2002). If usability guides the system, users feel comfortable and encouraged to make use of it (Shneiderman, 2004). In order to build systems with sound usability, it is important that they be centered on the user (Norman, 1999); towards this end, one should get to know the final users, know how they perform their tasks, and the types of impositions and limits that they are subject to. Because graphic interfaces are a hurdle for visually challenged users, they must interact with the system using support technology capable of capturing interfaces and making them accessible.

USABILITY

Only recently the matter of usability has been perceived as important to information systems professionals. Driven by the market, organizations are going online in order to position themselves on a new way of performing business. Since the technology infrastructure used to construct Web sites can deal with images, sounds and text composition, it became more evident that the output of information should be treated with care (Ferreira, 2003).

Information system must be designed with the purpose of establishing a productive interaction between the system and their users in order to increase people's productivity while performing their tasks. They must satisfy the expectations and needs of their users. To achieve this end, the NFR (non functional requirement) usability must be present in any method for systems construction.

The communication between users and an Information System (IS) is established by means of the IS interface. A good IS design must guarantee a transparent communication, that is, it must assure that when a user accesses the IS to perform any task, he or she needs only to focus their energy on the intended task (Norman, 1986), (Norman, 1999) & (Jokela, 2004), (Seffah, 2004).

To have users focusing their attention mainly on their tasks, the process of software development must be "user centered", that is, its interface must be designed with the objective of satisfying the expectations and needs of users. The design of an interface that considers user characteristics and the NFR usability is a difficult process for many reasons, but most of this difficulty can be traced to the lack of attention on NFRs during the system definition process. Building systems that take in consideration NFRs, require the availability of a corpus of knowledge to help the engineer in the task of defining the system to comply with those requirements.

Usability Directives Focused on Accessibility

A usability oriented site is not necessarily accessibility oriented, and *vice versa*; a page may be easy to use for ordinary users, but inaccessible for those with special needs (Hanson, 2004). The directives recommend that site administrators check accessibility by accessing them through a screen reader; the problem is that those visually disabled, besides having special abilities, also use certain combinations of keys that a non-disabled person would not be able to simulate; hence, usability aspects differ from one user profile to that of another.

The authors of the current study agree that accessibility should be verified through a screen reader but that in order to obtain a universal access site geared to usability, it is essential that the difficulties and abilities of users be modeled as well, as these guide the mental model of their interactions. With such modeling, it would be possible to obtain usability directives that, in conjunction with the W3C accessibility directives, would provide for harmonious interaction of those disabled while guaranteeing understandable navigation sites.

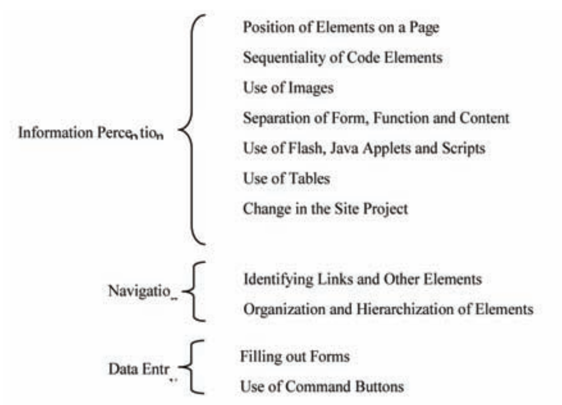
Usability issues occur, generally speaking, due to three reasons: the main aim of accessibility evaluation programs is compliance with directives where usability aspects are overlooked; many evaluation programs rely only on syntax verification techniques for sites and so detectable errors are limited to the tag description layer where the users' mental models are not taken into account (Takagi, 2004).

The literature (Ferreira, 2003), (Pressman, 2004) groups human-machine interaction sequences under two categories: *information visualization* and *data entry.* Based on the field work conducted at the *Instituto Benjamin Constant*, the present study proposes that, in interactions between visually impaired humans and machines, these sequences be grouped under three categories: *information perception*, *navigation* and *data entry.* After observing these interactions it was possible to identify certain usability directives, shown in the following sections (Figure 1).

Information Perception

Users accessing the Web by means of screen readers do not need to listen to all the words contained on the page: they only need to hear enough to determine where they want to go from there. It is, therefore, important that the interfaces be simple, because screen readers process the contents of a site differently from a visual reading process; to use them requires training and experience. Screen readers capture the HTML code, analyze the code and line up elements in the same order they appear in the code. As a result, visually impaired users perceive the page as if it were a text that they can

Figure 1. Usability directives



read line by line (by means of arrows) or link by link (by means of the tab key) (Nevile,2005).

Position of Elements on a Page

The order in which a screen element appears to the user accessing a page by voice reader is not the same order in which it appears visually on a navigator. What determines when it appears on the screen reader is the position it holds in the source code. An important element may very often have a prominent position on the page shown on a navigator, but when analyzed by the screen reader, ends up being one of the last elements users notice, as it is positioned at the end of the code. Aside from losing its prominent position, it will be perceived only after a number of less important information appears (Hildebrandt, 2005).

Sequentiality of Code Elements

Sequentiality is one of the barriers found by those visually disabled when browsing a site with a screen reader (or by means of a program that amplifies the interface): the user is only able to access a limited portion of the screen and, thus, loses out on the idea of the general context of the page at hand (Leporini, 2004). This is why HTML language tags that can be identified by the screen reader become an important element in the information perception process of visually impaired users. They allow the reader to provide information about the structure of the site (Leporini, 2004).

Hence, when designing an interface to be accessed by visually impaired users, one should be careful since many visual characteristics, such as bold, underline, italics, font styles, etc. are not detectable to the reader and so become imperceptible; on the other hand, other invisible elements, such as labels, link titles and alt attributes for images can be used to highlight information (Leporini, 2004).

Use of Images

The W3C 1.1 directive states that all non-textual information must be accompanied by a text. It is recommended that the alt attribute should always be present so that contents can be read by a screen reader. The alt attribute (alternate text) provides an alternative text associated to an image; in most navigators, the alt textual content is shown to the ordinary user when the mouse passes over the image. When the mouse moves away from the image, this window disappears. Likewise, when the page is being downloaded and an image is still being downloaded, the block with the alt text content appears until the final figure is shown.

If the texts for these image attributes are written appropriately, they can provide useful information for disabled users with respect to the meaning of the images being read by the screen readers (Queiroz, 2007), (Hanson, 2004) and (Harrison, 2005).

When a site contains an image without an alt attribute, this image may be detected or not; it will depend on the screen reader program being used. Some readers don't pick up anything, while others may indicate the existence of an "image" to the visually impaired user, but cannot furnish any information about the nature of the figure (Queiroz, 2007).

For instance, an interviewee reported not being able to register with an Internet provider as the link was hidden, probably because it was represented by an image without alt attribute. In her opinion the descriptive text of an image is only useful if it brings relevant information not available in the text (Livramento, 2005).

Interviewees say they prefer texts without images. What is more, they also state they cannot often find subtitles that might justify the presence of the image on the page. Perhaps bad use of alt attributes is in part responsible for the rejection of images, found to be the case in the current research work (Hildebrandt, 2005; Coube, 2005) (Livramento, 2005).

Separation of Form, Function and Content

Although a union of form, function and content are essential to obtain a complete and accessible site, the intersection between these elements may result in inconsistencies among different navigators or even among distinct means of access, as PDAs and cell phones. One should separate these elements, restricting the use of HTML to a description of the content and the use of CSS (Cascading Style Sheets) to the formatting (http_5).

As it offers many more resources for formatting and a more precise control for exhibiting each element in comparison with HTML formatting tags, the CSS standard plays an important role in Web accessibility. When using only CSS to format a page, the HTML code is restricted to the function of gathering and ranking the content, thus enabling navigators that do not depend on formatting – as is the case of screen readers for the visually impaired – to ignore the CSS code and concentrate only on what is contained in the HTML.

Use of Flash, Java Applets and Scripts

As HTML is not a programming language, in order to make sites more interesting, solutions were found to allow them to contain programs. Among these solutions, one can find scripts, small programs incorporated in Web pages with a capacity to generate special animated effects, formatting and forms.

As time passed, more powerful technology, like Flash and Java, began to be used to animate sites, and to make them interactive and more attractive. However, most flash and java applet files cannot be deciphered by screen readers. Though recent versions of Flash include resources that enable integration with accessibility support mechanisms, there are still limits to the relationship between flash components and text navigators, as is the case of certain screen readers. One solution would be to create a link for a text version without these resources (Queiroz, 2007).

Use of Tables

When a screen reader is used, tables are read horizontally, line by line. As visually impaired users cannot visualize the whole table and so have to rely on their memory to know in which position different columns are to be found, it would be preferable to re-read the heading of each column (the first cell of each column) before reading the data contained in each cell (Livramento, 2005). HTML offers resources that allow distinguishing the heading of the remaining cells, paving the way for this type of reading, as long as this distinction is correctly applied in the font code. It is, therefore, good usability practice to identify the names of each column and line by means of the “th” tag (table header).

Change in the Site Project

There are two types of alterations that can be made in sites. The first and more frequent one consists of simply updating the content without modifying the page layout. Newspapers, for example, are updated continually. This does not cause any uneasiness for visually impaired users.

The second and more problematic one occurs when the project of the site is altered. This obliges the visually impaired user to relearn the name and position of all the key elements on the page. Though this was not considered an obstacle, the visually impaired interviewees (Coube, 2005), (Hildebrandt, 2005) and (Livramento, 2005) wished to be notified every time a new version of the site came into effect. One interviewee reported that, when the page of her provider was modified, she and her husband, also blind, had no idea what was going on, unsure of whether it was an error on the part of the program or something they had done wrong (Livramento, 2005). One suggestion to cater to the needs of the visually impaired user is to put some identification inside the page containing the number and date of the current version.

Navigation

Visually disabled users do not use the mouse to navigate, since this device requires visual coordination (aim) (Queiroz, 2007). They mainly use the tab key and combinations of keys, called *shortcut keys*. These keys can also be used by non-disabled users to expedite certain tasks (Leporini, 2004). Using these keys requires learning one more skill, which leads partially sighted users to prefer using their residual sight (Hanson, 2004). This is why one should select a background color that will create a contrast between the background and the text to facilitate reading (Hanson, 2004).

Identifying Links and Other Elements

When navigating by means of keys, it is essential that the text describing the link be identified in an informative and useful way (Harrison, 2004); this text will be picked up by the screen reader and it is by this means that the disabled user will know what the link is for. So, simply identifying links with words like "click here" or "next" are an obstacle for users who rely on voice readers, as is the case of the visually impaired (http_1).

Organization and Hierarchization of Elements

Screen readers provide functions that enable users to jump directly to the various heading tags, a key element in structuring easy-to-navigate sites. By means of a tag, visually impaired users can navigate using the titles so as to get a general idea of the page (Takagi, 2004).

According to Livramento (2005), sites structured as paragraphs provide more objective navigation. Visually impaired users like to have the option of navigating by jumping from one paragraph to another, only reading through the ones they consider important; experienced users are quickly able to identify if they wish to continue reading a paragraph or skip to another, in this way approximating their method of reading to that of a person with ordinary eyesight (Livramento, 2005).

This is why it is essential to adopt the practice of signaling each paragraph in HTML code by means of a "p" tag instead of a "br" tag, which only enables a line break.

One of the problems in using a screen reader is that navigation on links is sequential (Leporini, 2004). This can slow down navigation. For instance, to return to a link to one's left, one would have to jump over all the links in order to restart reading the page and finally arrive at the desired content. Sites should provide resources that would enable users to jump links repeatedly, accelerating interaction. Hence skip links should be used, speeding up navigation and allowing users to jump links repeatedly and go directly to the desired content (Harrison, 2004). Skip links are not noticeable when a site is exhibited on an ordinary navigator, and are only useful when the site is being accessed by a screen reader (Takagi, 2004).

Data Entry

On entering data, visually impaired users do not use the mouse, but the keyboard, which has become a facilitator capable of being used by any visually disabled user due to an international typing norm: all keyboards produced in conformity with regional technical norms have, on the lower part of the J and F keys (on the alphanumerical side) and 5 (on the numerical side), high-relief to guide blind people when positioning their hands, just as people do when learning to type (Queiroz, 2007).

Relying solely on the keyboard, one could spend a long time choosing commands, typing data and inputting other things. Added measures should be taken to promote accessibility in interfaces dealing with data entry (Ferreira, 2003).

Filling Out Forms

If filling out forms can be a constraint for just any user, it is much worse for those visually impaired, forcing them at times to abandon the site. The simple fact that many sites have restricted access requiring passwords, which, due to security reasons are not spelled out by screen readers, already hinders user access.

One way people send data over the Internet is by filling out forms. Since the user navigates through forms by using the tab key, in order to facilitate data entry, the fields to be filled out and search buttons, if important, should preferentially be located at the top of the page. (Leporini, 2004).

Some interfaces are made in such a way that very often, when a visually impaired user locates the field to be filled out, no voice indication is made to explain what needs to be done; the user only hears a standard notification from the reader: “edit box.” The “label” tag would allow placing a text to be read by the user, giving information on what needs to be filled out (Queiroz, 2007). This tag also permits attributing a rapid access key to each field on the form, in addition to enlarging the click field for *selection box* and *radio box*, which would make filling out forms easier for those with only partial visual impairment. One should also avoid using a default value in the field, because even when read by the reader, it would require the user to erase the value (Harrison, 2004).

Another error found in forms is the indication of fields where one is required to make use of different color or font formatting. An alternative to this would be to use an asterisk, but screen reader users often disable the punctuation. Ideally this should be indicated by a letter that would represent the word “obligatory” (Harrison, 2004).

Use of Command Buttons

Another way of sending data is by means of command buttons, such as the “send” or “submit” button; these do not require a “label,” since they can be read by means of the “value” attribute; however, one should avoid using words such as “click here” or “continue” with this attribute, because they indicate nothing about the purpose of the button. If the button has an image instead of a text, it would suffice to use the “alt” attribute (Queiroz, 2007).

CONCLUSION

The current paper was motivated by the law decreed and signed in December 2004 that defined a twelve-month deadline for accessibility of public sites. Many organizations were prompted to try to adapt their information systems as a result of this decree in an attempt to obtain the accessibility certification.

Concern over accessibility at the moment of designing or redesigning a site does not guarantee this accessibility is maintained later when the site is updated. Constant and continual verification of accessibility should be made in order to avoid modifications in content or structure that would compromise the initial accessibility of the project. A new challenge emerges: that of designing, administering and maintaining sites in conformity with accessibility directives that are not only current and easy but also attractive.

Field research was conducted at the Instituto Benjamin Constant, the main reference center in Brazil for the education and re-education of those visually impaired, in order to get to know visually disabled users better. The reason for choosing visually disabled people was the fact that the Internet has done much to contribute to the improvement in the quality of their lives, allowing them to engage in new forms of relationships, find work opportunities and alternate forms of entertainment. This research work has prompted the understanding of how these users perceive and interact with sites and has identified certain hurdles that they need to overcome in order to access information. Through the understanding acquired in the field work and based on the literature, different types of impositions and limits that these users are subject to have been identified, enabling a better perception of their needs and special abilities. As a result, impaired user-machine interaction sequences have been grouped together into three categories: *information perception*, *navigation* and *data entry*. These interactions were observed and analyzed, which enabled the identification of certain usability directives that could contribute to the accessibility of sites in alignment with W3C directives, with emphasis on facilitating visually impaired user access to the Web.

REFERENCES

Enap. (2007, January). Material do curso "e-MAG-Modelo de Acessibilidade de Governo Eletrônico"- ministrado pela Escola Nacional de Administração Pública.

Ferreira, S.B.L., & Leite, J.C.S. (2003). Avaliação da usabilidade em sistemas de informação: o caso do sistema submarino - Revista de Administração Contemporânea - RAC. Publicação da ANPAD, v.7, n 2, - Abril/Junho.

Hanson, V. L. (2004). The user experience: Designs and adaptations. *ACM International Conference Proceeding Series - Proceedings of the International Cross-Disciplinary Workshop on Web Access*.

Harrison, S. M. (2005). Opening the eyes of those who can see to the world of those who can't: A case study. *Technical Symposium on Computer Science Education - Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education*.

Leporini, B., Andronico, P., & Buzzi, M. (2004). Designing search engine user interfaces for the visually impaired. *ACM International Conference Proceeding Series - Proceedings of the International Cross-disciplinary Workshop on Web Accessibility*.

Modelo de Acessibilidade. (2005). Recomendações de Acessibilidade para a Construção e Adaptação de Conteúdos do Governo Brasileiro na Internet - Departamento de Governo Eletrônico - Secretaria de Logística e Tecnologia da Informação - Ministério do Planejamento, Orçamento e Gestão - Documento de Referência - Versão 2.0 14/12/2005.

Nevile, L. (2005, November). Adaptability and accessibility: A new framework. *Proceedings of the 19th Conference of the Computer-Human Interaction Special Interest Group (CHISIG) of Australia on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future*, Canberra, Australia (Vol. 122, pp. 1-10).

Nicholl, A. R. J. (2001). O Ambiente que Promove a Inclusão: Conceitos de Acessibilidade e Usabilidade". Revista Assentamentos Humanos, Marília, v3, n. 2, p49–60.

Nielsen, J., & Tahir, M. (2002). Homepage: Usabilidade – 50 Websites desconstruídos. RJ: Editora Campus.

Norman, D. A. (1999). *The invisible computer: Why good products can fail, the personal computer is so complex, and information appliances are the solution*. MIT Press.

Petrie, H., Hamilton, F., King, N., & Pavan, P. (2006). Remote - usability evaluations with disabled people. *Proceedings of the SIGCHI Conference on Human Factors in Computing systems*, Canada.

Pressman, R. (2007). *Software engineering - a practioner's approach, 6th ed*. McGraw-Hill, Inc.

Queiroz, M. A. (2007). Material de um curso ministrado no SERPRO RJ por Marco Antonio de Queiroz, especialmente para cegos e pessoas de baixa visão. Tirado 02 de Abril 2003 de 2007.

Sales, M.B. De, & Cybis, W.de A. (2003). Desenvolvimento de um checklist para a avaliação de acessibilidade da *Web* para usuários idosos - *ACM International Conference Proceeding Series Proceedings of the Latin American Conference on Human-Computer Interaction.*

Shneiderman, B. (2004). *Designing the user interface: Strategies for effective human-computer interaction, 4th ed*. MA: Addison-Wesley.

Spelta, L.L. (2003). O Papel dos Leitores de Tela na Construção de *Sites* Acessíveis - Anais do ATIID 2003, São Paulo-SP, 23–24/09/2003.

Takagi, H., Asakawa, C., Fukuda, K., & Maeda, J. (2004). Accessibility designer: Visualizing usability for the blind. *ACM SIGACCESS Conference on Assistive Technologies - Proceedings of the ACM SIGACCESS Conference on Computers and Accessibility*.

Web Sites http4: http://www.*daSilva*.org.br/: February 5, 2006.

KEY TERMS AND DEFINITIONS

Accessibility: Term used to indicate the possibility of any person to make use of all the benefits of society.

Usability: Term used to denote the ease with which people can employ a particular tool or other human-made object in order to achieve a particular goal.

Visually Impaired: Vision loss that constitutes a significant limitation of visual capability resulting from disease, trauma, or a congenital or degenerative condition that cannot be corrected by conventional means, including refractive correction, medication, or surgery.

Instituto Benjamin Constant (IBC): Center of excellence and national reference in matters related to studies of visual impairment.

World Wide Web Consortium (W3C): The main international standards organization for the World Wide Web (abbreviated WWW or W3).

Web Accessibility Initiative (WAI): An effort to improve the accessibility of the World Wide Web (WWW or Web) for people using a wide range of user agent devices, not just standard Web browsers. This is especially important for people with physical disabilities which require such devices to access the Web.

INTERVIEWS

Coube, José Elias - Teacher in Informatics of the IBC (blind) – June 31, 2006.

Ferreira, Gerson F. - General Coordinator in Informatics of the IBC (blind) – June 10, 2006.

Hilderbrandt, Hercen - Teacher in Informatics of the IBC (blind) – June 29, 2006.

Cerqueira, Maria de Fátima Carvalhal - ex-learner of the IBC (blind) – July 9, 2006.

Souza, José Francisco - Director of Rehabilitation Division of the IBC (ten percent of vision) – July 22, 2006.

Livramento, Maria Luzia - Text Reviewer of Braille Press of the IBC (blind) – July 5, 2006.

This work was previously published in Multimodal Human Computer Interaction and Pervasive Services, edited by P. Grifoni, pp. 265-277, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 8.9
Web Application Server Clustering with Distributed Java Virtual Machine[1]

King Tin Lam
The University of Hong Kong, Hong Kong

Cho-Li Wang
The University of Hong Kong, Hong Kong

ABSTRACT

Web application servers, being today's enterprise application backbone, have warranted a wealth of J2EE-based clustering technologies. Most of them however need complex configurations and excessive programming effort to retrofit applications for cluster-aware execution. This chapter proposes a clustering approach based on distributed Java virtual machine (DJVM). A DJVM is a collection of extended JVMs that enables parallel execution of a multithreaded Java application over a cluster. A DJVM achieves transparent clustering and resource virtualization, extolling the virtue of single-system-image (SSI). The authors evaluate this approach through porting Apache Tomcat to our JESSICA2 DJVM and identify scalability issues arising from fine-grain object sharing coupled with intensive synchronizations among distributed threads. By leveraging relaxed cache coherence protocols, we are able to conquer the scalability barriers and harness the power of our DJVM's global object space design to significantly outstrip existing clustering techniques for cache-centric web applications.

DOI: 10.4018/978-1-60566-661-7.ch028

INTRODUCTION

Scaling applications in web server environment is a fundamental requisite for continued growth of e-business, and is also a pressing challenge to most web architects when designing large-scale enterprise systems. Following the success of the Java 2 Platform, Enterprise Edition (J2EE), the J2EE world has developed an alphabet soup of APIs (JNDI, JMS, EJB, etc) that programmers would need to slurp down if they are to cluster their web applications. However, comprehending the bunch of these APIs and the clustering technologies shipped with J2EE server products is practically daunting for even those experienced programmers. Besides the extra configuration and setup time, intrusive application rework is usually required for the web applications to behave correctly in the cluster environment. Therefore, there is still much room for researchers to contribute improved clustering solutions for web applications.

In this chapter, we introduce a generic and easy-to-use web application server clustering approach coming out from the latest research in distributed Java virtual machines. A *Distributed Java Virtual Machine (DJVM)* fulfills the functions of a standard JVM in a distributed environment, such as clusters. It consists of a set of JVM instances spanning multiple cluster nodes that work cooperatively to support parallel execution of a multithreaded Java application. The Java threads created within one program can be distributed to different nodes and perform concurrently to exploit higher execution parallelism. The DJVM abstracts away the low-level clustering decisions and hides the physical boundaries across the cluster nodes from the application layer. All available resources in the distributed environment, such as memory, I/O and network bandwidth can be shared among distributed threads for solving more challenging problems. The design of DJVM adheres to the standard JVM specification, so ideally all applications that follow the original Java multithreaded programming model on a single machine can now be clustered across multiple servers in a virtually effortless manner.

In the past, various efforts have been conducted in extending JVM to support transparent and parallel execution of multithreaded Java programs on a cluster of computers. Among them, Hyperion (Antoniu et al., 2001) and Jackal (Veldema et al., 2001) compile multithreaded Java programs directly into distributed applications in native code, while Java/DSM (Yu & Cox, 1997), cJVM (Aridor, Factor, & Teperman, 1999), and JESSICA (Ma, Wang, & Lau, 2000) modify the underlying JVM kernel to support cluster-wide thread execution. These DJVM prototypes debut as proven parallel execution engines for high-performance scientific computing over the last few years. Nevertheless, their leverage to clustering real-life applications with commercial server workloads has not been well-studied.

We strive to bridge this gap by presenting our experience in porting the Apache Tomcat web application server on a DJVM called JESSICA2. A wide spectrum of web application benchmarks modeling stock quotes, online bookstore and SOAP-based B2B e-commerce are used to evaluate the clustering approach using DJVMs. We observe that the highly-threaded execution of Tomcat involves enormous fine-grain object accesses to Java collection classes such as hash tables all over the request handling cycles. This presents the key hurdles to scalability when the thread-safe object read/write operations and the associated synchronizations are performed in a cluster environment. To overcome this issue, we employ a home-based hybrid cache coherence protocol to support object sharing among the distributed threads. For cache-centric applications that cache hot and heavyweight web objects at the application-level, we find that by using JESSICA2, addition of nodes can grow application cache hits linearly, significantly outperforming the share-nothing approach using web server load balancing plug-in. This is attributed to our *global object space (GOS)* architecture that virtualizes network-wide memory resources for caching the application data as a unified dataset for global access by all threads. Clustering HTTP sessions over the GOS enables effortless cluster-wide session management and leads to a more balanced load distribution across servers than the traditional sticky-session request scheduling. Our coherence protocol also scales better than the session replication protocols adopted in existing Tomcat clustering. Hence, most of the benchmarked web applications show better or equivalent performance compared with the traditional clustering techniques.

Overall, the DJVM approach emerges as a more holistic, cost-effective and transparent clustering technology that disappears from the application programmer's point of view. With efficient protocol support for shared object access, such a middleware-level clustering solution is suitable for scaling most web applications in a cluster environment. Maturing of the DJVM technology would bring about stronger server resource integration and open up new vistas of clustering advances among the web community.

The rest of the chapter is organized as follows. In Section 2, we survey the existing web application clustering technologies. Section 3 presents the system architecture of our JESSICA2 DJVM. In Section 4, we describe Tomcat execution on top of the JESSICA2 DJVM. Section 5 discusses JESSICA2's global object space design and implementation. In Section 6, we evaluate the performance of Tomcat clustering using the DJVM. Section 7 reviews the related work. Section 8 concludes this chapter and suggests some possible future work.

EXISTING WEB APPLICATION CLUSTERING TECHNOLOGIES

In the web community, clustering is broadly viewed as server load balancing and failover. Here, we discuss several widely adopted clustering technologies under the hood of J2EE.

The most common and cost-effective way for load balancing is to employ a frontend web server with load balancing plug-ins such as Apache mod_jk (ASF, 2002) to dispatch incoming requests to different application servers. The plug-ins usually support *sticky-sessions* to maintain a user session entirely on one server. This solution could make the cluster resource utilization more restricted and is not robust against server failures.

More advanced solutions need to support application state sharing among servers. Large-scale J2EE server products generally ship with clustering support for HTTP sessions and stateful session beans. One traditional approach is to serialize the session contents and persist the states to a data store like a relational database or a shared file system. However, this approach is not scalable. *In-memory session replication* is an improved technique also based on Java serialization to marshal session-bound objects into byte streams for sending to peer servers by means of some group communication services such as JGroups (Ban, 1997) (based on point-to-point RMI or IP multicast). Such a technique has been implemented in common web containers such as Tomcat. However, scalability issues are still present in group-based synchronous replications, especially over the general all-to-all replication protocols which are only efficient in very small-size clusters.

Enterprise JavaBeans (EJB) is a server-side component architecture for building modular enterprise applications. Yet the EJB technology itself and its clustering are both complicated. Load balancing among EJB containers can be achieved by distributed method call, messaging or name services which correspond to the three specifications: Remote Method Invocation (RMI), Java Messaging Service (JMS) and Java Naming and Directory Interface (JNDI). In particular, JNDI is an indispensible element of EJB clustering as EJB access normally starts with a lookup of its home interface in the JNDI tree. For clients to look up clustered objects, EJB containers implement some global JNDI services (e.g. cluster-wide shared JNDI tree) and ship with special RMI compilers to generate *replica-aware stubs* for making user-defined EJBs "cluster-aware". The stub contains the list of accessible target EJB instances and codes for load balancing and failover among the instances. EJB state changes are serialized and replicated to peer servers after the related transaction commits or after each method invocation. Undoubtedly, this clustering technology is expensive, complicated and with application design restrictions.

In recent years, a growing trend in web application development has begun to adopt lightweight containers such as the Spring Framework (Johnson, 2002) to be the infrastructural backbone instead of the EJB technology. On such a paradigm, business objects are just *plain old Java objects* (POJOs) implementing data access logic and running in web containers like Tomcat. Caching POJOs in a collection object like Hashtable is also a common practice for saving long-latency access to database and file systems. To support clustering of POJOs which conform to no standard interface, it seems almost inevitable that application programmers have to rework their application code to use extra APIs to synchronize object replicas among the JVMs. Though distributed caching libraries (Perez, 2003) can facilitate POJO clustering, these solutions again rely on Java serializations and require complex configurations. The cache sizes they support are usually bounded by single-node memory capacity as a result of employing simplistic all-to-all synchronization and full replication protocols.

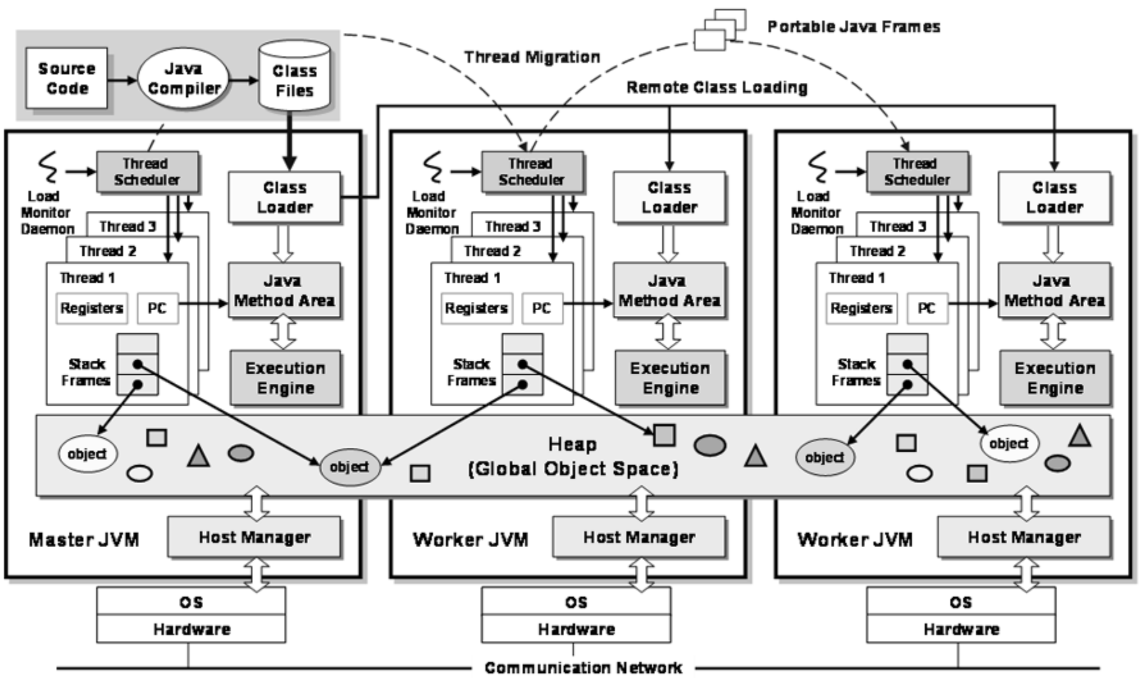
Although the clustering solutions surveyed so far have their own merit points, most of them share several significant shortcomings.

- **Restrictions on application design:** Many object sharing mechanisms rely on Java serializations which pose restrictions on application design and implementation. They cannot easily work in a cluster environment.
- **Possible loss of referential integrit**y: Most solutions suffer the break of referential integrity since it creates clones of the replicated object graph at deserialization and may lose the original object identity. That's why when a shared object undergoes changes, it must be put back into the container object by an explicit call like setAttribute() to reflect the new referential relation. Likewise, consistency problems occur when attributes with cross-references in HTTPSession are modified and unmarshaled separately.
- **Costly communication:** Object serialization is known to be hugely costly in performance. It performs a coarse trace and clones a lot of objects even for one field change. So there is certain limit on the number and sizes of objects that can be bound in a session.
- **No global signaling/coordination support:** Subtle consistency problems arise when some design patterns and services are migrated to clusters. For example, the *singleton pattern* sharing a single static instance among threads as well as some synchronization codes become localized to each server, losing global coordination. Event-based services like timers make no sense if they are not executed on a single platform. Only a few products (e.g. JBoss's clustered singleton facility) ship with configurable cluster-wide coordination support to ease these situations.
- **Lacking global resource sharing:** Most clustering solutions in the web domain put little focus on global integration of resources. They cannot provide a global view of the cluster resources such as memory, so each standalone server just does its own work without cooperation and may not fully exploit resources.

JESSICA2 DISTRIBUTED JVM

JESSICA2 (Zhu, Wang, & Lau, 2002) is a DJVM designed to support transparent parallel execution of multithreaded Java applications in a networked cluster environment. It was developed based on

Figure 1. JESSICA2 DJVM System Architecture



Kaffe JVM (Wilkinson, 1998). The acronym JESSICA2 spells as Java-Enabled Single-System-Image Computing Architecture version 2; this architecture promotes the *single-system image (SSI)* notion when connecting Java with clusters. Such a design concept is helpful to take away the burden of clustering by hand from application developers. The key advantage with using JESSICA2 is its provision of transparent clustering services which require no source code modification and bytecode preprocessing. It will automatically take care of thread distribution, data consistency of the shared objects and I/O redirection so that the program will run under an SSI illusion with integrated computing power, memory and I/O capacity of the cluster.

Figure 1 shows the system architecture of the JESSICA2 DJVM. JESSICA2 has bundled a number of salient features extended from the standard JVM that realize the SSI services. To execute a Java application on JESSICA2, a tailored command is called to start the master JVM on the local host and the worker JVMs on remote nodes, based on the specified list of hostnames. In each JVM, a class loader is responsible for importing bytecode data (of both the basic Java class library classes and the application classes) into its method area where a Java thread can look up a specific method to invoke. The class loader of JESSICA2 is extended to support *remote class loading* which ensures when a worker JVM cannot find a class file locally, it can request the class bytecode on demand and fetch the initialized static data from the master JVM through network communication. This feature greatly simplifies cluster-wide deployment of Java applications and hence transparently provides the *web farming* support which traditionally requires application server extension to fulfill.

When the Java threads of the application are started, the thread scheduler of the JVM will put their contexts (e.g. program counter and other register values) into the execution engine in turns. The Java methods invoked by the running thread will be compiled by the Just-In-Time (JIT) compiler into native codes for high-speed execution. JESSICA2 incorporates a cluster-aware JIT compiler to support lightweight *Java thread migration* across node boundaries to assist global thread scheduling. Java threads will be assigned to each worker JVM at the startup time in a round-robin manner to strike a raw load balance. Dynamic load balancing during runtime can be done by migrating Java threads that are running into computation hotspots to the less loaded nodes. For detecting hotspots, each JVM instance has a *load monitor* daemon that periodically wakes up and sends current load status such as CPU and memory utilization to the master JVM which is then able to make thread migration decisions with a global resource view.

Java threads migrated to remote JVMs may still be carrying references to the objects under the source JVM heaps. For seamless object visibility, JESSICA2 employs a special heap-level service called the *Global Object Space (GOS)* to support location-transparent object access. Objects can be shared among distributed threads over the GOS as if they were under a single JVM heap. For this to happen, the GOS implements object packing functions to transform object graphs into byte streams for shipping to the requesting nodes. The shipped object data will be saved as a *cache copy* under the local heap of the requesting node. Caching improves data access locality but leads to cache consistency issues. To tackle the problem of stale data, the GOS employs *release-consistent* memory models stemmed from software Distributed Shared Memory (DSM) systems to preserve correct memory views on shared objects across reads/writes done by distributed threads.

JESSICA2 offers parallel I/O and location-transparent file access. We extend JESSICA2 to support *transparent I/O redirection* mechanism so that I/O requests (file and socket access) can be virtually served at any node. Our system does not rely on shared distributed file systems such as NFS, nor does it need to restrict a single IP address for all the nodes in the running cluster. Rather, we extend each JVM to run a *transparent I/O redirection* mechanism to redirect non-home I/O operations on files or sockets to their home nodes. To attain I/O parallelism atop transparency, read-only file operations and connectionless network I/O can be done at the local nodes concurrently without redirection.

Finally, all inter-node communication activities required by the subsystems at upper layers like the GOS and I/O redirections are supported by a common module called the *host manager* which wraps up the underlying TCP communication functions with connection caching and message compression optimizations.

On the whole, we can see that DJVM is a rather generic middleware system that supports parallel execution of any Java program. Since the unveiling of DJVMs, their application domains remain mostly in scientific computing over the last few years. They were used to support multithreaded Java programs that are programmed in a data-parallel manner. These applications tend to be simple, embarrassingly parallel so that DJVMs could offer good scalability. However, much more mainstream applications are business-oriented, centered at server-side platforms and run atop some Java application servers. Their object access and synchronization patterns are far more complex. In the next sections, we will elaborate on the common runtime characteristics of application servers and their impacts on the DJVM performance through a case study of Apache Tomcat running on JESSICA2.

APACHE TOMCAT ON DISTRIBUTED JVM

Apache Tomcat is a Java servlet container developed at the Apache Software Foundation (ASF). It serves as the official reference implementation of the Java Servlet and JavaServer Page (JSP) specifications. Tomcat is the world's most widely used open-source servlet engine and has been used by renowned companies like WalMart, E*Trade Securities and The Weather Channel to power their large-scale and mission-critical web applications in production systems.

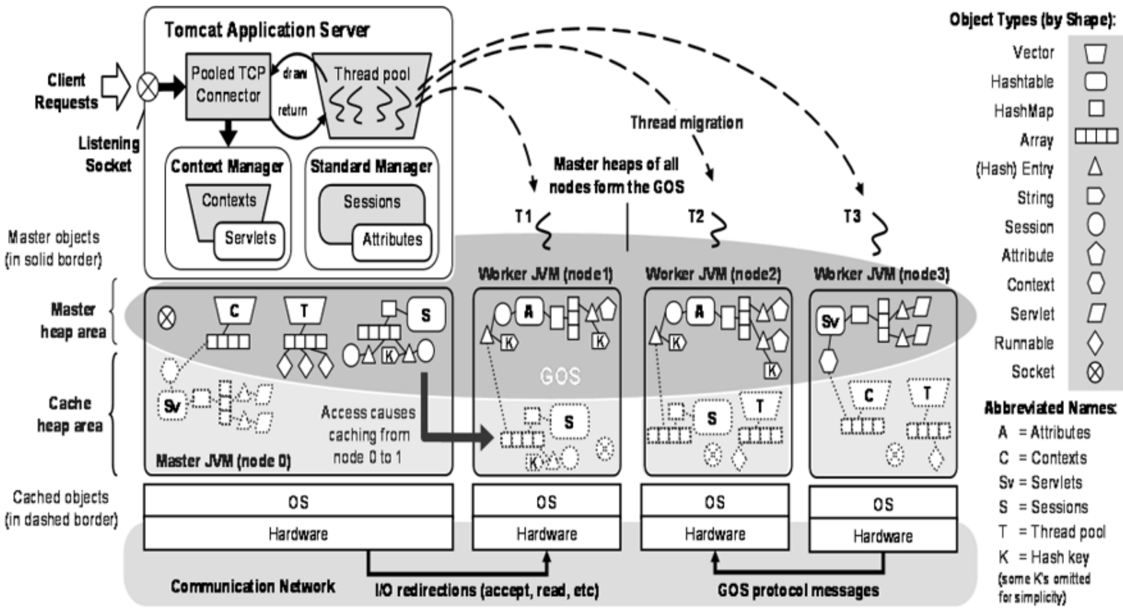
As a common design in many servers, Tomcat maintains a thread pool to avoid thread creation cost for every short-lived request as well as to give an upper bound to the overall system resource usage. Upon an incoming connection, a thread is scheduled from the pool to handle it. The web container then performs various processing such as HTTP header parsing, sessions handling, web context mapping and servlet class loading. The request eventually reaches the servlet code which implements application logics such as form data processing, database querying, HTML/XML page generation, etc. Finally, the response is sent back to the client. This request service cycle is complex, comes across many objects throughout the container hierarchy and imposes multithreading challenges to the DJVM runtime.

Being a classical and large-scale web application server, Tomcat reflects an important class of real-life object-oriented server execution patterns that are summarized as follows.

1. **I/O-intensive workload**: Most web server workloads are I/O-bound and composed of short-lived request processing. The per-request computation-communication ratio is usually small.
2. **Highly-threaded**: It is common that a server instance is configured with a large number of threads, typically a few tens to a hundred per server to hide I/O blocking latency.
3. **High read/write ratios**: Shaped by customer buying behaviors and e-business patterns, web applications usually consist of high read/write ratio, say around 90/10; the dominant reads come from browsing while only a few writes owing to ordering happen over a period.
4. **Long-running**: Typically a server application runs for an indefinitely long time, processing requests received from the client side.
5. **High utilization of collection framework**: Tomcat makes extensive use of Java collection classes like Hashtable and Vector to store information (e.g. web contexts, sessions, attributes, MIME types, status codes, etc). They are accessed frequently when checking, mapping and searching operations happen inside the container. To reduce object creation and garbage collection costs, many application servers apply the *object pooling* technique and use collection classes to implement the object pools.
6. **Fine-grain object access**: Fine-grain object access has two implications here: (1) the object size is small; (2) the interval between object accesses to the heap is short. Unlike many scientific applications which have well-structured objects with size of at least several hundred bytes, Tomcat contains an abundance of small-size objects (average about 80 bytes by our experience) throughout the container hierarchy. Object accesses are very frequent due to object-oriented design of Tomcat.
7. **Complex object graph with irregular reference locality**: Some design patterns such as facade and chain of interceptors used in Tomcat yield ramified object connectivity, cross-referencing and irregular reference locality among objects throughout the container hierarchy. By property 5, heavy use of Java Hashtable or HashMap also intensifies the irregularity of reference locality as hash entries are accessed in a shuffling pattern, contrasting with the consecutive memory access pattern in array-based scientific computations.

Figure 2 depicts the execution of the Tomcat application server on top of a 4-node cluster. When

Figure 2. Execution of Tomcat on JESSICA2 DJVM



Tomcat is executed atop JESSICA2, Tomcat is exposed to an SSI view of the integrated resources of the cluster nodes as if it was in one powerful server. A customized Tomcat startup script is used to bring up the server, running atop the master JVM. The script is tailored to supply the DJVM runtime parameters (e.g. the port number for master-worker handshaking) and to read a host configuration file which defines the hostnames or IP addresses of the worker nodes the DJVM would span.

When the server spawns a pool of threads, the threads will be migrated to the worker nodes. They will load the classes of the Java library, Tomcat and the web applications deployed dynamically through the cluster-aware class loader of JESSICA2. In this way, "virtual" web application server instances are set up on the worker nodes. The virtual server instances pull workload continuously from the master node by accepting and handling incoming connections through transparent I/O redirections. On each worker node, I/O operations (accept, read/write and close) performed on the shared server socket object (wrapped in the pooled TCP connector) will be redirected to the master node where it was bound to the outside world. Most other I/O operations can be performed on I/O objects created locally; so each cluster node can serve web page requests and database queries in parallel.

When a client request is accepted, the context manager of Tomcat will match it to the target web application context. If the request carries session state such as a cookie, the standard manager will search for the allocated session object from the sessions hash table. In essence, all Tomcat container objects including the context manager, the standard manager, the sessions hash table and web contexts allocated in the master JVM heap are transparently shared among the distributed threads by means of the underlying GOS service mentioned in section 3. When a thread gets the first access to a non-local object reference, it will encounter an *access fault* and send a fetching request to the object's home node. The home node will respond with the up-to-date object data and export the local object as the home copy of the shared object. Cluster-wide data consistency will be enforced on the home copy and all cache copies derived from it thereafter. Since each thread will be able to see the shared object updates made by others through synchronization, the global shared heap creates an *implicit cooperative caching effect* among the threads. The power of this effect can be exemplified by collection classes like hash tables.

As illustrated, all HTTP sessions stored in a Tomcat-managed hash table can be globally accessible. The responsibility of maintaining HTTP session data consistency across servers has transparently shifted to the GOS layer. In other words, every server is eligible to handle requests belonged to any client session. This leads to more freedom of choice in request scheduling policies over sticky-sessions load balancing which can run into hotspots. Another useful scenario is using the GOS to augment the effective cache size of an application-level in-memory Java cache (e.g., a hash table for looking up database query results). The fact that every thread sees the cache entries created by one another contributes to secondary (indirect) application cache hits through remote object access. The cache size can now scale linearly with additional nodes, so we can greatly take the load off the bottlenecked database tier by caching more data at the application tier.

The DJVM approach inherits most advantages of clusters. However, the aforesaid server runtime properties bring additional design challenges on the DJVM runtime. First, I/O intensive workloads are known to be more difficult to scale efficiently over a cluster. Second, the high thread count property implies higher blocking latency if contention occurs. More memory overheads would be resulted from any per-thread protocol data structures. High read/write ratio is a positive news to the GOS as it implies shared writes are limited, so our protocols can take this property as a design tradeoff. Next, for long-running applications, we need to make sure the memory overhead induced by the coherence protocol data structures scales up slowly for less frequent garbage-collection cycles. Property 5 puts up the biggest barrier to scalability. Frequent synchronizations on the globally shared thread pool and object pools produce intensive remote locking overhead. Worse still, these pools are usually built from Java collection classes which are not scalable. For example, fine-grain accesses to hash entries of a Java hash table are all bottlenecked around the single map-wide lock contention which will be much intensified by distributed locking. Finally, properties 6 and 7 together issue enormous remote access roundtrips and demand smart object prefetching techniques for aggregating fine-grain communications. These observations call for a renovation of JESSICA2's global object space (GOS) architecture.

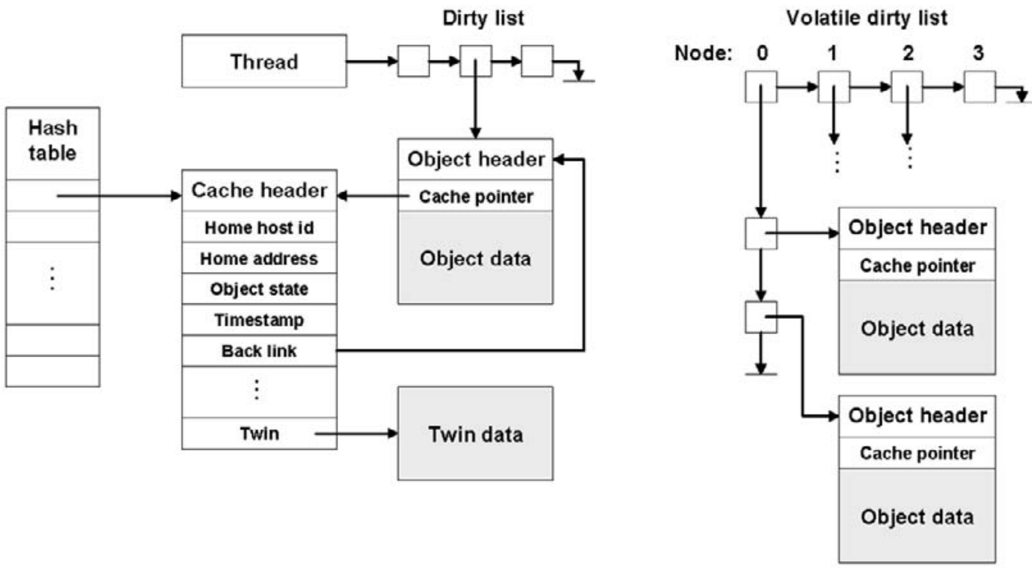
GLOBAL OBJECT SPACE

In this section, we elaborate on the design and implementation of our enhanced GOS system. We discuss the structure of the extended JVM heap, a home-based cache coherence protocol tailored for managing locks and a cluster-wide sequential consistency protocol for handling volatile field updates.

5.1 Overview of the Extended JVM Heap

To support cluster-wide object modification propagation and consistency maintenance, the heap of the standard JVM should be extended to make it "cluster-aware". In JESSICA2, each JVM heap is logically divided into two areas: the *master heap area* and the *cache heap area*. The master heap area essentially rides on the unmodified JVM heap, storing ordinary local objects. To make it "cluster-aware", when the local objects are being shared with some remote threads, they are exported as home objects with special flags marked in their object headers. The cache heap

Figure 3. GOS internal data structures



area manages cache objects brought from the master heap of a peer node. It consists of extra data structures for maintaining cluster-wide data consistency. The original GOS follows an intuitive design in which each thread has its own cache heap area, resembling the thread-private working memory based on the *Java memory model (JMM)*. This design prevents local threads from interfering each other's cache copies (such as during invalidations) but wastes precious memory space to keep redundant per-thread cache copies on the same node. So we adopt a unified cache design in the enhanced GOS which allows all local threads running on a single node to share a common cache copy. This design not only makes better usage of available memory resources but also reduce remote object fetching since when a thread faults in an object, other peer threads at the same node requesting the same object could find it in place. We also switch to a release consistency memory model in which the dominant read-only objects are never invalidated, so the interference among local threads is practically small. These modifications potentially could accommodate a high server thread count and achieve better memory utilization.

Figure 3 shows the internal data structures of the extended JVM heap. The object header of every object is augmented with special fields such as the *cache pointer*. A local or home object has a null cache pointer whereas a cache object has its cache pointer points to an internal data structure called *cache header* that contains the state and home information of the object. A node-level hash table (shared by all local threads) is used to manage and to look up cache headers during fetching events. In order to tell the home nodes of the modifications made on cache objects, each thread maintains a *dirty list* that records the ids of cache objects it has modified. At synchronization points, updates made on the dirty objects are flushed back to their home nodes. A similar per-node *volatile dirty list* is used to record updates on objects with volatile fields which are maintained by a separate single-writer protocol to be explained in section 5.3.

Object state is composed of two bits: valid/invalid and clean/dirty. The JIT compiler is tweaked to perform *inline checking* on each cache object access to see if its state is valid for read/write. Read/write on an invalid object will trigger appropriate interface functions to *fault-in* the up-to-date copy from its home. For efficiency, the software check is injected as a tiny assembly code fragment to the relevant bytecode instructions (GETFIELD, PUTFIELD, AALOAD, AASTORE, etc), testing the last two bits of the cache pointer. Valid object access passing the check will not impose any GOS interface function call overhead and is thus as fast as local object access.

Creating a single-heap illusion to distributed threads entails an advanced design of distributed cache coherence protocol as it has to be compliant to the Java memory model that defines the memory consistency semantics across multiple threads. The Java language provides two synchronization constructs for the programmers to render thread-safe code – the synchronized and volatile keywords. The synchronized keyword guarantees a code fragment or method with atomicity and memory visibility while volatile ensures that threads can see the latest values of volatile variables. We will discuss our enhancements of the GOS for handling the two types of synchronizations in Section 5.2 and 5.3 respectively.

5.2 Home-based Lazy Release Consistency Protocol

Entering and exiting a synchronized block or method correspond to acquiring and releasing the lock associated with the synchronized object. To fulfill the Java memory model, the original GOS implements an intuitive solution that works as follows. Upon a lock release, all updates to cache objects are flushed to their home nodes. Upon a lock acquire, all cache objects are invalidated, so later accesses will fault in the up-to-date copies from the home nodes. However, this would incur significant object fault-in overheads after every lock acquire. Thus, we renovate the original global object space by adopting a more relaxed *home-based lazy release consistency (HLRC)* memory model.

Contrary to the intuitive solution, upon a lock acquire, we confine invalidations to cache copies of shared objects that have been modified by other nodes only, rather than invalidating the total cache heap area. Our home-based cache coherence protocol guarantees memory visibility based on *Lazy Release Consistency (LRC)* (Keleher, Cox, & Zwaenepoel, 1992). LRC delays the propagation of modifications to a node until it performs a lock acquire. Lock acquire and release delimit the start and end of an *interval*. Specifically, LRC insures that the node can see the memory changes performed in other nodes' intervals according to the *happened-before-1* partial order (Adve & Hill, 1993), which is basically given by the local node's locking order and the shared lock transfer event. This means all memory updates preceding the release performed by a node should be made visible to the node that acquires the same lock. HLRC is similar to LRC in the sense of lock management but shapes the modification propagation into home-based patterns.

Memory updates are communicated based on a *multiple-writer* protocol implemented using the *twin-and-diff* technique that allows two or more threads to modify different parts (i.e. different fields or array portions) of the same shared object concurrently without conflict. In this technique, a *twin* copy is made as a data snapshot before the first write to a cache object in the current interval. Upon a shared lock release, for each dirty cache object, the modified part, i.e. *diff*, is differentiated from the twin. The diff is eagerly flushed to the corresponding home node, keeping the home copy always up-to-date. The thread can then safely discard the twins and diffs and close the interval. When the lock is acquired by another thread, the releaser passes *write notices* along the lock grant to the acquirer. The acquirer uses the write notices to invalidate the corresponding cached objects. It also saves the write notices so that they can be passed on to the next acquirer enforcing the happens-before partial order. A later access on an invalidated cache object will fault in the up-to-date copy from its home.

Here, we have to deal with some tricky data-race problems arising from sharing a unified cache copy among local threads. First, for systems of object-based granularity as in our case, field-level false sharing may occur since protecting different fields of one object by different locks is reckoned as well-synchronized in Java. For example, while one thread T1 holds a lock for modifying field A of a cache copy and makes it becomes in dirty state, another local thread T2 may acquire a lock for modifying field B of the same object. If another node has modified field B using the same lock, then T2 will invalidate that cache copy and fault-in the home copy, overwriting those pending modifications made by T1. Second, in systems with object prefetching, it is possible for one thread faulting in a home object A with object B prefetched to overwrite the pending modifications on the shared cache copy B made by another thread. Currently, we deal with these hazards by reconciling the timestamp field associated to each object to resolve detectable version conflicts and by incorporating techniques similar to *two-way diffing* (Stets et al., 1997).

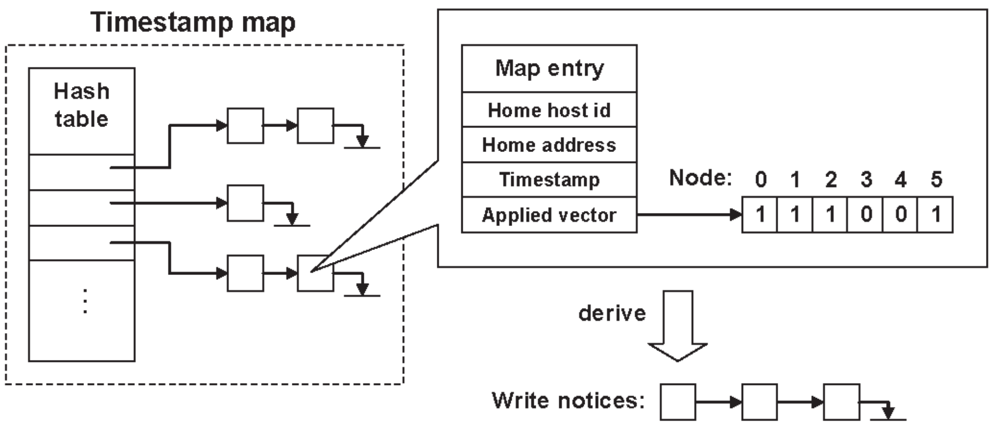
For home objects, local read/write can be done directly without generating and applying diffs. This benefit is usually known as the *home effect* (Zhou, Iftode, & Li, 1996). Some minor overhead that home nodes still need to pay is to keep record of the local writes for the next remote acquiring thread to invalidate the relevant cache copies. Locking of home objects resembles locking of local objects if the lock ownership has not been given to any remote nodes. Otherwise, it has to wait for the lock release done by the last remote acquirer.

Compared with homeless protocols, the advantages of HLRC are: 1. the home effect for reducing high diffing overheads; 2. fewer messages since an object fault can always be satisfied by one round-trip to home instead of diff request messages to multiple writers; 3. no *diff accumulation* and so no need for garbage collection of diffs. Hence, this becomes our protocol design choice for shorter latency seen by I/O-bound workload and less garbage accruing from long-running server applications.

Nevertheless, we depart from the usual HLRC implementations in some aspects. To track and enforce the happens-before partial order, traditional HLRC implementations rely heavily on *vector timestamps* to dig out the exact minimal intervals (and write notices) that the acquirer must apply. While this ensures the most relaxed invalidation, this entails complex data structures like interval records in Treadmarks (Keleher et al., 1994) or *bins* database (Iosevich & Schuster, 2005) to keep the stacks of vectors. The storage size occupied by them scales with the number of locking on shared objects. For lock-intensive applications, these stacks can grow up quickly and consume enormous space. For long-running server applications, the problem becomes more critical and systems that rely on pre-allocation schemes such as cyclic bins buffers (Iosevich & Schuster, 2005) will ultimately run out of space and result in runtime failure. Discarding interval records is possible if they have already been applied to all nodes. But some nodes may never acquire a particular lock while some nodes intensively acquire it. This issue is not ignorable particularly in multithreaded protocols where the length of vector timestamp scales with the number of threads. For highly-threaded applications like Tomcat, this has scalability impacts on both memory and network bandwidth. Therefore, our protocol eschews the use of vector timestamps.

Rather we employ object-level scalar timestamps to assist deriving the set of write notices. The basic idea is illustrated by Figure 4. Each node maintains a data structure called *timestamp map* which is essentially a hash table recording all shared objects that have once been modified. Each map entry consists of the object id, a scalar timestamp and a n-bit binary vector (n being number of nodes) and is used to derive the cor-

Figure 4. Timestamp map for implementing HLRC



responding write notice formatted as a couple (object id, timestamp). The *n*-bit binary vector is used to keep track of which node has applied the write notice (0 = not yet; 1 = applied). If all the *n* nodes have applied the write notice, it is considered *obsolete* and can be discarded. The size of this map scales with the number of modified shared objects rather than the number of shared locking. Repetitive locking on the same object will not generate separate interval records but go to update the same entry in the timestamp map. Due to high read/write ratios in web applications, the number of modified shared objects is limited. The timestamp map will also undergo a *periodic shrinking phase* to clean up those obsolete entries. So the map is practically small in size at most of the time.

Upon a shared lock release, modifications will be recorded into the local node's timestamp map. When a lock transfer happens, all non-obsolete map entries will be extracted as write notices and passed from the releaser to the acquirer. They will also be saved in the acquirer's map. Write notices with a newer timestamp will overwrite an old map entry if any and reset its *n*-bit vector to all zeros so that future acquirers will be able to know the changes. Without tracking the exact partial order, the set of write notices sent to an acquirer may not be minimal and possibly include modifications that "happens-after" the release of the lock being acquired. The drawback is that some cache objects at the acquirer side may be invalidated earlier than necessary. However this effect is insignificant since if the thread is really going to access the invalidated cache objects, it eventually needs to see the modifications. This effect will not accrue owing to our periodic cleanup of obsolete map entries and selective invalidations based on object timestamp comparison.

5.3 Volatile Consistency Protocol

Most DJVM prototype implementations enforce cluster-wide semantics of the volatile construct in a way that is stricter than necessary. For straightforward compatibility, the volatile construct is usually treated as if it was a lock, thus introducing unnecessary mutual exclusivity to the application. The latest Java concurrent utility package (JSR166, 2004), particularly the ConcurrentHashMap class shipped along, employs segment-based locks plus volatile count and object value fields to guard different hash bucket ranges. The advanced data structure offers much more scalable throughput than the conventional Java Hashtable. However, such a good design for concurrency will be smothered if the underlying DJVM handle the volatile fields as locks. So we decided to tailor consistency support to volatile fields.

Our new protocol for maintaining cluster-wide volatile field consistency is a *passive-based concurrent-read exclusive-write (CREW)* protocol. It enforces sequential consistency to ensure the next reader thread can see the updates made by the last writer on the same object.

To implement this model, we need to assign a manager for each object with volatile fields and it is naturally the home node where the object is created. For ease of explanation, we call an object with a volatile field as *volatile object*. The home node needs to maintain two states on the home copy of volatile object: *readable* and *exclusive*, as well as a list called *copyset* of the nodes that currently have a valid cache copy of this object. When the home node receives a fetch request from a node on a readable volatile object, the node's id will be added to the copyset list of the home copy. The consistency of a volatile object relies on the active writer to tell the readers of such an update. When a thread wants to write the object, no matter the home or cache copy, it must first gain the exclusive right on it from its home node. Before the exclusive right is granted to the candidate writer, the home will broadcast invalidations to all members of the copyset and clean up the copyset. The writer will record its modified objects into the per-node volatile dirty list. The exclusive right will be returned to the home along when the modification (diff) is flushed. Read/write on home objects similarly need to go through the state check except that they are done directly on the object data without diff generation and flushing. There is no need to generate any write notices because volatile cache copies are *passively* invalidated by the home when a writer exists.

Upon read on an invalid volatile object, it will need to contact the home for the latest copy and join the copyset again. If the state of the home copy is exclusive, then the fetch request will be put into a queue pointed by volatile object header. When the writer returns the diff and exclusive right to the home, the home will turn the object state back to readable and reply all queued readers with the updated object data. As long as the state of a cache volatile object stays valid, its consistency has been guaranteed and the thread can directly trust it until invalidation is received when some writer exists. This leads to the beauty of this protocol that results in much better concurrency. Reads on a valid volatile object are pure local operations without remote locks and any communications. For high application read/write ratio, our design tradeoff shifts the communication overhead of the dominant reads to writes.

PERFORMANCE ANALYSIS

In this section, we present the performance results obtained by running Tomcat on JESSICA2.

6.1 Experimental Setup

Our experimental platform consists of three tiers: 1. *web tier*: a 2-way Xeon SMP server with 4GB RAM for running the master JVM of JESSICA2 with Apache Tomcat 3.2.4 started up on it. 2. *application tier*: a cluster of eight x86-based PCs with 512 MB RAM serving as the DJVM worker nodes. 3. *data tier*: a cluster of four x86-based PCs with 2GB RAM supporting MySQL Database Server 5.0.45. All nodes run under Fedora Core 1 (kernel 2.4.22). A Gigabit Ethernet switch is used to link up the three tiers, while nodes within the same tier are connected by Fast Ethernet networks.

The initial and maximum heap sizes of each worker JVM are set to 128MB and 256MB respectively. Each database node has the same dataset replica with MySQL replication enabled to synchronize data updates across database servers at nearly real time. Jakarta JMeter 2.2 is used to synthesize varying workloads to stress the testing platform.

Table 1. Application Benchmark Suite

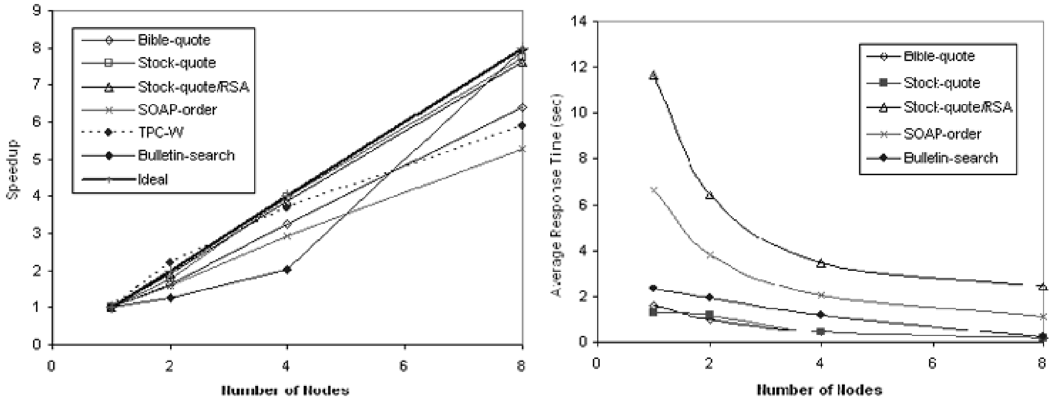
Application	Object Sharing	Workload Nature	I/O
Bible-quote	No sharing	I/O-intensive	Text files
Stock-quote		Relatively compute-intensive	Database
Stock-quote/RSA			
SOAP-order	HTTP session	I/O-intensive	
TPC-W			Database and image files
Bulletin-search	Cached database records	Memory-intensive	Database

Table 1 shows the application benchmark suite that we use to evaluate our clustering approach using the DJVM. They are designed to model real-life web application patterns.

1. *Bible-quote* characterizes applications like text search engines, news archives and company catalogs. The servlet application is I/O intensive, serving document retrievals and search requests over a set of text files of books.
2. *Stock-quote* models stock market data providers. We follow the trend of web services that deliver price data by XML messages. The application reads stock price data matching the input date range from the database and formats the query result into an XML response.
3. *Stock-quote/RSA* is secure version of Stock-quote involving compute-intensive operations of 1024-bit RSA encryption on the price data.
4. *SOAP-order* models a B2B e-commerce web service. A SOAP engine is needed to support the service. We choose Apache SOAP 2.3.1 and deploy it to Tomcat. The application logic is to parse a SOAP message enclosing securities order placements, validate the user account and order details and then put the successful transactions into the database.
5. *TPC-W* is a standard transactional web benchmark specification. It models an online bookstore with session-based workloads and a mix of static and dynamic web interactions. We adopt the Java servlet implementation developed by (ObjectWeb, 2005) but tailor the utility class for data access by disabling the default database connection pooling and utilizing thread-local storage to cache connections instead.
6. *Bulletin-search* emulates a search engine in a bulletin board or web forum system. We take the data dump from the RUBBoS benchmark (ObjectWeb, 2004) to populate the database. The application maintains a hash-based LRU-cache map of the results of the costly database searches, and is thus memory-intensive. In order not to lift up garbage collection frequency too much, we impose a capacity limit on the cache map, taking up about one-forth of the local JVM heap.

The original Tomcat is ported to JESSICA2 with a few customizations as follows: 1. the shared thread pool is disbanded. We replace the original thread pool by a simpler implementation which spawns a static count of non-pooled threads based on the server configuration file. 2. several shared object pools (e.g. static mapping tables for MIME types and status codes) are disintegrated into thread-local caches. The total lines of modified code including the new thread pool source file we introduce are less than 370 (about 0.76% of the Tomcat source base).

Figure 5. Scalability and average response time obtain by Tomcat on JESSICA2



6.2 Scalability Study

In this experiment, we measure the maximum throughputs and average response times obtained by scaling the number of worker nodes from two to eight. The speedup is calculated by dividing the baseline runtime of Tomcat on Kaffe JVM 1.0.7 by the parallel runtime of Tomcat on JESSICA2. Figure 5 shows the results obtained for each benchmark. We can see that most of the applications scale well and achieve efficiency ranging from 66% (SOAP-order) to 96.7% (Stock-quote). Bible-quote, Stock-quote and Stock-quote/RSA show almost linear speedup because they belong to the class of stateless applications, undergoing true parallelism without any GOS communications between the JVMs. In particular, Stock-quote and Stock-quote/RSA involve operations of coarser work granularity, such as string manipulations and RSA encryptions, and are hence more readily to attain nearly perfect scalability. The relatively poorer speedups seen by SOAP-order and TPC-W are expected as they are stateful applications and involve GOS overheads when sharing HTTP session objects among JVM heaps. We will further discuss the limited speedup obtained by SOAP-order in section 6.4.

Bulletin-search shows a nonlinear but steepening curve in speedup when the number of worker nodes scales out due to the implicit cooperative cache effect given by the GOS that we described in section 4. Along the scaling of nodes, when the cluster-wide aggregated available memory becomes large enough to accommodate most of the data objects cached in the application, the cache benefit will contribute an impulsive rise in speedup. Further study on this effect will be given in section 6.3.

Table 2 shows the cluster-wide thread count used in each application and the overall protocol messaging overheads inside JESSICA2 in the 8-node configuration. The count of I/O redirections is proportional to the request throughput and generally does not have impact on the scalability. The higher number of GOS protocol messages explains the poorer scalability obtained by the application if we reconcile with Figure 5. Bulletin-search is regarded as an exceptional case for its performance is more determined by its cooperative caching benefits which could supersede the cost of GOS communications.

6.3 Comparison with Existing Tomcat Clustering

A control experiment is conducted on the same platform to compare the DJVM approach with an existing clustering method for Tomcat using

Table 2. Protocol message overheads of JESSICA2 DJVM

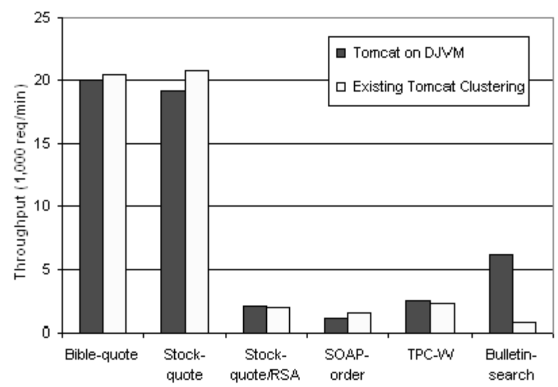
Application	# Threads	# GOS Messages / Sec	# I/O Redirections / Sec
Bible-quote	80	0	2006
Stock-quote	80	0	1791
Stock-quote/RSA	80	0	275
SOAP-order	16	979	146
TPC-W	40	351	1413
Bulletin-search	16	483	297

web load balancing plug-ins. We run an instance of Apache web server 2.0.53 on the web tier and eight standalone Tomcat servers on the application tier of our platform. The web server is connected to the Tomcat servers via the mod_jk connector 1.2.18 with sticky-session enabled (in-memory session replication is not supported in this comparison). The cluster-wide total number of threads and heap size configurations in this experiment are equal to the previous ones used in the DJVM approach.

Figure 6 shows the throughputs obtained by the two clustering approaches on eight nodes. We can see that both solutions achieve similar performance (within ±8%) for those stateless web applications (Bible-quote, Stock-quote and Stock-quote/RSA). These applications exhibit embarrassing parallelism and will not gain much advantage from the GOS. So putting the GOS aside, we can expect both solutions should perform more or less the same because both our transparent I/O redirection and mod_jk's socket forwarding are functionally alike for dispatching requests and collecting responses. Yet, extra overheads could be incurred in our solution when transferring big trunks of data via fine-grain I/O redirections and during object state checks.

TPC-W performs about 11% better on the DJVM than with mod_jk. One reason is that servers sharing sessions over the GOS are no longer restricted to handle requests bounded to their sticky sessions while load hotspots could happen intermittently when using mod_jk. On the other hand, SOAP-order performs 26% poorer on JESSICA2 than with mod_jk. The main factor that pulls down the performance is that the SOAP library has some code performing fairly intensive synchronizations in every request processing cycle. We will see later that the overhead breakdown presented in Section 6.5 echoes this factor. Bulletin-search performs 8.5 times better on the DJVM due to application cache hits augmented by the GOS. We will explain why the DJVM approach has significantly outplayed the existing solution in the next section.

Figure 6. Comparison of Tomcat on DJVM and existing Tomcat clustering



6.4 Effect of Implicit Cooperative Caching

Bulletin-search exemplifies the class of web applications that can exploit the GOS to virtualize

Table 3. Bulletin-search's cache size setting and hit rates augmented by GOS

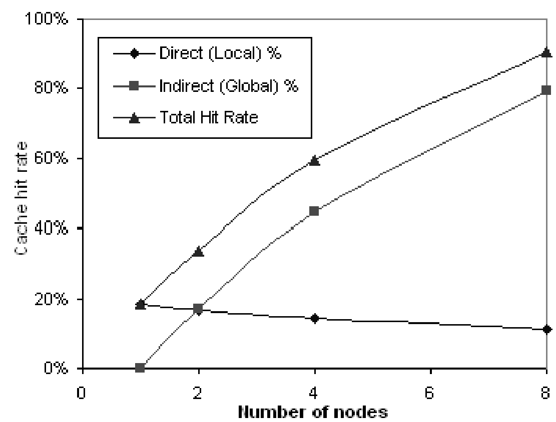
No. of Nodes	Cache Size (#Cache Entries)	Relative Cache Size	Total Hit Rate	Indirect Hit Latency (ms)	Cost Ratio of Miss: Indirect Hit	Throughput Speedup
1	512	12.5%	18.6%	N/A	N/A	N/A
2	931	22.7%	33.9%	9.07	40.79	1.26
4	1862	45.5%	59.3%	8.18	45.23	2.02
8	3724	90.9%	90.7%	11.74	31.52	7.96

a large heap for caching application data. Table 3 shows the application cache hits obtained by Bulletin-search when the number of cluster nodes scales from one to eight. With the GOS, the capacity setting of the cache map can be increased proportional to the node count beyond the single-node limit for different portions of the map are stored under different heaps. This is not possible without the GOS. Upon the creation of a new cache entry, its object reference built to the map is brought visible to all threads across synchronization points. So redundant caching is eliminated. Threads can exploit *indirect (or global) cache hits* in case the desired object is not in the local heap, easing the database bottleneck.

We can see from Figure 7 that the overall hit rate keeps rising along with the scaling of worker nodes of the DJVM and most of the cache hits are contributed by the indirect hits when the single-node capacity has been exceeded. This is the reason why our approach achieves a multifold throughput than the existing clustering approach in which there are only direct (local) hits that would level off or even drop slightly no matter how many nodes are added.

Here we define a term called *relative cache size (RCS)* that refers to the percentage of the aggregated cache size (combining all nodes) relative to the total size of the data set. When the RCS is below 50% in the 4-node case, the achievable cache hit rate is only around 60% and the 40% misses get no improvement such that the application obtains a speedup of merely two. But when the RCS exceeds certain level (e.g. 90% in the 8-node case), most of the requests are fulfilled by the global cache instead of going through the database tier. This explains the non-uniform scalability curve of this application in Figure 5.

Figure 7. Composition of application cache hits in Bulletin-search with GOS



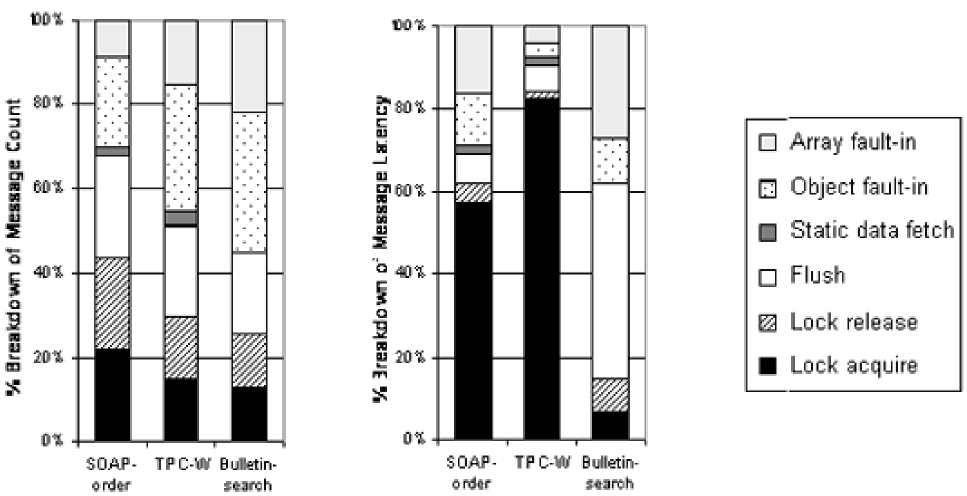
6.5 GOS Overhead Breakdowns

Table 4 shows the GOS overhead breakdowns in terms of message count per second for the three stateful applications. Figure 8 supplements with percentage breakdown of the message count as well as message latency. Lock acquire and release messages are issued when locking a remote object. Flush messages are sent upon lock releases but the flush message count is a bit more than lock release messages because in some cases updates are flushed to more than one homes. Other overheads are related to access faults which translate

Table 4. GOS overhead breakdown

GOS Message Type	# Messages / Sec		
	SOAP-order	TPC-W	Bulletin-search
Lock acquire	198	48	61
Lock release	198	48	61
Flush	217	70	92
Static data fetch	18	10	0
Object fault-in	197	99	160
Array fault-in	79	50	105

Figure 8. GOS percentage overhead



to communications with the corresponding home nodes. It is obvious that SOAP-order involves much more remote locking overhead than the other applications.

Our further investigation finds that one utility class of the deployed SOAP library would induce for each request about five to six remote locks on several shared hash tables and four remote locks on ServletContextFacade coming from the facade design pattern of Tomcat. Such heavy cluster-wide synchronization overheads justify the relatively poorer scalability given by this application.

Table 5 presents the local and remote locking rates for each application. We can see that local locks are much more than remote locks. The main reason behind this is that in Java-based servers, thread-safe reads/writes on I/O stream objects are exceptionally frequent, producing tremendous local locks. While local lock latency is very short (benchmarks shows an average of 0.2us), remote lock latency is however at least several thousand times longer in commodity clusters; yet remote locks are practically much fewer in most web applications. Another piece of information given by Table 5 is that SOAP-order and TPC-W have about 35% to 45% remote locks under cluster-wide contention, thus prolonging the wait time before locks are granted. This is why lock acquire has been the dominant part in the message latency for these two applications in Figure 8.

Table 5. Cluster-wide locking overheads

Application	# Local Locks / Sec	# Remote Locks / Sec	% Remote Locks Under Contention	Ratio of Local: Remote Locks
SOAP-order	232631	198	35%	1175:1
TPC-W	240470	48	45%	5010:1
Bulletin-search	27380	61	6.5%	449:1

RELATED WORK

Despite the boom of software DSM and the later DJVM research, it seems there have been only a few attempts at transparently supporting real-life server applications by means of shared virtual memory systems. Even fewer have been successful cases demonstrating good scalability though some of them had relied on non-commodity hardware to support their systems.

Shasta (Scales & Gharachorloo, 1997) is a fine-grained software DSM system that uses binary code instrumentation techniques extensively to transparently extend memory accesses to have cluster-wide semantics. Oracle 7.3 database server was ported to Shasta running on SMP clusters, albeit without success in achieving good scalability. They used TPC-B and TPC-D database benchmarks which model online transaction processing and decision support queries respectively. TPC-B failed to scale at all due to too frequent updates while TPC-D strived to achieve a speedup of one point something on three servers connected by non-commodity Memory Channel Network. To some extent, their experience and result exhibit many limitations of implementing a single system image at operating system level, compared to our approach of clustering at middleware level. For example, relaxed memory consistency model cannot be adopted at operating system level in usual cases, since correctness of binary applications often relies on consistency model imposed by hardware, which is generally much stricter than Java memory model. Being able to adopt relaxed memory model such as HLRC in our case is very important to server applications which may be intensive in synchronization.

cJVM (Aridor et al., 1999) is one of the earliest DJVM designed with intent to enable large multithreaded server applications such as Jigsaw to run transparently on a cluster. cJVM operates in interpreter-mode; it employs a *master-proxy model* and a *method shipping* approach to support object sharing among distributed threads. The system relies on proxy objects to redirect field access and method invocation to the node where the object's master copy resides. This model basically conforms to sequential consistency and is not efficient since every object access and method invocation may require communication although some optimization techniques were developed to avoid needless shipping. In contrast, our DJVM runs in JIT-compilation mode and conforms to release consistency, both propelling faster execution. In (Aridor et al., 2000), cJVM was evaluated by running pBOB (Portable Business Object Benchmark), a multithreaded business benchmark inspired by TPC-C, on a 4-node cluster connected by non-commodity Myrinet. They obtained an efficiency of around 80%. However, it is unclear that whether cJVM will perform such well if JIT gets enabled and commodity Ethernet is used as in our case.

Terracotta (Zilka, 2006) is a JVM-level clustering product emerged on the market for a couple of years. It applies *bytecode instrumentation* techniques similar to JavaSplit (Factor, Schuster, & Shagin, 2003) to a predefined list of common products and to user-defined classes for clustering among multiple Java application instances. Users need to manually specify shared classes as *distributed shared objects (DSOs)* and their cluster-aware concurrency semantics. Contrasting with our SSI-oriented approach, this configuration-driven approach may impair user transparency and create subtle semantic violation. Terracotta uses a hub and spoke architecture that requires setting up a central server, namely the "*L2 server*", to store all DSOs and to coordinate heap changes (field-level diffs) across JVMs. At synchronization points, changes on a DSO have to be sent to the L2 server that forwards the changes to all other clustered JVMs under the DSO's copyset to keep all replicas consistent. Our home-based protocol needs to keep only the home copy up-to-date by flushing diffs, then the next acquirer can see the changes by faulting in the whole object. Terracotta's centralized architecture may make the cluster susceptible to a global bottleneck when scaling out. Tailoring the bottleneck requires forklift upgrades on the L2 server (i.e. vertical scaling) that spoil the virtue of horizontal scaling using commodity hardware. We believe a home-based peer-to-peer protocol is a more scalable architecture for distributed object sharing.

CONCLUSION AND FUTURE WORK

In this chapter, we introduce a new transparent clustering approach using distributed JVMs (DJVMs) for web application servers like Apache Tomcat. A DJVM couples a group of extended JVMs for distributing a multithreaded Java application on a cluster. It realizes transparent clustering without the need for introducing new APIs and incorporates most of the advantages of a SSI-centric system such as global resource integration and coordination. Using DJVMs to cluster web application servers can enhance the ease of web application clustering and global resource utilization – both have been poorly met in most existing clustering solutions among the web community.

We port Tomcat to the JESSICA2 DJVM to testify this clustering approach. Our study addresses new challenges of supporting web application servers that characterize unique runtime properties of today's object-oriented servers over the classical scientific applications evaluated in the previous DJVM projects. The key challenge lies in making the system scalable with a large number of threads and offering efficient shared memory support for fine-grain object sharing among the JVMs. We enhance the cache coherence protocol design accordingly in several aspects: 1. adopt a unified cache among local threads to make better memory utilization; 2. implement a timestamp-assisted HLRC protocol to ensure release consistency of shared objects; 3. enforce sequential consistency among cluster-wide volatile fields via a concurrent-read exclusive-write (CREW) protocol. These improvements result in more relaxed coherence maintenance and higher concurrency. Our experimental result has illustrated significant cache hits obtained by using the global object space (GOS) to cache a large application dataset with automatic consistency guarantee.

Several trends have put forward the advent of the DJVM clustering technology. Today's web applications are becoming increasingly resource-intensive due to security enhancement, more complicated business logics and XML-based standards. Collaborative computing paradigm provisioned by DJVMs becomes vital to generate helpful cache effect across cluster nodes for efficient resource usage. Second, application logics tend to increase in complexity and now more and more application frameworks are POJO-based. Clustering at application level and adoption of proprietary clustering mechanisms shipped with particular application server products will tend to be laborious and error-prone, if not unfeasible. We foresee DJVMs, typifying the kind of generic clustering middleware systems, will be gaining more user acceptance. Third, design and development for user applications, server programs and library support nowadays have put more emphasis on scalability than ever. When scalability or performance portability is not a problem and meanwhile DJVMs are supreme in cost-effectiveness, this would have a catalytic effect that more applications readily go for the DJVM technology.

In future, we will investigate solutions to enhance fine-grain object sharing efficiency in the DJVM environment. In our research plans, we would consider incorporating transactional consistency (Hammond et al., 2004) into the cluster-wide memory coherence protocol.

REFERENCES

Adve, S. V., & Hill, M. D. (1993). A Unified Formalization of Four Shared-Memory Models. *IEEE Transactions on Parallel and Distributed Systems*, *4*(6), 613–624. doi:10.1109/71.242161

Antoniu, G., Bougé, L., Hatcher, P., MacBeth, M., McGuigan, K., & Namyst, R. (2001). The Hyperion system: Compiling multithreaded Java bytecode for distributed execution. *Parallel Computing*, *27*(10), 1279–1297. doi:10.1016/S0167-8191(01)00093-X

Aridor, Y., Factor, M., & Teperman, A. (1999). *cJVM: A Single System Image of a JVM on a Cluster*. Paper presented at the Proceedings of the 1999 International Conference on Parallel Processing.

Aridor, Y., Factor, M., Teperman, A., Eilam, T., & Schuster, A. (2000). Transparently obtaining scalability for Java applications on a cluster. *Journal of Parallel and Distributed Computing*, *60*(10), 1159–1193. doi:10.1006/jpdc.2000.1649

ASF. (2002). *The Apache Tomcat Connector*. Retrieved June 18, 2008, from http://tomcat.apache.org/connectors-doc/

Ban, B. (1997). *JGroups - A Toolkit for Reliable Multicast Communication*. Retrieved June 18, 2008, from http://www.jgroups.org/javagroupsnew/docs/index.html

Factor, M., Schuster, A., & Shagin, K. (2003). *JavaSplit: a runtime for execution of monolithic Java programs on heterogenous collections of commodity workstations.* Paper presented at the Proceedings of the IEEE International Conference on Cluster Computing.

Hammond, L., Wong, V., Chen, M., Carlstrom, B. D., Davis, J. D., & Hertzberg, B. (2004). Transactional Memory Coherence and Consistency. *SIGARCH Comput. Archit. News*, *32*(2), 102. doi:10.1145/1028176.1006711

Iosevich, V., & Schuster, A. (2005). Software Distributed Shared Memory: a VIA-based implementation and comparison of sequential consistency with home-based lazy release consistency: Research Articles. *Software, Practice & Experience*, *35*(8), 755–786. doi:10.1002/spe.656

Johnson, R. (2002). *Spring Framework - a full-stack Java/JEE application framework*. Retrieved June 18, 2008, from http://www.springframework.org/

JSR166. (2004). *Java concurrent utility package in J2SE 5.0* (JDK1.5). Retrieved June 24, 2008, from http://java.sun.com/j2se/1.5.0/docs/api/java/util/concurrent/package-summary.html

Keleher, P., Cox, A. L., Dwarkadas, S., & Zwaenepoel, W. (1994). *TreadMarks: Distributed Shared Memory on Standard Workstations and Operating Systems.* Paper presented at the Proceedings of Winter 1995 USENIX Conference.

Keleher, P., Cox, A. L., & Zwaenepoel, W. (1992). *Lazy release consistency for software distributed shared memory*. Paper presented at the Proceedings of the 19th annual international symposium on Computer architecture.

Ma, M. J. M., Wang, C. L., & Lau, F. C. M. (2000). JESSICA: Java-enabled single-system-image computing architecture. *Journal of Parallel and Distributed Computing*, *60*(10), 1194–1222. doi:10.1006/jpdc.2000.1650

ObjectWeb. (2004). *RUBBoS: Bulletin Board Benchmark*. Retrieved June 19, 2008, from http://jmob.objectweb.org/rubbos.html

ObjectWeb. (2005). *TPC-W Benchmark* (Java Servlets version). Retrieved June 19, 2008, from http://jmob.objectweb.org/tpcw.html

Perez, C. E. (2003). *Open Source Distributed Cache Solutions Written in Java*. Retrieved June 24, 2008, from http://www.manageability.org/blog/stuff/distributed-cache-java

Scales, D. J., & Gharachorloo, K. (1997). *Towards transparent and efficient software distributed shared memory*. Paper presented at the Proceedings of the sixteenth ACM symposium on Operating systems principles.

Stets, R., Dwarkadas, S., Hardavellas, N., Hunt, G., Kontothanassis, L., & Parthasarathy, S. (1997). Cashmere-2L: software coherent shared memory on a clustered remote-write network. *SIGOPS Oper. Syst. Rev.*, *31*(5), 170–183. doi:10.1145/269005.266675

Veldema, R., Hofman, R. F. H., Bhoedjang, R., & Bal, H. E. (2001). *Runtime optimizations for a Java DSM implementation*. Paper presented at the Proceedings of the 2001 joint ACM-ISCOPE conference on Java Grande.

Wilkinson, T. (1998). Kaffe - a clean room implementation of the Java virtual machine. Retrieved 2002, from http://www.kaffe.org/

Yu, W., & Cox, A. (1997). Java/DSM: A Platform for Heterogeneous Computing. *Concurrency (Chichester, England)*, *9*(11), 1213–1224. doi:10.1002/(SICI)1096-9128(199711)9:11<1213::AID-CPE333>3.0.CO;2-J

Zhou, Y., Iftode, L., & Li, K. (1996). Performance evaluation of two home-based lazy release consistency protocols for shared virtual memory systems. *SIGOPS Oper. Syst. Rev.*, *30*(SI), 75-88.

Zhu, W., Wang, C. L., & Lau, F. C. M. (2002). *JESSICA2: A Distributed Java Virtual Machine with Transparent Thread Migration Support*. Paper presented at the Proceedings of the IEEE International Conference on Cluster Computing.

Zilka, A. (2006). *Terracotta - JVM Clustering, Scalability and Reliability for Java*. Retrieved June 19, 2008, from http://www.terracotta.org

KEY TERMS AND DEFINITIONS

Copyset: The current set of nodes or threads that hold a valid cache copy of an object. This data structure is kept at the home node of the object and is helpful for sending invalidations in a single-writer-multiple-reader cache coherence protocol.

Distributed Java Virtual Machine (DJVM): A parallel execution environment composed of a collaborative set of extended Java virtual machines spanning multiple cluster nodes for running a multithreaded Java application.

Global Object Space (GOS): A virtualized memory address space for location-transparent object access and sharing across distributed threads. The GOS for distributed Java virtual machines is built upon a distributed shared heap architecture.

Java Memory Model (JMM): A memory (consistency) model that defines legal behaviors in a multi-threaded Java code with respect to the shared memory. The JMM serves as a contract between programmers and the JVM.

Lazy Release Consistency (LRC): The most widely adopted memory consistency model in software distributed shared memory (DSM) in which the propagation of shared page/object modifications (in forms of invalidation/update) is delayed to lock-acquire time.

Implicit Cooperative Caching (ICC): A helpful cache effect created by distributed threads through cluster-wide accesses to a collection of shared object references.

ENDNOTE

1 This research was supported by Hong Kong RGC grant (HKU7176/06E) and China 863 grant (2006AA01A111).

This work was previously published in Handbook of Research on Scalable Computing Technologies, edited by K.C. Li; C.H. Hsu; L.T. Yang; J. Dongarra; H. Zima, pp. 658-681, copyright 2010 by Information Science Reference (an imprint of IGI Global).

Chapter 8.10
Virtual Web Services:
Extension Architecture to Alleviate Open Problems in Web Services Technology

Julio Fernández Vilas
University of Vigo, Spain

Jose J. Pazos Arias
University of Vigo, Spain

Ana Fernández Vilas
University of Vigo, Spain

ABSTRACT

Several open issues in Web services architecture are being solved by using different kinds of solutions. Standard high-availability techniques based on the use of Web servers, business-logic-based caching systems, dynamic binding of Web services by programming the access to a SOAP message content from the business logic layer, and other kinds of current open problems can now be handled using a common unique technique. What we propose is to apply virtualization techniques to Web services.

DOI: 10.4018/978-1-60566-042-4.ch004

INTRODUCTION

When referring to current Web service architecture, a very important aspect to take care of is the one related to the separation of roles and the meaning of each role inside the architecture. Although the distinction between client, provider, and directory is clear, a great part of the Web services technology is based on a Web service offered by a provider (Booth, Haas, McCabe, Newcomer, Champion, Ferris et al., 2004). That is, according to the roles of the current proposed architecture, the provider is intimately related to the Web service it actually offers. In fact, both Web service and provider are used as only one role inside the architecture, called service provider. Several open problems of the current architecture can be solved by redefining this way of conceiving the roles inside the architecture.

Within the current architecture, the relation between client and provider has been established based on the use of two concepts:

- *Binding*. It is a process performed at development-time, consisting of adapting client software to the definition or description of a Web service.
- *Invocation*. It takes place on runtime, and it can be defined as the process by which a running client application calls a Web service.

The revised version of the W3C architecture redefines and merges these two concepts as "interaction." According to an "interaction" between client and provider, the binding is performed in a static way, so the way invocations must be performed is predefined. According to this:

- Dynamic binding cannot be performed, or, at least, not in an automatic way. There are different options based on the use of metadata ("Web services invocation framework," 2007) that enable the use of dynamic binding, but it is always mandatory to use metadata to access application data.
- Once a client application is bound to a concrete Web service, the execution of the application will be bound to the service provider selected, initially, at development-time. It will be necessary to bind a new Web service if we decide to use a new service provider. And this means that it will be necessary to develop new code to adapt the client application to the new interface of the new Web service. Although it is possible to change the location of the server providing the Web service without needing to make any modification to client application code, this will be only useful for those providers that have developed and published a Web service in the same way, that is, with the same parameters (its names and types), the same namespaces, and so forth.
- If a client application wants to use more than one provider in order to invoke the same equivalent service (offered by different service providers), the developer must bind the client application to each one of the Web services.

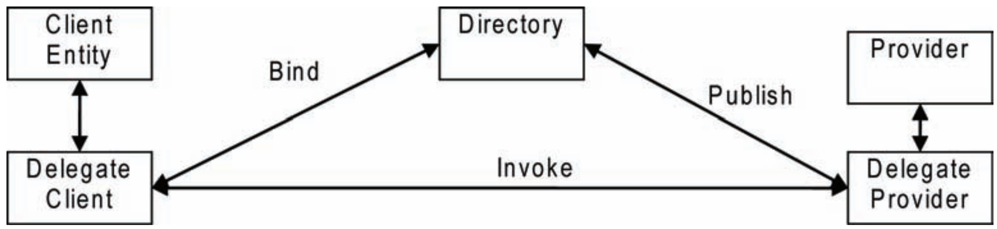
These and other less relevant problems have their origin in the fact that the role of "service provider" is not distinguished from the role of "Web service provider." When we talk about service-oriented architectures (SOA) (Barry, 2003; Vinosky, 2002), we usually use the terms "clients" and "providers," and we always relate these term with the use of services. We think it is necessary to slightly change the terminology, and use the terms "client application" when we refer to a SOA client, and "Web service provider" when we refer to a SOA provider.

In addition, we propose the creation of two new roles (as seen in Figure 1): the service provider (differentiated from the Web service provider), and the client entity (differentiated from the client application). The appearance of these two new roles will led us to a mandatory role separation. That is, Web services architecture is now composed of five roles: client entity, client application, directory, Web service provider, and service provider.

STATE OF THE ART

In the following subsections we will analyze the state of the art in several open issues in the Web services technology.

Figure 1. New roles



High Availability (HA)

When using production systems, the availability of a system is one of the most important issues to care about. HA is not considered as a part of the Web services architecture, but infrastructure providing companies (e.g., application servers, Web servers, load balancers, etc.) propose to increase the availability of a Web service by means of enhancing the availability of its underlying infrastructure.

There are two alternative ways to enhance the availability of a system:

- **Fault-tolerant systems**. All the hardware components in a fault-tolerant system are redundant (memory, processor, etc.), so, in case of a component fault, the whole system will continue running.
- **Cluster systems**. A cluster system is made of several computers (servers) that communicate using a high speed local area network (LAN). If a server fails, the whole cluster will continue running.

It is also possible to find mixed hardware configurations where fault-tolerant systems are used to build a fault-tolerant cluster, like Sysplex® (Nick, Chung, & Bowen, 1996), and cluster systems are made up of nodes build with redundant components.

As a result of hardware availability enhancement, providers protect their systems in order to offer the highest possible availability. But client's point of view is really different. If a provider stops accepting and responding invocations due to a problem in any hardware or software component (or even in the network), what clients notice is a system outage, that is, 0% availability, since the clients do not have any other way of invoking the service.

This situation can really happen, and it is due to the fact that the providers protect the implementation, while clients invoke an interface. Actually, all issues related with production systems of Web services (i.e., availability, scalability, etc.) have had no special attention inside current Web services architecture. The solution to all of them is always focused oi the implementation, not the interface.

Due to these lacks, client applications cannot have an assured minimum availability, since they are bound to a concrete provider, who may not comply with the conditions of a contract with that client. If clients could select automatically a provider when detecting an error condition, the availability of the service (independently of the provider used) would increase. This increment would be obtained independently of the availability of the implementation.

When enhancing the availability of a Web service, as we have mentioned above, protecting the implementation is not enough. The interface, what clients really use, should be equally protected. How? Just untying interface and implementation so the point where request from clients are received and the point where they are solved are really different and independent.

In order to reach this independence we propose the use of **indirect invocation**, so the interface can be published at an intermediate layer (responsible of dialoguing with clients), and the implementation is performed at a different layer, far from clients and invoked from the intermediate layer. This way, the intermediate layer can have some decision capabilities in order to, by publishing one only interface, invoke different implementations that reside in different points or nodes.

Quality of Service (QoS)

The term quality of service (QoS) refers to a set of techniques and technologies used with the aim of offering predictable results. For example, QoS techniques are used to assure a minimum bandwidth, a limited error rate, or a percentage of network availability. QoS is actually a concept that can be applied to any entity that provides a service, and from this point of view, a Web service can be analyzed from a QoS optic.

Variables like performance, response time, or availability can be used to measure the QoS of a Web service. But clients are also interested in other type of variables, that is, variables related to the provider of the service, and not just the Web service. That is, clients may be interested in things like provider cost or reliability. Once again, it is very important to separate the roles of service provider and Web service provider. Actually, we are talking about the need of differentiating "service metrics" (provider related) from "Web service metrics" (Web service related).

From clients' point of view, it is very important to know the value of a Web service's availability, for instance. But it is even more important to know the value of the availability of a service depending on the cost of that service. This assertion can be applied to any number of variables.

Independently of the number of variables that a client wants to handle, it is necessary for the client to handle them separately, differentiating the ones that define the behaviour of a Web service from those that define the behaviour of a Web service provider. The interface of a Web service has several associated metrics, but the implementation of that service has another different set of metrics.

There are several works published related to the QoS issue (e.g., Conti, Kumar, Das, & Shirazi, 2002; Menasce, 2002), but all of them only care about defining metrics, and they do not offer an architecture for implementing QoS. In addition, they do not offer the possibility to differentiate Web service metrics from provider metrics.

Weller (2002) analyzes a theoretical mechanism for qualifying Web services. According to this study, a new component is added to the standard architecture components, that is, the qualification repository, where clients store qualifications about providers, and other clients request qualifications about providers in order to perform an invocation. The repository computes and stores all qualifications about providers.

Our purpose of indirect invocation based on the use of expressions allows the use of any kind of variables in order to qualify a Web service and/or a service provider, and to obtain a unique value that can be used to select the most appropriate Web service at a given moment, according to a concrete QoS.

Multiprovider

As Web services technology grow, more provider companies will use it. That is, client entities will have an opportunity to enhance client applications, since they can use an increasing number of providers that will offer the same or equivalent service. But, for instance, they will differentiate in their availability, cost, or performance.

This ability to use more providers will become an additional new work when developing client applications, since each provider will use its own namespaces, schemas, parameters, and so forth. So, it will be mandatory to write ad-hoc code to integrate each one of the used providers into client applications, and developers will need to carry out more expensive and cost-consuming development processes.

Our purpose based on the virtualization of Web services, permits, in a natural way, to use multiple providers for only one Web service. As we will see below, Web services are published in a virtual way, while implementation of Web services can be performed by multiple different providers.

Error Management

The common lifecycle of a client application using Web services includes, at least, the following processes:

- To look for a Web service and a provider.
- To locate the description of the Web service (Web service definition language [WSDL] document).
- To perform a binding process, in order to build proxies, which will be the ones used to perform the invocations. This process will normally produce a set of modules that hold the definitions of: (i) the class of input data, (ii), the class of output data, and (iii) the invocable methods with its input and output parameters according to previous defined classes.

When a client application invokes a Web service, it is responsible of checking and handling any error condition that can occur. The following checklist should be implemented inside client applications:

- Invoke service.
- Check error.
- If invocation is not successful, determine if invocation can be retried.
- Retry the invocation a finite number of times.
- If invocation cannot be performed, send an error.

Let us consider having a client application bound to a Web service, and we want to stop using our current provider in order to use a new one (due to its low availability, for instance). We will have to modify our client application in order to delete all the references to the old provider and add new references (i.e., URL, types, namespaces, etc.) to the new provider.

In addition, it is usual (and it will become more usual in the future) to have more than one provider for the same service, in such a way that if one provider fails, client applications can invoke another provider.

With the aim of increasing the availability of client application, developers can decide to add more providers (and bind them) to that client application. This way, client application can retry invocation with a provider in case of an error of another one. But, according with what we have said above, a new checklist should be used:

- Invoke service.
- Check error.
- If invocation is not successful, determine if invocation can be retried.
- Retry the invocation a finite number of times.
- Select a new provider from the list of candidates and try again.
- If invocation cannot be performed, send an error.

As we can see, development and maintenance processes get more complex as we add more flexibility and availability to client applications. By using our purpose, we will be able to build Web services whose implementation is based on the selection and usage of several different and equivalent providers. This way, we will have a chance to enhance the availability of client applications without the need of complex development processes.

Caching

It is easy to understand how to implement a caching system when we talk about Web pages (HTML pages, images, etc., that is, static information). However, where (and how) can we apply caching techniques to Web services entirely based on the use of dynamic information (i.e., SOAP messages in response to different SOAP requests)? Our purpose enables the use of caching inside Web services technology, but it is important to take into account the fact that caching techniques can be applied to Web services only under specific circumstances:

- The number of invocation requests is large enough. If so, there will be a significant increase in performance when using a caching system.
- There is at least a method (a WSDL port, let us call it a "cacheable method") whose response to a request can be repeated along time without any variation in SOAP responses. If not, it may not be adequate for implementing caching mechanisms.
- The results received by clients invoking a cacheable method when the response is cached must not differ from those received by clients when the response is not cached.

Different applications in real life can be eligible to be implemented using cacheable Web services. Some examples of cacheable Web services are:

- A Web service that provides weather forecast, where changes in responses are not expected to occur in a short time.
- A Web service to obtain stock market share prices. Although this is a typical real time application (clients obtain real prices on each invocation), consider how the service runs when there are no changes to price, or when the stock market is close.
- A Web service to obtain the status of an item in relation to a stock management service (e.g., available quantity, type, model, etc.).

ARCHITECTURE DEFINITION

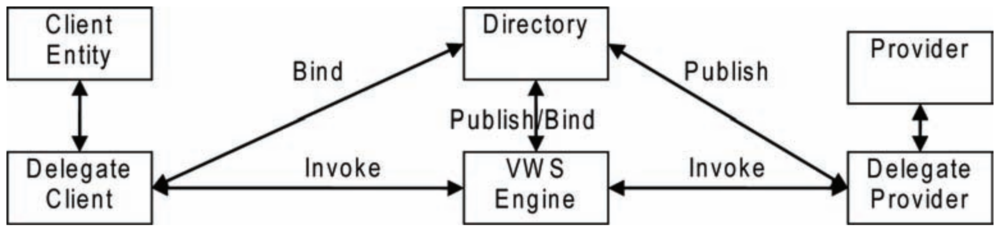
Our proposed architecture for the use of virtual Web services (VWS) determines the existence of, at least, five components (Figure 2): a client (client entity), a delegate client (client application), a service provider, a delegate provider (Web service provider), and a VWS engine. The client is the one who needs a service and, by the use of a delegate client, performs an invocation. The service provider is the one that publishes and offers a service through a delegate service provider. In our VWS architecture, invocations are not performed directly from the delegate client (typically a client application) to the delegate service provider (typically a Web service), but rather the delegate client performs an invocation to a VWS engine, who acts as an intermediary, and the engine performs an invocation to the delegate provider.

The VWS engine is not a virtual server. It is a standard server, and it should be implemented in the same way as a server used to process standard Web service invocations. The term "VWS engine" refers to a server that is capable of understanding virtual Web service definitions. That is, it can receive, process, and respond to standard Web service invocations, and, in order to process a request, the engine uses VWS definitions to select and invoke the most suitable Web service.

IMPLEMENTATION AND DEFINITION LANGUAGES

Any application that is able to use a standard Web service can be bound to a virtual Web service. Our virtualization technology provides a definition language (an XML-based language). This

Figure 2. Virtualization-based architecture



language (VWS definition language [VWSDL]) is used to write VWS documents. Clients do not use VWS documents, since these documents are just a definition of an implementation of a virtual Web service, and this definition is only useful to VWS engines. Our proposed language has been defined as a stand-alone language, that is, not as a WSDL extension, since its intended use is completely different. Even though the main objective of both languages is the same (describe a Web service), the way in which services are described are completely different.

VWS documents are used to describe virtual Web services, and they must contain, at least, a list of methods provided by the service (using the method elements, as shown in Figure 3). In addition, for each method published inside a service, we must specify the input and output parameters, and their corresponding data types.

As shown in Figure 3, each method element has its own name and type attributes. The type attribute is used to specify the type of implementation that is being defined. For instance, the equivalent value states that, in order to execute a virtual method, a set of "equivalent" services are available to accomplish the execution. The content of a method element is also used to specify a list of Web services and methods that will be invoked in order to complete the execution of a method. All those Web services and their methods represent in fact the implementation of a virtual method, and they are specified (inside the method element) using so many invoke elements as needed.

In SOAP language, parameters are nominal, that is, input parameters sent to methods use the name of the elements (XML elements). The same happens with return parameters. Let us suppose that we build a virtual method that receives a parameter called "P1" and returns a parameter called "RP." This represents a little restriction in the way we write the method and invoke elements, since the name of the parameters used in the virtual method must match the ones used in real methods. If a match can be found, the invoke elements can be used without a problem. When such a match cannot be found, a map element should be used to solve this situation. The map element can contain a set of in and out elements that are used to specify the correspondence between virtual parameters and real parameters. Inside each subelement (in or out), we must write an origin parameter name and a destination one, so the engine knows how to map the parameters before and after each invocation.

It is possible to find equivalent services that use similar parameters with different data types; a price is a typical example that can be handled using integer types or float types. In situations like this, a conversion process must be performed. This conversion is expressed via the type attribute. It is also possible to map a simple type to an element contained inside a complex one using XPath expressions, that is, assigning a simple value extracted from a DOM node.

The use of the map element brings much new functionality to the virtualization technique, since parameter mapping between virtual and

Figure 3. Sample method declaration

```
<method name="getPrice" type="equivalent" cos="COS#1">
      <select name="COS#1" expression="0.7*adjust(availability)+0.3*reverseAdjust(cost)" />
      <select name="COS#2" expression="0.5*adjust(availability)+0.5*adjust(security)" />
    <input>
      <parm name="ticker" type="xsd:string"/>
    </input>
    <output>
      <parm name="name" type="xsd:string"/>
      <parm name="value" type="xsd:float"/>
      <parm name="date" type="xsd:date"/>
    </output>
    <invoke id="2" method="sendItToMe" location="http://a.com/s1.wsdl">
      <map>
             <in origin="/ticker" target="t"/>
             <out origin="/name/shortName" target="/name"/>
             <out origin="/price" target="/value" type="relaxed"/>
      </map>
    </invoke>
      <...>
</method>
```

real services lessens the coupling level between client applications and service implementations. Some benefits we obtain include deploying new versions of implementation Web services with the same interface, and new interface versions using the same implementation Web services.

Expressions

In order to explain the expression system, let us consider the existence of three variables that can be used to define a Web service and/or a provider: availability, cost, and security.

When a client application needs to invoke a service, the client must decide which would be the most appropriate provider. Normally, a client would choose the one that has the lowest cost, the highest availability, and the highest security level. In our example, the client should choose service B according to the cost, service A according to the availability, service C according to the security level. To solve situations like this, what our VWS architecture proposes is to build an expression that combines all the variables and gives a unique value, pondering each one of the variables using factors. This way, a sample expression combining all three variables could be written as Expression (1).

```
0.6*A + 0.3*C + 0.1*S (1)
```

where A represents the availability, C is the cost, and S is the security level. To be able to write expressions in which we can mix different types of variables, having each one its own unit (milliseconds, dollars, etc.), we must apply a conversion factor in order to represent all the variables within the same scale. In addition, we should keep in mind the fact that some variables maximize the value of the whole expression by means of their maximum values (availability, for instance), while other variables do the same with their minimum values (cost, for instance).

In order to unify scales, ranges, and units for all variables used in an expression, the function adjust can be used. In addition, for variables whose minimum values represent a maximum scoring, the function reverseAdjust is available. Expression (1) can be rewritten as Expression (2) using these new functions.

```
0.6 * adjust(A) + 0.3 *
reverseAdjust(C) + 0.1 *
adjust(S) (2)
```

When an invocation request is received at the VWS engine, the engine must calculate the scaling factors to unify the scale of the variables that will be used in the expression. An example is shown in Figure 3 (select elements).

Complex Expressions

Two additional methods will allow writing complex expressions: complex variables and time-variables. A complex variable is a new variable that is created using a combination of simple ones. In Expression (3), a variable defined as G is used to hold a value that represents a global score for a provider, using A, C, and S.

```
G = 0.6 * adjust(A) +
0.3*reverseAdjust(C) + 0.1 *
adjust(S) (3)
```

$$adjust\ ((W*G_{i-1} + G_i) / (W + 1)) \quad (4)$$

On the other hand, time-variables allow writing expressions that involve present and past values of a variable. For example, Expression (4) can be used to obtain a global qualification for a provider, where time is taken into account by using two global-score values (i.e., G_{i-1} for the previous computed value of G and G_i for the present value of G), and a weighting factor (W) that is used as a stabilization mechanism.

External Variables

Our virtualization architecture does not impose any kind of constraint when writing expressions or using variables, and so we can use external variables managed by an external entity. For example, we could write an Expression (5) that includes an external status variable (T) indicating the status of the Web service. If a Web service is "out-of-service," we can set T to zero, and the VWS engine will stop selecting it.

T*(0.4*reverseAdjust(C) + 0.6*adjust(A)) (5)

Extension Mechanisms

Our virtualization architecture proposes the use of two additional mechanisms in order to help clients in having some decision capabilities over the evaluation process. The main objective of these mechanisms is to give the clients the ability to drive the evaluation process performed at the VWS engine.

The first mechanism is based on the use of classes of service (CoS). A CoS is defined (according to our purpose) by writing a concrete expression inside a method element (using a select element). More than one CoS can be defined by writing so many select elements as needed, identifying each of them with a name. Each select element represents in fact a given class of service. Once several classes of service have been defined inside a VWS document, clients can specify what CoS to use in each invocation by using SOAP extensions.

If a client does not specify what CoS to use, the VWS engine should use a default CoS in order to perform an invocation.

The second extension mechanism allows the client applications to prune the list of eligible Web services. Clients can send, attached to the SOAP requests, a set of conditions that implementation Web services must fulfill (Figure 4). This way, when the VWS engine has to perform an evaluation, it will consider only those services that fulfill the client's conditions.

The base for the implementation of our extension mechanisms is the use of SOAP header elements. Like other standards, we propose to add specific extension information to the header of SOAP response messages. The information contained in a SOAP header must obey a schema

Figure 4. Sample restriction

```
<retriction select="(availability &gt; 95) and (cost &lt; 20)" cos="COS#2" />
```

in order to use our cache purpose, that is, a virtual Web services extension language (VWSEL) (Fernández, Pazos, & Fernández, 2005), created as an extension of our VWSDL (Fernández, Pazos, & Fernández, 2004). VWSEL is used in our virtualization-based architecture to offer clients a chance to send processing information to VWS engines, such as classes of service, lists of providers, preferences about caching, and so forth.

Client applications must understand VWSEL in order to use VWS extensions. It is important to note that clients who do not understand VWSEL can still work without a problem, since a SOAP message (aside from including VWSEL data) is still a standard SOAP message that can be processed by client applications, engines, and providers.

It is responsibility of every Web service to decide what information will be included inside the VWSEL section of SOAP messages. Web services should use VWSEL documents to send only caching information to clients.

QUALIFICATION AND EVALUATION

Variables are used inside our solution to achieve the automation of two processes:

- A **qualification process:** The VWS engine computes and stores statistical information in a database after selecting and invoking a Web service. The same or another engine will use that statistical data in a later moment.
- An **evaluation process:** By means of this process, a VWS engine reads statistical information stored in the database, with the aim of selecting and invoking the most suitable Web service to accomplish the execution of a virtual service method.

We have already seen how to build expressions in which we can put together all we know about the way a given Web service works. The objective of expressions is to achieve quick and automated qualification and evaluation processes. Theoretically, every time a Web service is invoked, the VWS engine should update the statistical data related to that service.

Nevertheless, by means of a good performance, it could be good practice not to store all the gathered statistical data. As an alternative, we can decide to store the data associated to only a part of the invocations performed. This could be a recommendable way of working, but the percentage of data stored must be big enough, so they can be representative of the total number of invocations.

The process of updating the statistical database must be performed in an automated way, and this automation requires that the variables used in the process can be immediately computed after a service invocation. That is, the engine cannot update a variable that represents the time spent by a provider in serving orders. Actually, variables that cannot be computed from a service invocation are usually provider-related variables, and not service-related ones. Provider-related variables should be considered as external variables, that is, user-managed variables.

Let us suppose that we have developed a client application that uses a Web service to provide a book selling service over the Internet. Let us consider two sample variables: delivering time and package quality. Neither one can be computed when invoking the service since they are provider-related variables.

Provider-related variables, as it happens with service-related ones, can be either quantitative or qualitative, so the same conversion process used in service-related variables can be applied to provider-related ones. In the previous example, delivering time is a quantitative variable, while package quality is a qualitative variable.

The evaluation process of a provider-related variable can be performed the same as the service-related variables evaluation. Actually, both types of variables can be mixed in the same expressions, because one type of variables and the other should be correctly adjusted when performing an evaluation.

However, the qualification of a provider cannot be performed in an automatic way, because the values of variables (provider-related) can be unknown during the lifecycle of an invocation. When using expressions that include provider-related variables, the VWS must have access to data that represent the behavior of the providers in previous invocations. That is, a process external to the engine must update the statistics database. This way, when the VWS engine computes an expression, it can do it without caring about the type of variables used (service- or provider-related), since all the data will be available. From the point of view of the engine, there is no difference between one type of variables and the other.

Rating Subsystem

Our virtualization architecture proposes the use of an optional "rating server" as an extra component that should be in charge of managing statistical data. That is, the rating subsystem is in charge of the qualification and the evaluation processes in the previous section. The VWS engine and the rating server could be the same machine. However, in a big production system, an independent rating subsystem could be used.

The rating system defined inside our QoS architecture is not concerned with the quality of the server nor the client, but with the quality of the services and the providers. Our purpose does not care about the way the rating server is implemented.

When the VWS engine receives an invocation request, it must send a query to the rating server in order to select the most suitable Web service provider. Querying the rating server is a process that can be performed on each invocation or periodically, depending on the VWS engine criteria. Something similar occurs when the engine needs to update statistical data on the rating server. Additionally, in order to optimize the number of accesses to the rating server, the VWS engine could store the statistical data of all the invocations and update the server only at predefined intervals using summarized information. In production environments, it can be interesting to build rating systems that do not use real data obtained from the results of the execution of services, but only sampled data, because of performance reasons.

The rating server location is a major issue. A corporate server can be used, so it will be used as a "provider repository." Alternatively, a public rating server managed by a third-party company can also be used. In this case, the reliability of the rating system will depend on the trust placed in the provider of the rating server.

APPLICATIONS: HIGH AVAILIBLITY

Cluster Implementation

A VWS document that describes a virtual Web service whose invocations can be solved by a set of providers corresponds to what we call a "cluster of Web services." By using VWS documents, clusters can be built by using engines and providers. Our virtualization model establishes the existence of two types of nodes inside a cluster. The first type is the principal node. This node will be in charge of receiving service invocation requests. We must place the VWS documents inside the principal node. The second type is the nonprincipal node or provider node, and it is in charge of executing requests received from principal nodes.

Using the method and invoke elements of the VWS documents we can establish a relation between real and virtual services. This way, we can create a cluster architecture where the cluster's principal node (a VWS engine) will be the one in charge of receiving client requests and distributing the workload across cluster nodes (Web service providers).

According to this structure, if a provider (a cluster node) fails, the VWS engine will redistribute pending invocation requests, and the operative nodes in the cluster should take charge of unassigned workload. This way, the whole cluster continuity can be guaranteed. However, there are still some problems to solve. How can we accomplish such a load balancing system? How can we deal with a planned or unplanned node outage? How can we select the most suitable provider in each moment? The next sections will provide answers to these questions.

Building a Web Services-Based Cluster

Let us suppose that we have a Web service (WS1), with a method (M1). If we want WS1 to be a highly available Web service, we must deploy it to a cluster. The deployment process requires the Web service to be deployed to each one of the nonprincipal nodes in the cluster. This way, more than one instance of the Web service can be used, and these instances can be executed at different nodes. All of them are said to be equivalent.

To get our VWS engine up and running we must create a VWS document to define a virtual Web service (VWS1) that would contain, at least, a virtual method (VM1) (it is important to note that clients do not use the VWS documents to perform the binding; they use a standard WSDL document derived from the VWS document). Inside the VWS document, we must specify how the virtual method execution should be accomplished. We need to include, at least, three XML elements inside the VWS document: a service element describing the service; a method element that describes the method we want to publish (including its input and output elements); and one or more invoke elements. These invoke elements are the ones in charge of describing how and where the method implementation must be made.

When a method execution request for the VM1 method arrives to the VWS engine, the engine must select the most suitable provider in order to complete the request. Once a provider has been selected, a real service (implementation service) will be invoked, sending it input parameters as needed. After service execution, the engine will receive the return parameters from the real service and it will send them back to the client application. For the whole cluster to run accurately, the VWS engine has to decide which provider node would be the most suitable to perform an execution. The engine should use some selection criteria in order to maximize cluster performance.

Node Selection

After the VWS engine receives a request, the engine must select a cluster node that can accomplish the invocation request. To do it, the engine will examine the content of the method element included in a VWS document describing the virtual service. Among the providers detailed in the invoke elements (included in the method element), the VWS engine will choose the best prior to each real service invocation. This concept, the best, is a concept that can change along time.

Our virtualization model proposes the use of expressions to select the most appropriate provider prior to each invocation. Each method element inside a VWS document should include an expression. This expression must reflect which the priorities are when a provider has to be selected. Let us consider Expression (6)

```
0.7*A+0.3*R (6)
```

where A is the availability and R is the response-time, and they represent historical data about a given Web service. Prior to a Web service invocation, the VWS engine must compute the values associated with all services specified in the invoke elements inside a method element, then compare all result values, and, finally, invoke the service with the highest score.

In order to unify scales, ranges, and units for all variables used in an expression, the function adjust can be used. In addition, for variables whose minimum values represent a maximum scoring, the function reverseAdjust is available. So, Expression (6) could be perfectly adjusted and rewritten like Expression (7):

```
0.7*adjust(A)+0.3*reverseAdjust
(R) (7)
```

Cluster Maintenance

To succeed in having a 100% available system, methods that allow performing maintenance on the cluster must be available. When we need to perform any kind of maintenance, we meet with the need for stopping part of the system. If we want to stop a node without interrupting activity, we should use an "upDown" variable and add it to all expressions used inside the cluster. This way, we can rewrite Expression (7) as Expression (8):

```
upDown * (0.7*adjust(A)+0.3*reve
rseAdjust(R)) (8)
```

If we want to remove a node from the cluster, we must set the value of the upDown variable to zero, and the node will stop receiving new execution requests. That is, the node enters a draining state, and when there are no pending requests, the node can be removed from the cluster.

Node Error Detection

During normal cluster operation, errors can appear that can cause two different effects: increments in response time, and fatal errors like a node outage. The VWS engine is the element in charge of dealing with those kinds of errors.

If the response time of a given node is increased more than usual, the normal operation of the engine should make that node stop being used. This can be accomplished by adding a variable called R (response time) to all expressions. For example, using Expression (7) providers with the lowest response time will always get a high qualification value, while providers with a high response time would stop being selected for an invocation. If a node outage is detected, then the VWS engine should stop invoking services on that node. To deal with this situation the VWS engine must use the upDown variable, like in Expression (8).

When does the provider be used again? A simple procedure for dealing with this situation would consist on using a polling technique (PING or repetitive TCP-open). A most appropriate method consists on having a mechanism that allows the provider to send a notification to the engine, in order to notify the new state. VWS engines must provide a Web service, including a specific method (let us name it "upDownPort" method). Using upDownPort, providers can notify its actual state. Moreover, the upDownPort can also be used to stop a cluster node (by modifying the upDown variable value).

Scalability

Scalability problems are usually solved using two different types of solutions: vertical scalability, achieved by improving hardware configuration of the cluster nodes, and horizontal scalability, where new nodes can be added to the cluster. If we decide to use vertical scalability, we will find no problems when implementing virtualization, since Web services are completely independent on the hardware infrastructure. If we decide to use horizontal scalability as a way to extend the whole cluster capacity, we must search for alternative cluster structures, when building virtual services.

Our virtualization model sees the cluster as a tree, in which the root node is the VWS engine, and the provider nodes are leafs in that tree. The first way in we can extend a cluster consists on adding leafs to the tree (Figure 5).

The virtualization model also proposes another type of scalability, that is, hierarchical scalability. With VWS, we can publish virtual Web services and use them as if they were "traditional" (standard) Web services. A virtual service implementation can be done using another virtual service, and so we can build cluster structures that contain intermediate nodes. According to this, the root node of the cluster would be a VWS engine, leaf nodes would be provider nodes, and intermediate nodes should be implemented as a mixture of both a root node and a leaf node. Intermediate nodes should behave as a root node in order to send requests to leaf nodes under it, and as a leaf node that receives requests from a root node. In Figure 5, an expansion of "Node 2" has caused the creation of an intermediate node and the addition of two new nodes.

APPLICATIONS: QOS

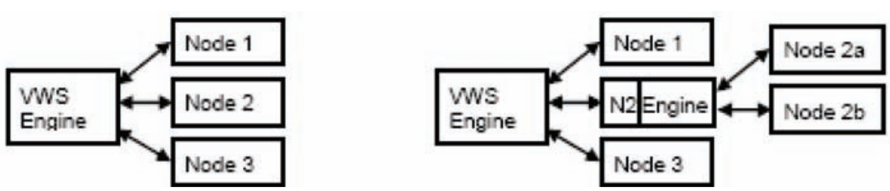
When developing software that uses Web services, things like the uncertainty associated to the Internet (e.g., unpredictable workload, uncertain and highly oscillating number of users, etc.), must be taken into account. In addition, the quality of the providers used (i.e., its availability, reliability, response time, etc.), will have a great impact on client's applications.

All the providers are different, and clients always want to use the best one. This concept, the best, is a concept that changes along time, due to network outages, server outages, high response times, and so forth. In order to use the best Web service from the best provider, we need to use systems that gather statistical information about providers and their services. That statistical information should be updated with time, so client applications always invoke the best service and provider according to that statistical data. As we have seen, our virtualization architecture proposes the use of three elements that enable the selection of the most suitable service and provider:

- A qualifying method for each one of the services and providers.
- A way to obtain historical qualifications of previous invocations.
- An element in charge of computing the qualifications and performing an evaluation using the information gathered previously.

A correct operation of the assign-and-compute-qualifications system depends directly on choos-

Figure 5. Hierarquical scalability



ing the appropriate variables, like availability, reliability, cost, or performance.

The election of the right units for the variables is a mandatory step before using a qualification system. Although there are variables like response time that can be easily measured, there are other variables like "provider security level," which are expressed using qualitative values (e.g., low level, high level, etc.). A simple way of managing the qualitative variables consists of assigning value ranges (using techniques similar to those used in fuzzy logic). This way we can map qualitative values to numeric values.

Let us suppose that we build a virtual service, and its implementation is made using three standard Web services (A, B, and C) that are equivalent (i.e., same functionality). We have decided to use the following variables to select the most appropriate service for each invocation: cost, availability, and security level. Cost is a variable whose value is established by the provider (provider-related variable), while availability and security level refer to a Web service (service-related variable).

Applying virtualization techniques to QoS problems is very easy. It only requires writing expressions that represent the QoS metrics that we want to take into account equally to the expression system we have showed for HA issues. Typycally, in clusters of Web services, we will use Web service-related variables, while in QoS problems we will use a mix of Web service provider-related variables and service provider-related variables.

APPLICATIONS: CACHING

It is possible to define at least two different caching structures in Web services architecture. The first one is a two level caching architecture (2LCA), which involves the existence of two entities: server and client. This caching architecture is not new, and caching is applied here using a typical caching implementation, like the one used by Web browsers that access Web servers directly. Three-level caching architecture (3LCA) is also possible. This kind of architecture is likely to occur in an environment where intermediate elements are an active part of the whole architecture. In a 3LCA, a third component appears, the intermediate element, which can also be an active actor in the caching system.

Our purpose is a very good solution to improve the performance of Web service-based applications, especially in stressed environments (with high volume of invocations). In addition, our purpose, far from the traditional HTTP cache system, gives a high degree of freedom to Web service developers, since our purpose, based on the use of extra information that is sent with SOAP messages (VWSEL), allows Web service programmers to control the way that messages are cached by client entities. That is, our purpose allows developers to control a cache system in a different way on each invocation, depending on the execution logic of the Web service.

There are some works on improving Web services performance by using caching; some of them are based on several different programming mechanisms. Goodman (2002) presents a solution based on the use of a cache object. This object is a Java object that must be managed from the Web service logic. This solution does not avoid network traffic, since invocation must be received at application code, in order to access the above-mentioned cache object.

Other kinds of solutions are based on the use of HTTP headers to manage the expiration of the HTTP responses. In "Perform Output Caching with Web Services in Visual C#.NET" (2003), the use of programming attributes (similar to compilation directives) is proposed. The use of the WebMethod attribute, together with the CacheDuration property, allows a simple way to control the TTL of the response. Again, this solution is based on the use of HTTP, enforcing a dependency between a Web service (and its business logic) and a transport layer (HTTP). In addition, it is a static mechanism, that is, programmers cannot control the duration of a response depending on the logic of the Web service.

RELATED WORK

Virtualization is being successfully applied to many other environments, such as storage virtualization, network virtualization, and hardware virtualization. What we propose here is to virtualice software, creating new virtual components (e.g., VWS) with which we can achieve a degree of decoupling and independence between clients and providers greater than the one we could achieve with standard Web services.

The architecture we propose is innovative as a global solution for a range of problems that have only been addressed individually so far. Problems related to SLA management, quality of service, or high availability are the subject of study by public and private entities, but the solutions proposed are specific for each one of these issues: architectures and languages to support SLA management (Dan, Ludwig, & Pacifici, 2003; Ludwig, Keller, Dan, King, & Franck, 2003; Sahai, Machiraju, Sayal, Jin, & Casati, 2002), metrics for QoS (Barry, 2003), and software and hardware architecture intended to improve the availability of the implementation of Web services, but not their interface, which is what clients perceive as a Web service. We have also discussed about these issues our 2004 work.

At the same time, the use of intermediate elements (the engines in our proposal) is a technique that is being implemented in some software platforms, but always with a specific use and using proprietary languages and/or systems. For example, WS-DBC (Brose, 2003) uses an intermediate element as a security system, while WS-Gateway (Venkatapathy & Holdsworth, 2002) isolates the private networks of clients and/or providers, also supporting certain protocol changes (e.g., from SOAP to HTTP/POST).

The work we have performed on caching systems has a close relation with the theories exposed by Cao and Irani (1997) and Cao and Liu (1997) and other works. Moreover, the same as the theories exposed by Chankhunthod, Danzig, Neerdaels, Schwartz, and Worrell (1996) and Worrell (1994), our caching purpose will add even greater benefits when applied to a hierarchic architecture, like the one we proposed in 2004, based on the use of VWS engines in charge of routing SOAP messages.

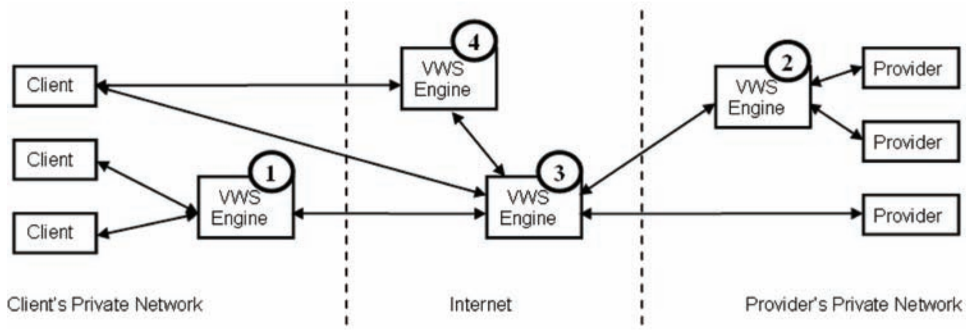
CONCLUSION AND FUTURE WORK

There exist several ways to apply the proposed architecture to the current Web services technology. Those forms represent different implementations and lead to obtaining different functionalities depending on the localization of the VWS engine. Basically, we can distinguish three alternatives, depending on whether the engine is located: (1) in the private network of the client, (2) in the private network of the service provider, and (3) in the Internet, being accessible to both client and service provider.

In case (1) above (point 1 in Figure 6), the use of an engine offers the following functionalities:

- It makes the software of the client independent of that of the delegate provider, because the engine allows both of them to change their interfaces without affecting the other, as long as the engine can map the different structures of the SOAP messages exchanged in the invocations.
- Provider independence. Virtual Web services allow the adding, modifying, and removing of providers without affecting the client, that is, without any need for the client to modify its applications.
- Proxy. The use of the engine allows client applications to invoke Internet Web services even when they have cannot connect to

Figure 6. Sample architecture implementation



anything beyond its private network.

- Automated error management. Because the methods of the virtual services are built from a list of Web service providers that is functionally equivalent, the engine can control the errors that take place and, when a provider fails, try to use another one. This occurs unnoticed for client applications.
- Cache. As we have seen, it is possible to cache client requests, noticeably improving response times. Take as an example the case of a Web service that offers share prices on the closing of financial markets, or a weather forecast Web service.

An engine placed inside the provider's private network (point 2 in Figure 6) fundamentally allows the building of Web services cluster systems, where the VWS engine acts as a controller node for the cluster, in charge of receiving requests and routing them to a certain node that is selected depending on the workload of each node (point 3 in Figure 6). Other possible functionalities are:

- Firewall. The VWS engine allows providers residing in a private network to be invoked from outside that network, keeping a high security level inside of it.
- When used as a cluster controller, it allows for the introduction of modifications in a node of the cluster while keeping the others unchanged, making it possible to perform software testing with a minimal impact in the construction of a new Web service in case of an error.

Last, in case (3) (Internet engine), its main use is that of a broker, that is, the engine acts like an intermediate component in the network that puts clients and service providers in contact (point 3 in Figure 6), moreover offering the following main functionalities:

- Decoupling between delegate client and delegate provider, due to the fact that the definition of virtual services makes it possible to modify the interface of the delegate providers (Web services) without changing the client software.
- Use of multiple providers.
- Error control and management.

What we propose is to use a common language for the description of virtual Web services, which at the same time provides a standard way to construct the interfaces that intermediate elements must offer through standard WSDL documents. We also propose an extension of the standard architecture in order to support VWS in such a way that it be compatible with current architecture.

VWS can help developing Web services with rich features like high availability, performance optimization, QoS, error management, and so forth.

The overall performance of the proposed architecture (whichever its use) will greatly depend on the variables and expressions used for the description of virtual Web services. It has to be noted that the VWS engine introduces a new overhead inside the execution architecture, since requests must be received and rerouted to the appropriate provider. However, this overhead is not significant when compared to the benefits obtained with our architecture.

Using VWS developers can build atomic Web services that can be published and subsequently consumed by resource-constrained devices like mobile phones or PDA; that is, virtual Web services can be used as a personalization mechanism regarding client requirements in order to simplify its use in such device types.

VWS technology is the base for other works that extend the use of our model. Regarding these other features of our model:

- We can use the VWS documents to build composite Web services. This work is in progress, and we are defining a set of different types of invocations. Our goal is to develop a Web services programming language (WSPL, as an extension of VWSDL) that supports basic programming structures (e.g., if-then-else, do-while, etc.). Its objective is to provide a simple composition method.
- We plan to integrate WSLA with our Web service descriptions. This way, a VWS engine can be used to analyze each invocation of a Web service and evaluate SLOs after each invocation.

Our proposal is not disruptive in its implementation, because it can coexist with the current architecture with no problems at all. Ideally, in fact, both architectures should coexist, because the standard one shall be used for easy problems in controlled environments, like the invocation of Web services in a corporate network.

REFERENCES

Barry, D. (2003). *Web services and service-oriented srchitectures: The savvy manage's guide.* Morgan Kauffman.

Booth, D., Haas, H., McCabe, F., Newcomer, E., Champion, M., Ferris, C., et al. (2004). *Web services architecture.*

Brose, G. (2003). Securing Web services with SOAP security proxies. In *Proceedings of the International Conference on Web Services* (pp. 231-234).

Cao, P., & Irani, S. (1997). Cost-aware WWW proxy caching algorithms. In *Proceedings of the 1997 Usenix Symposium on Internet Technologies and Systems (USITS-97).*

Cao, P., & Liu, C. (1997). Maintaining strong cache consistency in the World-Wide Web. In *Proceedings of the 7th International Conference on Distributed Computing Systems.*

Chankhunthod, A., Danzig, P., Neerdaels, C., Schwartz, M., & Worrell, K. (1996). A hierarchical Internet object cache. In *Proceedings of the 1996 USENIX* (pp. 153-163).

Conti, M., Kumar, M., Das, S. K., & Shirazi, B. (2002). Quality of service issues in Internet Web services. *IEEE Transactions on Computers, 6*(51), 593–594. doi:10.1109/TC.2002.1009145

Dan, A., Ludwig, H., & Pacifici, G. (2003). *IBM.* Retrieved May 25, 2008, from http://www-106.ibm.com/developerworks/library/ws-slafram

Fernández, J., Pazos, J., & Fernández, A. (2004a). An architecture for building Web services with quality-of-service features. In *Proceedings of the 5th WAIM.*

Fernández, J., Pazos, J., & Fernández, A. (2005). Optimizing Web services performance using caching. In *Proceedings of the 2005 NWeSP* (pp. 6-16).

Fernández, J., Pazos, J. J., & Fernández, A. (2004b). High availability wit clusters of Web services. In *Proceedings of the 6th APWeb* (pp. 644-653).

Goodman, B. D. (2002). Accelerate your Web services with caching. *developerWorks Journal*.

Ludwig, H., Keller, A., Dan, A., King, R. P., & Franck, R. (2003). *Web Service Level Agreements (WSLA) Project*. Retrieved May 25, 2008, from http://www.research.ibm.com/wsla

Menasce, D. A. (2002). QoS Issues in Web services. *IEEE Internet Computing*, *6*(6), 72–75. doi:10.1109/MIC.2002.1067740

Nick, J., Chung, J. Y., & Bowen, N. S. (1996). Overview of IBM system/390 parallel sysplex. In *Proceedings of 10th International Parallel Processing Symposium (IPPS)* (pp. 488-495). IEEE Computer Society.

Perform output caching with Web services in visual C#.NET. (2003). Microsoft Support.

Sahai, A., Machiraju, V., Sayal, M., Jin, L. J., & Casati, F. (2002). *Automated SLA monitoring for Web services.* Paper presented at the 13th IFIP/IEEE International Wokshop on Distributed Systems (DSOM*)* (pp. 28-41).

Venkatapathy, S., & Holdsworth, C. (2002). *An introduction to Web services gateway.* Retrieved May 25, 2008, from http://www-106.ibm.com/developerworks/Webservices/library/ws-gateway/

Vinosky, C. (2002). Web services interaction models. Current practice. *IEEE Internet Computing*, *3*(6), 89–91. doi:10.1109/MIC.2002.1003137

Web services invocation framework. (2007). Retrieved May 25, 2008, from http://ws.apache.org/wsif

Weller, S. (2002). *Web services qualification.* Retrieved May 25, 2008, from http://www-106.ibm.com/developerworks/library/ws-qual

Worrell, K. (1994). *Invalidation in large scale network object caches.* Unpublished master's thesis, University of Colorado.

This work was previously published in Managing Web Service Quality: Measuring Outcomes and Effectiveness, edited by K. Khan, pp. 74-94, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 8.11
Web-Based Corporate Governance Information Disclosure:
An Empirical Investigation

Yabing Jiang
Fordham University, USA

Viju Raghupathi
City University of New York, USA

Wullianallur Raghupathi
Fordham University, USA

ABSTRACT

Effective and timely use of the Internet, with a navigable and regularly-updated corporate Web site, can improve perceptions about a company's transparency vis-à-vis corporate governance practices. Based on an empirical analysis of data from 30 company Web sites in the DJIA, we conclude that a majority of companies underutilize their Web sites in communicating corporate governance information. Nearly all companies provide only routine content disclosure using minimal Web design features. A proactive display of compliance and due diligence content is largely absent, and most companies have yet to exploit advanced Web technologies to the fullest. Companies with better quality content and design appear to have higher corporate governance quotient (CGQ) ratings thereby implying greater transparency. The resultant framework can help companies vastly improve their Web sites by including more content that reflects due diligence and transparency and implementing Web 2.0 and other advanced technologies. Companies serious about transparency will adopt a more strategic approach to Web content and design. The lists of variables we identify can serve as checklists and/or templates for executives and researchers.

INTRODUCTION

The importance of proper governance has been underscored recently by a wave of well-publicized corporate scandals including Enron, Tyco, WorldCom and Global Crossing, to name a few. In response, firms have taken steps to strengthen their governance not only by making boards more independent but also by explicitly charging directors to enhance corporate transparency through the adoption of higher disclosure standards. Advances in information technology (IT), such as financial information systems (FIS), extensible business reporting language-enabled (XBRL) reporting, open conference-call technology, and Web 2.0 have greatly facilitated this endeavor by ensuring the timely production and cost-effective dissemination of corporate governance information. The key role of the corporate Web site, in particular, lies in facilitating communication between the public company and its stakeholders (Jones, 2002, 2003a, 2003b; Prentice, Richardson & Scholz, 1999; PricewaterhouseCoopers, 2004; Schultz, 2003).

This article draws, and merges, two streams of the literature, namely the corporate governance disclosure stream, and the role of IT in corporate governance stream (e.g., the Web) to identify and analyze the variables (features) with respect to content and design of corporate governance information on the companies' Web sites. The focus is to show prevailing Web-based mechanisms and practices and to suggest opportunities for improvement. We provide managers and researchers with a framework blueprint on the effective use of their Web sites to communicate and disseminate corporate governance information.

The rest of the article is organized as follows: First, we provide a comprehensive review of the literature in the areas of corporate governance disclosure and transparency and the use of the Web in information dissemination. Next, we describe the research framework for this study. Third, we outline our methodology and analyze the data collected. We then formulate a prescriptive framework for the proper use of the Web in enabling corporate governance disclosure. Fifth, we discuss the scope and limitations of our research. Lastly, we discuss our conclusions and suggest future research.

LITERATURE REVIEW

Background Information

Monks and Minow (2004) define corporate governance as "the structure that is intended to make sure that the right questions get asked and that checks and balances are in place to make sure the answers reflect what is best for the creation of long-term, sustainable value." Put more simply, governance refers to a body of rules and regulations that corporations must follow to protect the rights of their stakeholders (shareholders, creditors, employees, and others). A 2002 Global Investor opinion survey conducted by McKinsey & Co. highlighted corporate governance as a significant investment criterion (Bhat, Hope, & Kang, 2006). The survey found investors have higher confidence in companies with good corporate governance. Steps to meet increased public interest in governance transparency are reflected in recent governance regulations introduced by stock exchanges and regulators worldwide. The Sarbanes-Oxley Act 2002, for example, is a federal law that sets new and stricter standards for corporate governance practices in the United States (Bhat, Hope, & Kang, 2006).

Good corporate governance implies greater transparency. Berardino (2001) points out that "investors want governance that is designed and administered to protect the interests of all shareholders. They want companies to accurately disclose their financial position and business performance (SCMP, 2001, p. 12)." Thus, a major issue facing the business community is the need to increase company transparency and directors' accountability (Brooker, 2000) to meet or even exceed stakeholders' expectations.

Bushman, Piotroski and Smith (2004) define corporate transparency as "the availability of firm-specific information concerning publicly listed firms in the economy to those outside the firm, and it is conceptualized as an output from a multifaceted system whose components collectively produce, gather, validate, and disseminate relevant information." In other words, transparency is about providing the context of business operations that validates a company's financial projections, including such details as strategic plans, risk management, policies and corporate governance information specific to the company. But taking real steps toward greater transparency is not that easy exercise (Schultz, 2003). It is challenging to introduce practices that lead to greater transparency, especially when they aren't necessarily asked for. What type and/or amount of transparency will impress stakeholders but not give away too much information, such as details that might help the competition? Institutional Shareholder Services (ISS) (Institutional Shareholder Services, 2005) suggests that in order for companies to be effective at corporate governance programs they must certainly include a broad, proactive communication agenda that maximizes the effectiveness of traditional forms of communication, such as disclosures in public filings. And yet, companies must also embrace progressive practices, including accessibility to management and board members via new technology. A 2002 KPMG International report concurs and suggests broader and deeper disclosure on corporate Web sites and in annual reports. Web sites, the report states, should have a corporate governance section, which explains, for instance, procedures for getting a motion onto a proxy ballot (DiPiazza & Eccles, 2002; The Economic Intelligence Unit, 2002). In the next section we review and discuss the relevant literature on information disclosure, information asymmetry, and transparency.

Disclosure, Information Asymmetry and Transparency

There are several recent trends in disclosure regulation (Bethel, 2007) that impact the level and nature of disclosure. These trends have implications for how the Web may be used to disseminate information. The first trend in disclosure regulation is requiring firms to provide more comprehensive disclosure. The second, which has emerged over the last seven or eight years, is a shift from periodic and transactional disclosure to on-going or continuous disclosure. The third trend is a push towards broader dissemination and accessibility to information by investors. And the fourth trend is greater personal accountability by CEOs and CFOs for the quality of accounting statements (Bethel, 2007).

Disclosure regulation is usually associated with informational asymmetry or some form of externality (Healy & Palepu, 2001). When asymmetry of information exists between investors and internal management, adverse selection problems surface and bid-ask spread in securities price can widen. Not knowing the qualification of a company's true value, investors might attempt to shun the risk from asymmetry of information by overstating the discount rate on equities or risk premium on bonds (Chen & Jian, 2005). The market microstructure literature indicates that market liquidity increases as information asymmetry is reduced (Kanagaretnam, Lobo, & Whalen, 2007). An emphasis on information disclosure could eliminate such asymmetry and reduce capital and operating costs (Verrecchia, 2001).

Also, an agency problem surfaces when asymmetry of information exists between information providers and internal management. Instead of creating maximum value for shareholders, self-interested parties might opt for wasteful, privileged expenditures or overinvestment, a typical form of moral hazard cause by information inconsistency. One well-known solution to asymmetry is regulated disclosure. Both Leftwich (1980) and Watts and Zimmerman (1986) suggest that by creating minimum disclosure requirements, regulation of disclosure reduces the information gap between “informed” and “uninformed” agents in the economy. On this topic, Healy and Palepu (2001) point out that there are three reasons for companies to volunteer their corporate information: to increase liquidity of stocks, to reduce capital costs, and to gain media coverage. Other studies (Diamond, 1985; Verrechia, 2001) examined the consequences of voluntary information disclosures and concluded that such releases benefit investors. Information asymmetry declines as these public information releases “level the playing field” for all-investors (Pozen, 2006). In addition, the need for investors to engage in costly information gathering is reduced. Transaction cost theory too, suggests that greater transparency goes hand in hand with lower capital costs because of reduced transaction costs and/or reduced estimation risk (Habib, 2006). Callahan, Lee and Yohn (1997) conclude that, “by improving the information environment for companies through better disclosure, accountants can contribute to a reduction in transaction costs, and hence the cost of capital” .

Lastly, prior studies also have documented that the quality of a firms’ mandatory and voluntary disclosures both increase with the quality of the firms’ corporate governance. In the case of mandatory financial reports, better quality governance is associated with a lower likelihood of financial statement fraud (Beasley, 1996). And better quality governance is associated with higher overall level of voluntary disclosure (Eng & Mak, 2003). In summary, both mandatory and voluntary disclosures have a positive impact on the overall performance of a company.

In the next section we examine the role of the Web in governance disclosures.

Web-Based Disclosure

IT such as the Internet and the Web has great potential to disseminate corporate governance information effectively leading to better quality disclosure and improved transparency. For example, XML and extensible business reporting language (XBRL) provide financial information users with a standardized Internet-based method to prepare, publish, and exchange financial information (Boritz & No, 2005; Debreceny et al., 2001a).

Several studies have attempted to identify the determining factors of Internet-based corporate disclosure (ICD). Ashbaugh, Johnstone and Warfield (1999), Craven & Marston (1999), and Ettredge, Richardson and Scolz (2001) found that larger U.S. and U.K. companies are more likely to provide financial information on the Internet. Ettredge, Richardson and Scolz (2002) studied the determinants of voluntary ICD in the Unites States. A distinction was made between disclosure of items that are mandated by the SEC, and those not required by regulation. The findings indicated that the former are associated with company size and information asymmetry between the company and investors whereas the latter vary with company size, information asymmetry, demand for external capital, and disclosure reputation. However, the study only examined the content aspect of disclosure.

In a study of the application of the Internet in reporting among Chinese-listed companies, Xiao, Yang and Chow (2004) developed a disclosure index of 82 items based on the framework of Web-based disclosure proposed by Debreceny, Gray and Mock (2001b) with further references to other prior studies by Deller, Stubenrath and Weber (1999), Marston and Polei (2002), and Pirchegger and Wagenhofer (1999). Their index encompassed 58 items of disclosure content, and 24 items about presentation format. The content items showed what the companies disclosed on their Web sites. These included items required by the China Securities Regulatory Commission (CSRC) and non-CSRC-required items. The presentation format items dealt with how the information was presented (e.g., whether it was in processable format) and its convenience of use (e.g., whether there was an internal search engine).

A study by The Economist Intelligence Unit looked at the Web pages of the top 10 firms by market capitalization in each of five countries: the U.S, U.K., Japan, France and Germany over a three-week period in July 2002 (The Economic Intelligence Unit, 2002). Each company was assessed for the provision and accessibility of information on 29 different governance issues, including disclosure on executive pay, information on non-executive directors, retention of auditors and ease of voting at the annual general meeting (AGM). The study found most companies (two-thirds of the 50 reviewed) offered a separate and easy-to-find section on corporate governance. The exceptions were in Japan and the Unites States, where such sections were only available on half of the Web sites reviewed. Nevertheless, transparency in some areas was spotty. If, for example, one wanted a record of how often non-executive directors attended board meetings, most companies (94%) did not disclose the information, and only in two cases was one able to find the information in less than 30 minutes of searching. Governance information from CEO searches to selection of directors and auditors to shareholder voting rights typically was found either buried in or missing from the Web pages and annual reports (The Economic Intelligence Unit, 2002).

In 2002, Blunn and Co. (2003), an investor relations consulting firm, conducted a study to gauge the extent to which companies had improved their disclosure of governance information on their Web sites. The surveyors inspected the participant Web sites against the NYSE, SEC and Sarbanes-Oxley requirements for Web site disclosure. The result: Blunn found that very few companies complied with the new NYSE governance listing standards. Of the 135 companies reviewed, 84% did not have a corporate governance section on their Web sites; only 14% published their corporate governance policies prominently; and less than 25% published their companies' corporate codes of ethics (Jones, 2003a).

In a follow-up study by the same company conducted from March 3 to 26, 2003, the corporate Web sites of 250 public companies in several countries were surveyed to assess the breadth of corporate governance information disclosed online (Blunn & Co., 2003; Jones, 2003b). This study concluded there was room for improvement across the board in the breadth and depth of what companies publish online. Only 33% included at least their corporate governance policies, key committee charters and a code of conduct applicable to senior officers. Only 10% provided insider transaction information. A large number of the companies did not identify either the financial experts on the board or the audit committee nor did they provide information on their managements' internal control structure, systems or reporting processes. Further, very few companies had posted procedures for reporting accounting complaints to the audit committee. Lastly, the study showed that except for a handful of best practice companies, governance information was for the most part difficult to find on corporate Web sites. Typically, the information was buried among a variety of print documents that had been scanned and posted to sites (Blunn & Co., 2003) without attention to accessibility.

The investor relations global rankings (IRGR) is a comprehensive audit and ranking system for IR Web sites, corporate governance, and earnings release and disclosure procedures. As far as corporate governance is concerned, IRGR's evaluation criteria include only major content issues (MZ Bulletin, 2006). A recent study by the *Financial Times* assessed top-class corporate Web sites (Bowen, 2007) and concluded that the best Web pages were managed according to well-observed rules and processes. These top sites also offered high-quality service to stakeholders and provided good contact points. More importantly, they made exceptional use of Web technology, with broadband spreading, video, pod casts and other features. Missing, however, was much evidence of Web 2.0 technology and its interactive functions for building online communities (This may be because these corporate giants are laggardly or because they are rationally cautious.).

These studies have several limitations. First, nearly all of them studied content alone. By overlooking design, researchers ignored an important aspect of Web site value. Second, with the exception of the FT and IRGR studies, earlier research is outdated. Regulatory requirements and Web technology have changed dramatically in recent years. Finally, there's been an attempt to link, at least minimally, Web pages' content and design to some measure of overall corporate performance.

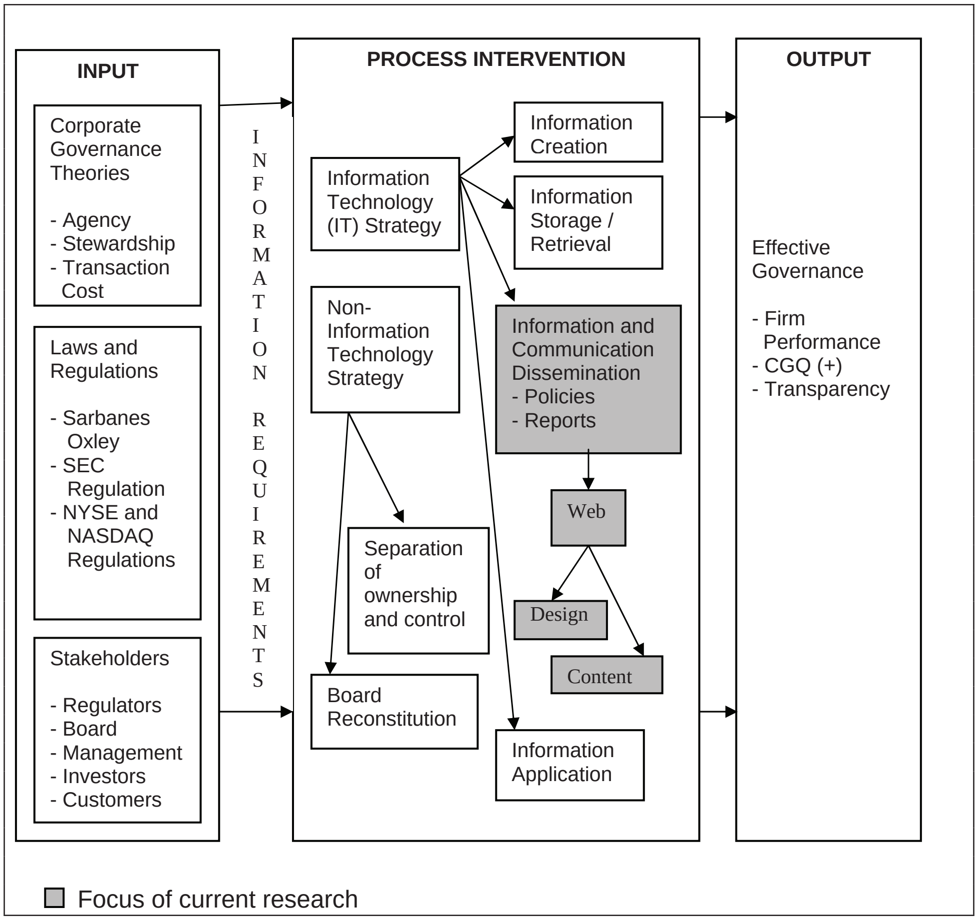
Our study revisits the question of the successfulness of corporate Web sites by looking at a more representative sample and addressing content *and* design issues. Also, we attempt to find a correlation between Web pages and a company's corporate governance quotient (CGQ) as a performance measure. The CGQ is discussed under methodology.

CONCEPTUAL FRAMEWORK FOR THIS RESEARCH

Our conceptual framework for the relationship between the Web and disclosure builds on the work of Bushman et al. (2004). They provide a framework for conceptualizing the informational environment of the firm. They classify information mechanisms in their framework into three categories: corporate reporting (e.g., Web content), private information acquisition, and information dissemination (e.g., Web design). Corporate reporting involves periodic disclosure of firm-specific information on a voluntary or mandatory basis. With regards to private information acquisition, relations between public information disclosure and the private information processing and gathering activities of investors have been recognized as important determinants of information allocations in an economy (e.g., Verrecchia, 1982). The inclusion of the third component, information dissemination was motivated by Bushman et al.'s (2004) perspective that lack of a well-developed communication infrastructure (e.g., Internet-based) may impede the flow of information reported by firms, limiting the availability of the information to economic agents. One could measure information dissemination by analyzing the content and design features of the Web sites.

Figure 1 represents a conceptual framework of corporate governance and its relationship to Web pages. The *inputs* represent corporate governance theories, such as agency theory and stewardship theory that provide a framework in which managers need to perform mindful of the interests of the stakeholders. First, typically governance is the charge of executives at the highest level. Secondly, various laws and regulations now specify financial and other procedures regarding disclosure of corporate information to shareholders, creditors, employees and others. The Sarbanes-Oxley Act and the Securities and Exchange Act (SEC) are examples of newer regulations for governance. Another input for governance is the information requirements of stakeholders (Hermansen, Hill, & Ivancevich, 2000). Shareholders, creditors, customers and employees expect access to documents that will give them an accurate view of a company's position and standing. Such documents include corporate business policies, statements, and the charge of various committees, corporate ethics, and social responsibilities, among others. These inputs are transformed by *process interven-*

Figure 1. Conceptual Framework of Relationship between Corporate Governance and Web Sites



tionists, which facilitate the execution of various tasks and enable conversion of primary inputs into the desired output, namely effective governance. The inputs have a common element in that they emphasize the flow of information to and from companies (e.g., to stakeholders). Information technology (IT), such as the Web, provides the tools and resources necessary to provide a smooth and efficient flow and exchange of information. In addition, non-IT strategies include ensuring effective governance through reconstitution of the board, separation of ownership and control and others. This focus deals with the board of directors, executives, and their activities. The IT strategy, on the other hand, focuses on intervention in four corporate governance processes involving information flow: information creation, information storage and retrieval, communication and dissemination/transfer, and application. Creation involves development of new content or replacement of existing content to make better decisions (Alavi & Leidner, 2001), such as using report generators to create reports. Storage and retrieval refers to the means by which information repositories are developed and retrieval mechanisms are set up to retrieve and utilize information. An example might be software that allows companies to store financial information in an easily retrievable format. Communication and dissemination refers to the presentation and distribution of governance information to stakeholders thereby ensuring increased accessibility and availability beyond the barriers of time and space (Alavi & Leidner, 2001). IT, (for e.g., say the Web), increases a company's reach beyond the formal communication lines in an organization. Web and pod casting, blogs, discussion forums and electronic bulletin boards are examples of Web facilities that increase and improve the diffusion of governance information. An important component of information dissemination is its relevance and timeliness.

The research in this study focuses on the use of Web pages to communicate and disseminate governance information. The desired *output* of effectively designed Web pages is a high level of corporate governance transparency. And by extension, this perception of transparency correlates with a firm's overall performance and measures (e.g., CGQ). The next section describes the methodology and data analysis for our study.

METHODOLOGY AND DATA ANALYSIS

The 30 companies that make up the Dow Jones Industrial Average (the Dow 30) were used for this study. These companies are also included in the S&P 100 and provide a fair representation of large to mid-cap companies.

We recorded the CGQ (corporate governance quotient) for each company as of March 1, 2007. The CGQ metric was developed by Institutional Shareholder Services (www.riskmetrics.com) to measure and compare the corporate governance structures of public companies. The ratings are relative and are reported on a percentile basis ranging from zero to 100%. For this study we created two groups: companies with a CGQ greater than 75% (high) and companies with CGQ less than 75% (low). We decided on this classification after reviewing the literature on CGQ available at www.riskmetrics.com. Generally speaking, companies with higher CGQ scores have better governance. A company's CGQ ratings appear on the first page of each ISS proxy analysis. Each company receives two CGQ ratings: The first score compares the company's corporate governance practices with a relevant index such as the S&P 500, the S&P (mid-cap) 400, the S&P (small-cap) 600, or the Russell 3000; the second score compares the company's corporate governance practices with its industry peers using S&P's 23 sector groupings.

According to ISS, the CGQ ratings comprise eight core topics (and 61 sub-issues), namely board structure and composition; charter and bylaw provisions; laws of the state of incorporation; executive and director compensation; quantitative factors such as financial performance; director and officer stock ownership; director education; and audit. The score for each core topic represents a set of governance variables. The eight core topics currently include six rating variables. If ISS indicates that a company's index ranking was 37.1, it means, in terms of governance, that the company had outperformed only 37.1% of the companies in the S&P 500. ISS assigns overall corporate governance grades to over 5,100 public companies, and it provides them free on the 'company profile' pages on Yahoo Finance (http://yahoo.finance.com). Due to our limited sample size, we recorded only the percentage comparison to the S&P 500.

Building on prior literature discussed in the section of Web-based disclosure, we address two categories of variables: content and design. The set of individual variables constituting 'content' and 'design' were identified based on a large literature review of previous studies, information obtained from federal laws (e.g., Sarbanes-Oxley), regulatory bodies' (e.g., SEC, NYSE, Nasdaq), and consultant/service provider organizations (e.g., Institutional Shareholder Services, Governance Metrics International). Content variables refer to the various corporate governance disclosures and documents, voluntary and mandatory, displayed on a company's Web sit (Blunn & Company, 2003; Jones, 2002, 2003a; 2003b; Institutional Shareholder Services, 2005; MZ Bulletin, 2006;

Table 1. Descriptive Statistics for Content – CGQ

	Group1 (n=16) (CGQ > 75%)		Group 2 (n=14) (CGQ < 75%)		Total (N=30)	
CONTENT VARIABLES	**#**	**%**	**#**	**%**	**#**	**%**
Board						
Direct Link to Contact Directors/Board	8	50.0	8	57.1	16	53.3
Director Biographies	12	75.0	12	85.7	24	80.0
Director Compensation	7	43.8	4	28.6	11	36.7
Duties and Responsibilities of Lead Director	3	18.8	1	7.1	4	13.3
Independent Director Information	3	18.8	4	28.6	7	23.3
Information on Committees They Serve	15	93.8	13	92.9	28	93.3
Committees / Charters						
Audit Committee	16	100.0	14	100.0	30	100.0
Charter of the Lead Independent Director	2	12.5	0		2	6.7
Compensation Committee	15	93.8	13	92.9	28	93.3
Executive Committee	7	43.8	8	57.1	15	50.0
Governance Committee	16	100.0	13	92.9	29	96.7
Nominating Committee	12	75.0	6	42.9	18	60.0
Specify Financial Experts in Audit Committee/Board	4	25.0	2	14.3	6	20.0
Compliance Policies						
Information on comprehensive Compliance Program	2	12.5	2	14.3	4	13.3
Procedure on Reporting Acct. Issues to Audit Committee	3	18.8	4	28.6	7	23.3
Protection to Report Abuses / Anonymity / Whistleblower Policy / Ombudsman	4	25.0	0	0.0	4	13.3
Ethics						
Code of Business Conduct And Ethics	14	87.5	12	85.7	26	86.7
Political Donations/Contributions Statement/Report	6	37.5	6	42.9	12	40.0
Code of Business Conduct/Ethics for Board/Executives	5	31.3	6	42.9	11	36.7
Governance Documents						
Articles/Certificate of Incorporation/Bylaws	11	68.8	11	78.6	22	73.3
CEO and CFO Certifications	3	18.8	4	28.6	7	23.3
Corp. Gov. Guideline/Principles	16	100.0	14	100.0	30	100.0
Corp. Gov. Guidelines/Principles in Interactive Format	8	50.0	3	21.4	11	36.7
Director/Executive Stock Ownership/Pension Guidelines	6	37.5	3	21.4	9	30.0
Securities Transactions by Directors and Officers	15	93.8	13	92.9	28	93.3
Stock Option/Ownership Details	1	6.3	1	7.1	2	6.7
Analyst/Investor Conference Call	15	93.8	11	78.6	26	86.7
Management Team						
Executive Compensation Guidelines	5	31.3	1	7.1	6	20.0
Management Team Biographies	11	68.8	11	78.6	22	73.3
Press Room						
News Releases	16	100.0	14	100.0	30	100.0

continued on following page

Table 1. continued

Proxy						
Link to Proxy Statement	14	87.5	14	100.0	28	93.3
Proxy in Interactive Format	6	37.5	8	57.1	14	46.7
Social Responsibility						
Dedicated Section for Social Responsibility	11	68.8	14	100.0	25	83.3

Table 2. Descriptive Statistics for Design – CGQ

	Group1 (n=16) (CGQ > 75%)		Group 2 (n=14) (CGQ < 75%)		Total (N=30)	
DESIGN VARIABLES	**#**	**%**	**#**	**%**	**#**	**%**
Interactivity						
Blogs	1	6.3	2	14.3	3	10.0
Charting Capability/Analysis Tool	15	93.8	11	78.6	26	86.7
Facility to Download in More than one Format	12	75.0	12	85.7	24	80.0
Facility to Fill in an On-Line Request Form	8	50.0	9	64.3	17	56.7
Email Link from Corporate Governance Page	8	50.0	8	57.1	16	53.3
Availability of Survey (eg., Visitor Satisfaction Survey)	5	31.3	2	14.3	7	23.3
Software Download Support	11	68.8	13	92.9	24	80.0
Location						
Location of Corporate Governance Link in Main Page	6	40.0	4	28.6	10	33.3
Multimedia						
Podcast Feature	4	25.0	6	42.9	10	33..
RSS Feeds	6	37.5	8	57.1	14	46.7
Webcast Archive Features	16	100.0	14	100.0	30	100.0
Access Webcast without Registration	5	31.3	2	14.3	7	23.3
Navigation						
Capability to View Page in Multiple Languages	6	37.5	5	35.7	11	36.7
Index/Site Map/Directory	11	68.8	10	71.4	21	70.0
Last Updatedd Indicator	8	50.0	6	42.9	14	46.7
Search Box	16	100.0	13	92.9	29	96.7

Schultz, 2003). The individual and composite content variables are listed in Table 1. Typically, past studies have looked only at content issues. In this study we also look at various design variables listed in Table 2. The design variables were also grouped into composites, each characterizing a certain theme. Using company home pages as starting points, two researchers collected content and design information by independent manual inspection. This is consistent with how content analysis is performed. Differences were reconciled by re-visiting the Web pages. We looked for the presence or absence of each content variable and coded it in binary format (1=present, 0=not present). Likewise, the presence or absence of each design feature was recorded. The data was summarized and tabulated to develop frequency counts and percentages for the two CGQ groups. Composite percentage scores were also calculated to obtain a big picture view. Apart from the independent study and findings of the content and design variables, we also wanted to identify correlation between content, design and a company's CGQ.

CONTENT

Table 1 summarizes the descriptive statistics for the corporate governance content features grouped by the two CGQ categories of companies. This classification provides a basis for comparing the content vis-à-vis the CGQ. A composite picture of the corporate governance content is the result.

Board

In this group of variables, only four companies provide information on the duties and responsibilities of the lead director position (13.3%). For the related variable of independent director information, only seven companies were found to display that information (23.3%). A majority of companies (53.3%) provide a direct link to contact directors/board. Further, 80% (n=24) provide director biographies, and an overwhelming number of companies, (n=28) 93.3%, display information on committees they serve. Only 11 companies (36.7%) give any information regarding director compensation. It's interesting to note that more companies among the higher CGQ group provided director compensation information than in the lower CGQ group.

Committees/Charters

All companies provide information about their audit committees, and an overwhelmingly majority devotes online space to their compensation (n=28) and governance (n=29) committees. However, correlated to lead director information under "Board," only two companies (6.71%) provide the charter of the lead independent director. Surprisingly, only 50% (n=15) give information on the all-important executive committee, and 20% (n=6) specify the financial experts on the audit committee board. Sixty percent (n=18) provide information on the nominating committee. Furthermore, at least two companies with CGQ > 75% gave the charter of the lead independent director. None of the companies with CGQ < 75% did so. Additionally, 75% of companies with higher CGQ provided information on the nominating committee compared to 42.9% of companies with lower CGQ. There is, then, scope for communicating additional committee/charter related governance information in this group of variables.

Compliance Policies

Overall, companies have a poor record for giving structured, visible information when it has to do with compliance. Only 13.3% (n=4) give insight into how to report accounting abuses, explain whistle blower policy & procedures or describe the role of an ombudsman, and all of these companies had CGQ > 75%. Only 23.3% of the companies (n=7) give information on the procedures to report accounting issues to the audit committee. Obviously, companies need to do a better job of providing appropriate information on compliance.

Ethics

The variables here are indicative of ethical issues related to corporate governance. While a good number of companies (n=26, 86.7%) have a code of business conduct and ethics, only 40% (n=12) declared political donations. Thirty-six point seven percent (n=11) indicated a code of business conduct/ ethics for the board and executives in particular. Companies might consider using their Web sites to make ethics policies more transparent.

Governance Documents

This important group of variables relates to a number of mandatory requirements of Sarbanes-Oxley, SEC, NYSE and Nasdaq having to do with the display and dissemination of important governance related documents. Seventy-three point three percent (n=22) of companies give the articles/certificate of incorporation /bylaws information while all companies do provide their corporate governance guidelines and principles. Additionally, 93.3% (n=28) give at least some information on securities transactions by directors and officers. However, insider transaction details are not given. This blurred view, along with the need to inform about stock option/ownership (only 6.7%, n=2, give this information), shows there is room for significant transparency, especially considering the recent buzz about stock options. It is also worthwhile to note that 37.5% of companies with CGQ > 75% give director/ executive stock option information while only 21.4% (n=3) of CGQ < 75% do so. Fifteen companies with high CGQ have some form of analyst/investor conference call, and 11 companies with low CGQ indicate this feature.

Management Teams

While 73% (n=22) provide biographies of the management team, only 20% (n=6) give information on executive compensation guidelines. Also 31.3% of companies with CGQ > 75% give this information while only 7% of companies with CGQ < 75% do so.

Press Room

All companies satisfy the basic requirement of providing press releases.

Proxy

While 93.3% (n=28) of companies have a link to the proxy statement, only 46.7% of the total set the proxy in an interactive format. Companies can use advanced Web technologies such Web 2.0 to disseminate proxy information more flexibly.

Social Responsibility

Lastly, a large number of companies (n=25) have a dedicated section on social responsibility which is tied to overall corporate governance.

Summary

Overall, with respect to content and with the exception of a few variables, there appears to be no significant differences between companies with high CGQ and companies with low CGQ. This may imply that companies with low CGQ have structural governance problems that cannot be addressed simply by displaying content on their Web pages. It is quite possible that a majority of companies provide basic content and follow basic benchmarks for what content to display on their Web pages. It is also conceivable that most use one or two Web application providers who use standardized templates. The data and analysis clearly show deficiencies and areas where companies can improve and move forward to the next level of communication and dissemination.

DESIGN

Table 2 summarizes the frequency counts and percentages for the design-related variables we identified. These are grouped by high and low CGQ companies. The data speaks for itself. A majority of companies have used basic Web design capabilities and features such as charting/analysis, multiple formats download, email alert sign up, software download support, search box and Web cast archives. But a wide range of useful tools, such as advanced Web technologies like blogs, pod casts, RSS feeds, and pop-up surveys, will do much to improve the ready availability of content on corporate sites.

Interactivity

The group of variables in this composite looks at features that make the corporate governance Web pages dynamic. While a majority of companies have charting/analysis capability at 86.7% (n=26), multiple download format (80%, n=24), e-mail alert sign up capability (83.3%, n=25) and software download support (80%, n=24), very few have blogging capability at 10% (n=3). Some have online request form feature (56.7%, n=17) and e-mail link from corporate governance page (53.3%, n=16). An effective feature more companies should consider is a pop-up online visitor satisfaction survey that assesses visitor satisfaction with corporate governance information. Only five companies with CGQ > 75% and only two with CGQ < 75% presently have this feature. Slightly more companies with high CGQ have charting capability and email alert features.

Location

This is an important variable indicating how easy (visible) or difficult (multiple levels/links) it is to access corporate governance information. Surprisingly, only 33% (n=10) of the companies have corporate governance links on their home page. For many companies, the visitor must dig down several levels (clicks) to access the information. Companies may consider elevating corporate governance to the main page to facilitate more effective dissemination.

Multimedia

In this important category, variables represent use of more advanced Web technologies. Again, this is an area companies should put to greater use more than they have been. A mere 33.3% (n=10) of companies utilized pod casting. Simultaneously, only 46.7% (n=14) of companies have RSS feeds and only 23.3% (n=7) of companies enable Web cast access without registration. Freely available Web casts clearly are less cumbersome. On the positive side, all companies employ a Web cast archive feature.

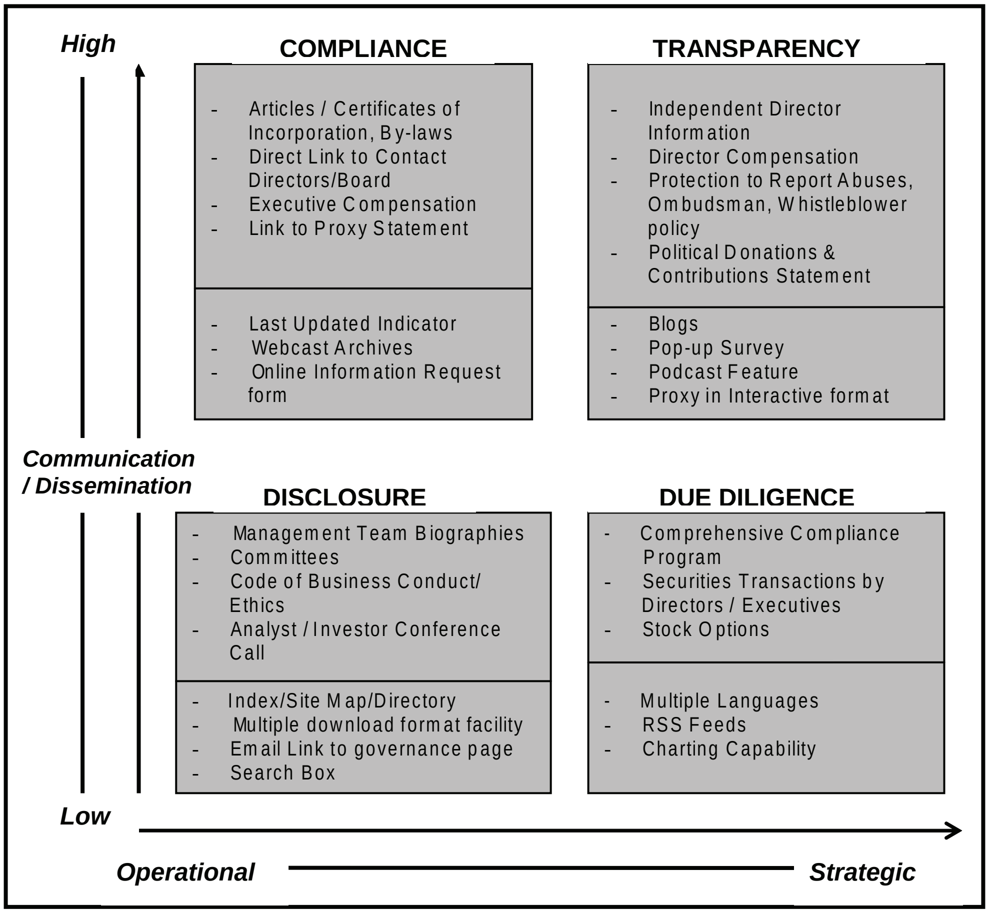
Navigation

Another critical category of design is navigation. Nearly all companies had a search box (n=29, 96.7%), while only 36.7% (n=11) of the companies had capability to view some of the corporate governance information such as ethics and governance principles in multiple languages. This was a puzzling discovery since many of the companies have a global presence. Also, only 46.7% of the companies (n=14) had a last-updated indicator. Because currency and timeliness are critical to corporate governance, companies might consider updating to include this feature. Lastly, 70% (n=21) companies had some kind of index/site map or directory. This important feature would help visitors guide through the vast amount of governance information. Slightly more companies with higher CGQ have these features compared to companies with lower CGQ.

FRAMEWORK FOR CORORATE GOVERNANCE AND WEB PAGES

While this research does not claim predictive capability, we developed a prescriptive framework based on analysis of the data. Boards of directors

Figure 2. A matrix framework for corporate governance and Web sites



and executives can use this framework to assess the present and future potential of their corporate Web sites to enable communication and dissemination of governance information and processes.

Figure 2 is a prescriptive framework of the role Web sites play in communicating and disseminating corporate governance information with the goal of greater transparency. Essentially, the X axis is a continuum from operational to strategic focus for the company while the Y axis is a continuum from low to high communication and dissemination. The resulting matrix has four quadrants. The lower left quadrant represents content and design features indicative of primarily mandatory disclosure items of corporate governance. This is the minimum functionality that a company's Web pages should possess. Primarily, static information is displayed via the links. Examples of content features include management team biographies, code of business conduct/ethics and others. Standard design features include index/site map/directory, multiple download format facility and a search box. As we progress up and the focus shifts to compliance with regulations, there's increased interactive communication and dissemination. Some of this increase is mandated, some is voluntary. It appears, based on our analysis above that most companies today would find themselves in one of these two quadrants. In the compliance quadrant, we see companies providing additional content information, such as articles/certificates of incorporation and by-laws, executive compensation, link to proxy statement and others. Design features are more advanced, with the use of Webcast archives, as well as the provision of an online information request form. A last-updated indicator feature, while simple, is a powerful feature indicating timeliness and recency. The activities in these quadrants are operational, structured and tangible and nowadays are considered routine. They are also mostly reactive (e.g., to regulation). As a company moves to the right on the continuum the goal is to enable sophisticated due diligence and achieve greater transparency. The company engages in outreach programs to connect with stakeholders proactively. Companies in the lower right quadrant are concerned with activities relating to scanning the environment, actively seeking information on corporate governance as well as contributing to industry benchmarks and standards. In terms of content, the company proactively seeks to disclose securities transactions by directors and executives as well as provide information on stock options. Important design features include a multiple languages option, RSS feeds and dynamic charting capabilities. This quadrant focused on performing effective due diligence.

For a company whose scope puts it in the upper right quadrant, the key strategic objective is greater transparency and perception of effective corporate governance. This indicates sensitivity to stakeholder needs and objectives. Strategic value is obtained by involving a stakeholder in the governance process. For example, a company can use interactive blogs and pod casting for ongoing conversation with shareholders. Additionally, proxy information is provided in an interactive format. The company lays out clearly whistleblower policy and anonymous reporting of malfeasance. For example, a Web-based anonymous reporting tool can be designed. Further, lead director information and director compensation can be disclosed.

By no means are the four quadrants boxed and isolated. Rather, they are connected and interrelated as companies evolve, moving from the bottom left to the top right. The implementation and effective use of Web pages is not a one-shot deal but a continually evolving and progressive initiative, as new requirements and advances in technology are assimilated (Pozen, 2006).

SCOPE AND LIMITATIONS

The focus of this paper is the study of Web disclosure practices of the 30 Dow Jones companies. We examined various governance-related content and design features found on the companies' Websites and compared the disclosure differences between the high and low CGQ groups. Specifically, we studied the online disclosures of corporate governance-related information. It is beyond the scope of this study, however, to explain the rationale behind a firm's disclosure practices. Nevertheless, we can suggest several insights that may explain why companies do not take advantage of Internet technology to make full and effective online disclosures.

First, concerns about the costs of compliance have been expressed in the popular press (Akhigbe & Martin, 2006; Roberts, 2004a, 2004b; Solomon & Bryan-Low, 2004). Compliance with SOX regulations requires significant, non-recurring investment costs "upfront." These include costs of building and maintaining governance Web sites. Ashbaugh et al. (1999) found that the average cost of Web site development by U.S. firms was 3.8% of the firms' total assets. Even those Web sites already in place require expertise and resources for periodic updates. According to Ghose (2006), a survey of 224 public companies by Financial Executive International (FEI) in July, 2004, found that the average cost of complying with section 404 was approximately $4 million, and that the average cost varied with firm size.

Second, disclosure of information may be costly because it weakens a firm's competitive position and reduces firm value. Verrecchia (1983,

1990) contended that firms do not fully disclose information when doing so entails proprietary costs. Bamber and Cheon (1998) provided evidence supporting this view. They demonstrated that companies with high proprietary information cost (i.e., those with few competitors) disclose less precise management earnings forecasts. Botosan and Stanford (2005) also found that firms withhold segment information when proprietary costs are high. Thus, it is unclear whether better corporate governance will enhance in every case the quality and frequency of information released by management, and/or that it will reduce information asymmetry (Kanagaretnam et al., 2007). Hence, although providing high quality financial reports was shown to reduce information asymmetries as well as the cost of capital, many firms decide against maximum reporting because of the proprietary nature of many disclosures (Habib, 2006). Chen and Jian (2005) discuss the 'lemon problem' of information disclosure as a result of conflicts of interest and asymmetry of information between internal management and external investors. Improperly-handled disclosure may fail the capital market mechanism (Akerlof, 1970). These issues warrant further research but are beyond the scope of this article.

Last but not least, while Web-based disclosure has its benefits, there are some disadvantages compared to old-fashioned hard copy corporate reporting. For one, there is the potential for information overload (Debreceny, Gray & Rahman, 2002). Also, the non-regulated nature of the Web can make controlling the context of online information difficult. Information found online can easily be taken out of context and used unfavorably by third parties. In turn, this lack of control can give rise to security and trust issues (Boritz & No, 2005). The existence of both advantages and disadvantages indicates that not all companies necessarily will benefit from Web-based disclosures.

In terms of limitations, the scope of our findings is limited by the size of our sample. A larger sample size may help generalize the results more effectively. We followed prior research in the development of the content and design templates. A survey of CIOs and other executives in charge of corporate governance may allow further insight into the role of IT in general and the Web in particular in enabling governance. This data will also enable more advanced statistical analysis. We examined the data on the corporate Web sites, and so the quality of our results is only as good as the presented data. In this exploratory work we looked at only classification variable, that is, CGQ. Consideration of such variables as corporate culture and IT acceptance in that culture, company sector (e.g., manufacturing or service), company size, and so on may have an impact on Web usage.

CONCLUSION AND FUTURE RESEARCH

Although our research is limited to the Dow 30 component companies, the sample is representative of large cap companies in the S&P 500 and Fortune 500. The findings have important implications for the use of corporate Web sites in the presentation and dissemination of corporate governance information. This impacts transparency in general.

Our analysis and results present a mixed picture. While companies in our sample use their Web sites to present corporate governance related information, they do so in the most basic (read: limited) way. In most cases, advanced content and design features are absent or found lacking. Companies have a long way to go to take full advantage of the Web and related technologies to effectively organize, present and disseminate corporate governance information. These technological enhancements would enable companies to be more transparent and thereby satisfy stakeholder needs for timely and quality corporate governance information.

Presently, the rules of the NYSE regarding Web site requirments are narrowly-focused minimum standards designed for all companies. As such they ignore some issues important to investors. They also do not provide guidance on *how* the information should be posted online. This leaves a lot of room for companies that are serious about being transparent to stand out from their peers by adopting best practices in how and what they publish online. Good disclosure of corporate governance principles reduces stock price volatility because stakeholders know their company is trying to do the right thing. Further studies with larger samples may reveal additional patterns. One might note, for example, vast cultural differences in a future study of multinational companies. Additional empirical work can provide insight into the relationship between effective communication (via the Web) and higher governance ratings. Another area for research is the examination of IT's role in enabling corporate governance by reducing the transaction and agency costs (Strebel, 2004). The use of the Web and other IT types in enabling corporate governance is in a primitive stage compared to other areas, but the use of the framework and templates developed in this article can accelerate their maturing process.

REFERENCES

Akerlof, G. (1970). The market for 'lemons:' quality uncertainty and the market mechanism. *Quarterly Journal of Economics, 90,* 629-650.

Akhigbe, A., & Martin, A. D. (2006). Valuation impact of Sarbanes-Oxley: Evidence from disclosure and governance within the financial services industry. *Journal of Banking & Finance, 30,* 989-1006.

Alavi, M., & Leidner, D.E. (2001). Review: knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly, 25*(1), 107-136.

Ashbaugh, H., Johnstone, K. M., & Warfield, T. D. (1999). Corporate reporting on the Internet. *Accounting Horizons*, *13*, 241-258.

Bamber, L. S., & Cheon, Y. S. (1998). Discretionary management earnings forecast disclosures: Antecedents and outcomes associated with forecast venue and forecast specificity choices, *Journal of Accounting Research*, *36*, 167-190.

Berardino, J. F. (2001). Corporate governance on the agenda, South China Morning Post, Markets, p. 2.

Bethel, J. E. (2007). Recent changes in disclosure regulation: description and evidence. *Journal of Corporate Finance*, *13*, 335-342.

Bhat, G., Hope, O-K., & Kang, T. (2006). Does corporate governance transparency affect the accuracy of analyst forecasts? *Accounting and Finance*, *46*, 715-732.

Blunn & Company. (2003). Standards of online corporate governance disclosure, April 24. http://www.blunnco.com

Boritz, J. E., & No, W. G. (2005). Security in XML-based financial reporting services on the Internet. *Journal of Accounting and Public Policy*, *24*, 11-35.

Botosan, C. A., & Stanford, M. (2005). Managers' motives to withhold segment disclosures and the effect of SFAS No. 131 on analysts' information environment. *The Accounting Review*, *80*, 751-771.

Bowen, D. (2007). A message to the world, FT Digital Business. *The Financial Times*, *March 28*, p.1.

Brooker, M. (2000). Corporate revamp stutters along. *South China Morning Post*, Review, p. 10.

Bushman, R. M., Piotroski, J. D., & Smith, A. J. (2004). What determines corporate transparency? *Journal of Accounting Research*, *42*, 207-252.

Callahan, C. M., Lee, C. M. C., & Yohn, T. L. (1997). Accounting information and bid-ask spreads, *Accounting Horizons*, *11*, 50-60.

Chen, Y-M., & Jian, J-Y. (2005). The impact of information disclosure and transparency rankings system (IDTRs) and corporate governance structure on interest cost of debt, National Yunlin University of Science and Technology, Tawain, R.O.C.

Craven, B. M., & Marston, C. L. (1999). Financial reporting on the internet by leading UK companies. *European Accounting Review*, *8*, 321-333.

Debreceny, R., & Gray, G. L. (2001a). The production and use of semantically rich accounting reports on the Internet: XML and XBRL. *International Journal of Accounting Information Systems*, *1*, 47-74.

Debreceny, R., Gray, G. L., & Mock, T. J. (2001b). Financial reporting web sites: What users want in terms of form and content. *The International Journal of Digital Accounting Research*, 1, 1-23.

Debreceny, R., Gray, G. L., & Rahman, A. (2002). The determinants of internet financial reporting. *Journal of Accounting and Public Policy*, *21*, 371-394.

Deller, D., Stubenrath, M., & Weber, C. (1999). A survey on the use of the internet for investor relations in the USA, UK and Germany. *European Accounting Review*, *8*, 351-364.

Diamond, D. (1985). Optimal release of information by firms. *Journal of Finance*, *40*, 1071-1094.

DiPiazza, Jr., S. A., & Eccles, R. G. (2002). *Building public trust*. John Wiley & Sons.

Ettredge, M., Richardson, V. J., & Scolz, S. (2001). The presentation of financial information at corporate web sites. *International Journal of Accounting Information Systems*, *1*, 149-168.

Ettredge, M., Richardson, V. J., & Scolz, S. (2002). Dissemination of information for investors at corporate Web sites. *Journal of Accounting and Public Policy*, *21*, 357-369.

Ghose, A. (2006). Information disclosure and regulatory compliance: economic issues and research directions. New York University.

Habib, A. (2006). Information risk and the cost of capital: review of the empirical literature. *Journal of Accounting Literature*, *25*, 127-168.

Healy, P. M., & Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital market: A review of the empirical disclosure literature. *Journal of Accounting and Economics*, *31*, 405-440.

Hermansen, D.R., Hill, M.C, & Ivancevich, D.M. (2000). Information technology-related activities of internal auditors. *Journal of Information Systems (Supplement)*, 39-53.

Institutional Shareholder Services. (2005). Good corporate governance is enabled by communication. February. http://www.issproxy.com/pdf/GoodCommunication.pdf

Jones, D. (2002). Communicating corporate governance via the Web. December. http://accounting.smartpros.com/x36294.xml

Jones, D. (2003a). Best practices for online governance disclosure. *The Corporate Board*, *March/April*, 21-25.

Jones, D. (2003b). Corporate governance Web page practices. *The Corporate Governance Advisor*, *May/June*, 26-27.

Kanagaretnam, K., Lobo, G. J., & Whalen, D. J. (2007). Does good corporate governance reduce information asymmetry around quarterly earnings announcements? *Journal of Accounting and Public Policy*, *26*, 497-522.

Leftwich, R. (1980). Market failure fallacies and accounting information. *Journal of Accounting and Economics*, *2*, 193-211.

Marston, C., & Polei, A. (2002, July 4-5). Corporate reporting on the Internet by German companies. In *6th Annual Conference on Financial Reporting and Business Communication*, Cardiff University, Cardiff.

Monks, R. A. G., & Minow, N. (2004). *Corporate Governance* (3rd ed). UK: Blackwell Publishing.

MZ Bulletin. (2006). *IR Global Rakings, 6*(23), June. http://www.irglobalrankings.com/mzsites/irgs/arquivos/MZB2006_Webvrs.pdf

Pirchegger, B., & Wagenhofer, A. (1999). Financial reporting on the internet: A survey of the homepages of Austrian companies. *European Accounting Review, 8*, 383-395.

Pozen, R. C. (2006). Investors' web. *The Wall Street Journal*, *June 12*, p. A12.

Prentice, R.A., Richardson, V.J., & Scholz, S. (1999). Corporate Web site disclosure and rule 101b-5: An empirical evaluation. *American Business Law Journal, 36*(4), 531-578.

PricewaterhouseCoopers. (2004). Integrity-driven performance. A new strategy for success through integrated governance, risk and compliance management, A white paper.

Roberts, D. (2004a). GE says it faces $30m bill for governance. *The Financial Times*, *April 29*, p. 15.

Roberts, D. (2004b). Businesses hit by escalating costs of implementing Sarbanes-Oxley measures. *The Financial Times*, *November 12*, p. 1.

Schultz, S. (2003). Driving shareholder value through investor communication. *Investor Relations Update*, *February*, 14-16.

Solomon, D., & Bryan-Low, C. (2004). Companies complain about cost of corporate-governance rules. *The Wall Street Journal*, *February 10*, p. 1.

South China Morning Post (SCMP) (2001). Markets, p. 12.

Strebel, P. (2004). The case for contingent governance. *MIT Sloan Management Review*, *Winte*r, 59-66.

The Economist Intelligence Unit. (2002). Corporate governance: The new strategic imperative, A white paper sponsored by KPMG International.

Verrecchia, R. (1983). Discretionary disclosure. *Journal of Accounting and Economics*, *5*, 179-194.

Verrecchia, R. (1990). Information quality and discretionary disclosure. *Journal of Accounting and Economics*, *12*, 365-380.

Verrecchia, R. (2001). Essays on disclosure, *Journal of Accounting and Economics*, *32*, 97-180.

Watts, R., & Zimmerman, J. (1986). Positive accounting theory. Englewood Cliffs, NJ: Prentice-Hall.

Xiao, J. Z., Yang, H., & Chow, C. W. (2004). The determinants and characteristics of voluntary Internet-based disclosures by listed Chinese companies. *Journal of Accounting and Public Policy*, *23*, 191-225.

This work was previously published in Information Resources Management Journal, Vol. 22, Issue 2, edited by M. Khosrow-Pour, pp. 50-68, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 8.12
Using Web Service Enhancements to Establish Trust Relationships with Privacy Protection:
(Extended and Invited from ICWS 2006 with id 47)

Zhengping Wu
University of Bridgeport, USA

Alfred C. Weaver
University of Virginia, USA

ABSTRACT

The lack of effective trust establishment mechanisms impedes the deployment of diverse trust models for web services. One issue is that collaborating organizations need mechanisms to bridge extant relationships among cooperating parties. We describe an indirect trust establishment mechanism to bridge and build new trust relationships from extant trust relationships with privacy protection. Another issue is that a trust establishment mechanism for web services must ensure privacy and owner control. Current web service technologies encourage a service requester to reveal all its private attributes in a pre-packaged credential to the service provider to fulfill the requirements for direct trust establishment. This may lead to privacy leakage. We propose a mechanism whereby the service requester discovers the service provider's requirements from a policy document, then formulates a trust primitive by selectively disclosing attributes in a pre-packaged credential to negotiate a trust relationship. Thus the service requester's privacy is preserved.

INTRODUCTION

Web services facilitate collaborations and interoperations between business partners, software agents, service providers and consumers, which promote loosely coupled and dynamic service-oriented architectures. But they do not address the business aspects of interactions such as security, access control, business partner selection, service level agreement monitoring, and auditing – the activities that build trust between a consumer and a provider of a web service and that will ultimately determine which services are used and which are not. The issues of trustworthiness are tightly bound in the minds of consumers. For example, a consumer would generally trust the Citibank online services to conduct online banking in a secure and responsible manner, because Citibank has a massive physical network of banks and has been in the financial market over one hundred years. On the other hand, a consumer probably wouldn't entrust his/her savings to a newly launched financial institution with no obvious connections to any legitimate business. The point here is that as web services begin to gain a foothold in electronic business, critical services will probably be limited to extensions of pre-existing business relationships with already trusted companies.

How can a business service provider engender new trust or transfer trust to a new consumer via an existing agent? And how can two companies establish a trust relationship in order to provide and consume business services or share information over web services? They have to negotiate in order to establish a conventional business trust relationship, and so they would almost certainly want to do the same for a trust relationship using web services. Web service standards UDDI (Atkinson, 2003), WSDL (Christensen, 2001), and SOAP (Box, 2000) say nothing about this. Thus, although it would be possible to find a web service to use just by examining a UDDI registry, it is unlikely to be used prior to investigation of its reputation and reliability. To establish a trust relationship, the consumer and the service provider require a negotiation process. The negotiation process needs to exchange trust-related information between the two parties. The parties can exchange private attributes to build the trust relationship directly, or they can use pre-established relationships to build a new one via a trusted third party. Exchange of private attributes may put the privacy of the consumer and the service provider at risk. For example, a hacker may pretend to be a consumer to access useful information from the service provider. More seriously, a hacker can pretend to be a service provider to gather private attributes from consumers for malicious usage. We need to reduce that risk. As an alternative, using pre-established relationships may be a more convenient and more secure way.

To provide additional functionalities for security, privacy and many other purposes, "web service enhancements" appear. Web service enhancements are a series of specifications describing security, privacy and other contexts applied to web services by several industrial practitioners. In this article we describe an indirect trust establishment mechanism using web service enhancements for bridging extant trust relationships to produce new trust relationships. Using an exchange of privileges obtained from a commonly trusted third party (who has established trust relationships with both parties) avoids disclosure of any private attributes. Meanwhile this mechanism still allows free negotiation and trust agreement selection between the involved parties when subjective judgments have to be made. We also propose a privacy protection mechanism that reveals the minimal number of attributes necessary to build the desired trust relationship for the web service environment. Using this mechanism, a set of attributes signed by the service requester's digital signature is associated with a trust primitive element. This element is used to negotiate a trust relationship. Any changes of policy requirements associated with this trust relationship are dynamically enforced using a trust group element. The Related Work section provides an overview of related work. The next section, Using Web Service Enhancements To Bridge Trust Relationships, describes how to use web service enhancements to bridge extant trust relationships. The Privacy Protection In Indirect Trust Establishment section introduces privacy protection in the proposed indirect trust establishment mechanism. The Using Web Service Enhancements To Protect Privacy For Direct Trust Establishment section describes a proposed parallel mechanism for privacy protection in direct trust establishment. The Prototype System section illustrates an implementation of these mechanisms in a context of healthcare applications. The Discussion section discusses the performance of the proposed mechanisms and a case study. The article concludes with contributions and future work.

RELATED WORK

Several types of trust establishment mechanisms have been described in the literature for service-oriented computing and cross-domain applications. The most basic way is to map the identity of the service requester (or one identity in the service requester's security domain) to one identity in the service provider security domain. Other extant trust management systems build trust relationships between different security domains using the requester's role as a basis for mapping. More recently, group-based mechanisms have been proposed to build trust relationships.

While "role" is an abstract concept, in a complex organizational setting such as a healthcare environment one might assign differing roles, and hence differing access permissions, to physicians, technicians, and patients. In Sandhu, Coyne, Feinstein, and Youman's (1996) article on role-based access control, role-based trust establishment process is implied by setting a mapping between local roles and roles within remote domains, which we call a role-to-role mapping. In Chadwick, Otenko and Ball's article (2003), the authors propose using X.509 certificates to manage trust relationships. This trust establishment process also employs a manual configuration of static role-to-role mappings to form a trust relationship before an actual access occurs. In the mechanism proposed by Freudenthal, Pesin, Port, Keenan and Karamcheti (2002), predefined trust relationships are used to complete the trust establishment process for dynamic cross-domain environments. The common approach of (Sandhu et al., 1996), (Chadwick et al., 2003), and (Freudenthal et al., 2002) is that the authors try to decide whether to grant access at run-time by deciding what permissions each requester has according to the assigned role and the predefined trust relationship associated with that role. The use of role as a basis for trust establishment creates a problem in domain-to-domain interactions, which is the potential misalignment of the precise definition of roles from one domain to another. One consequence could be that the domain whose users are requesting access might legitimately need to create special roles that map more precisely to the agreed intent of the requested operation. In that case, the users would be enabled by the policies of the domain they are accessing to take actions beyond those allowed them by their own domains' policies. A more general problem with roles is that their use generally does not conform to the principle of least privilege, as promoted by Schneider (Schneider, 2003), which limits the security privileges to actual needs.

Vandenwauver, Govaerts, and Vandewalle (1997) use group-based mechanisms to describe a collection of security requirements agreed to by the administrators for a group of domains. This group-based mechanism reduces the administrator's burden of creating an explicit policy to manage each trust relationship. The authors also assume that a recognized consortium of group members has created a trust group with predefined membership requirements. The service provider has to verify some non-identity attribute information about each service request. This trust group mechanism lacks the process of negotiation before building a trust relationship. Li, Mitchell, & Winsborough (2002) proposes an entire trust management framework that can group logically related objects so that permissions about them can be assigned in one operation. These logical groups are defined by an authority, not via a negotiation process. Both group-based mechanisms mentioned above require some superior authority to create or predefine group information, and thus the resulting trust establishment mechanisms are not fully dynamic; they may not keep pace with the changing policies and requirements of the service providers.

Using indirect mechanisms to bridge extant trust relationships is a convenient and efficient method to produce new trust relationships, but very few indirect trust establishment mechanisms have been designed for a web services environment. However, several types of indirect trust models and the corresponding trust establishment mechanisms have been proposed for service-oriented computing (Papazoglou & Georgakopoulos, 2003), which are the foundations of web services. The trust relationships involved in the interactions between web services are enabled by separate authorities issuing security tokens (A Joint White Paper, 2002), which certify the identities or other non-identity attributes for the requesters or providers of web services. There are two major types of trust models for these trust relationships. One type is the centralized model and the other type is the distributed model.

In centralized trust models, a common trusted intermediary, called the "Trust Authority," is used for establishing trust relationships between any two entities. However, it may be difficult to find an ideal central authority if the community of trustees is large and heterogeneous. If a trust authority is determined, token recipients are typically able to ascribe a sufficient level of trust to a security token because they can be confident of its origin. For example, they know and trust the authority that issued the token and can verify the token's origin through cryptographic means. It is through the existing trust they have in the third party security token issuer that they are able to derive indirect trust for the holder of a security token created by the same issuer. The Privacy Enhanced Mail (PEM) certification (Kent, 1993) assumes that everyone in the world trusts one ultimate authority to verify the identities of other certificate senders (an assumption we find unrealistic). The PEM model does not allow for multiple levels of trust within its certification hierarchy. Unlike PEM, the X.509 authentication framework and its variant for web services (Hallam-Baker, 2004) follows a multiple trust authority structure. It postulates that everyone will obtain certificates from an official certifying authority (CA). The CAs are organized into a global hierarchy of certifying authorities. All users within a "community of interest" have keys that have been signed by CAs with a common ancestor in this global hierarchy, which forms a tree structure.

In distributed trust models, a static or dynamic web of trust is woven with less structured logical interactions between networks (Dimmock, 2004) compared with the centralized trust models. It is assumed that trust is transitive under certain contexts, because trust-related information can be propagated through one or more chains of trusted intermediaries in the networks. In the Pretty Good Privacy (PGP) system (Zimmermann, 1995), an entity generates a public/private key pair. Each entity is responsible for acquiring the public key certificates needed and for assigning degrees of trust to their source. There is no common ancestor to act as the trust server for grouping the users within a community of interest. Instead, trust is propagated through chained structures formed by individual entities. When comparing X.509 with PGP, it has been pointed out that the most apparent difference is the architecture (Josang, 1996) (Ellison, 1999). X.509 has hierarchic structures for professional or government organizations with liabilities, whereas PGP has anarchic structure based on informal relationships and undefined roles. Like the logic-based systems described in (Rangan, 1992) (Abadi, 2003) (Pimlott & Kiselyov, 2006), there is no partial trust (degree of trust) in this kind of system; trust is either complete or nonexistent. In the solution proposed by Tarah and Huitema (1992), in the degrees of trust from different entities could have conflicts, and the final degree of trust needs to be composed from different trust values (Damiani, Vimercati, Samarati, & Viviani, 2005).

Yet none of the above frameworks or systems can accommodate all trust models. However, all the trust models co-exist in the real world and all the trust models are used in daily life. So a flexible framework to accommodate all these trust models is desirable. Other associated questions remain open, for example, the actual meaning of a trust value and how different trust values can be combined to yield a composite value. Users generally prefer to control their own private information. But none of the above mechanisms address privacy issues, which could lead to information leakage during the propagation of trust-related information in direct and indirect trust establishment processes. To solve these issues, we introduce an indirect trust establishment mechanism using a bridging protocol to augment direct negotiation and establishment of trust. We use an alternative method to supplement indirect trust establishment with lightweight negotiation to provide owner control in the process of bridging extant trust relationships. At the same time, to guarantee privacy protection, we use privileges granted by a common third party as a substitute for exchanging private attributes during the trust negotiation. Parallel in direct trust establishment, all the identity-based, role-based and group-based mechanisms do not address privacy issues, which could lead to superfluous information disclosure during the trust establishment process. We propose a privacy protection mechanism for trust establishment, which maintains the idea of negotiation but without inadvertent disclosure of unnecessary information.

USING WEB SERVICE ENHANCEMENTS TO BRIDGE TRUST RELATIONSHIPS

Web Service Enhancements

Web services use SOAP to exchange information over the network. SOAP is a lightweight method for exchanging structured information in a decentralized environment. XML (eXtensible Markup Language) is used in SOAP to define a flexible messaging framework that can exchange a message over a variety of underlying protocols. Although SOAP is the basic infrastructure for information exchange between web services, it does not provide any privacy and security mechanisms for the information exchanged. To provide privacy, security, trust, and other functionalities for web services, web service enhancements are proposed.

Among this set of enhancements, the web service security specification requires that an incoming access request message prove a set of claims such as name, public key, permission, capability, or an existing trust relationship to guarantee security. A web service indicates its requirements and other security related information in its policy document together with the privileges to be granted for the entities satisfying these requirements (e.g., a WS-Policy file). If an access request arrives without having the required proof of claims, the service provider ignores or rejects the request. These claims are contained in security tokens. A security token is a representation of security-related information conveyed within the format of a SOAP message (Bajaj, 2004). If an issuer cryptographically endorses a security token, the token is called a signed security token. A security token service (STS) is a web service that issues security tokens (Rangan, 1992). That is, it makes assertions based on evidence that it trusts to whomever trusts it. To communicate trust, a security token service requires proof, such as a security token or a set of security tokens, and issues a new security token with its own trust statement (note that for some security token formats this can be just a re-issuance or co-signature). Another important related service is the attribute service. An attribute service is a web service that maintains attribute information about entities within a security domain.

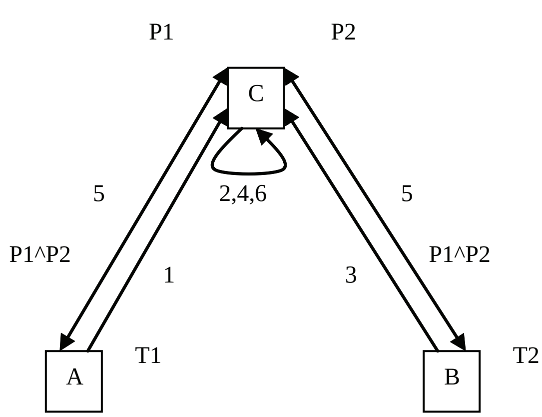
With these services one entity can rely upon a second entity to execute a set of actions or to make a set of assertions about a set of subjects or scopes (Anderson, 2004), which is called trust establishment. Trust relationships can be established by exchanging private attributes or bridging existing trust relationships; these techniques focus on owner control and utilizing extant trust relationships respectively. We propose a new indirect trust establishment mechanism to incorporate owner control into the process of bridging extant trust relationships.

A Bridging Protocol

The proposed indirect trust establishment mechanism uses a common third party as the anchor to bridge two extant trust relationships. The central part of this mechanism is a bridging protocol. An extant trust relationship is represented as a trust group element, which includes a trust relationship name (T), a list of participants involved in this relationship, and a list of privileges (P) granted for that relationship. This trust relationship can be established via an on-line trust negotiation or a contract written on paper. In the bridging protocol, the common third party needs to discover any difference in privileges granted to the two participants in order to provide the two participants equal standing and the opportunity to make their own subjective decisions for the new trust relationship. Figure 1 shows the workflow of the protocol to establish a new trust relationship between A and B using extant trust relationships between A and C, and B and C. The step numbers of the protocol correspond to the numbered arrows in Figure 1.

1. *A sends to C a request to establish a trust relationship with B, using an extant trust relationship (T1) with C.*
2. *After C receives A's request, C waits for a similar request from B. If a time limit expires, C sends a fail message to A and exits its waiting state.*
3. *B sends to C a request to establish a trust relationship with A, utilizing its trust relationship (T2) with C.*
4. *After receiving the requests from both A and B, C compares its granted privileges P1 and P2 for T1 and T2, and calculates the intersection of P1 and P2 (P1^P2).*

Figure 1. Workflow of the bridging protocol



5. *C sends P1^P2 to A and B, and asks whether they agree to build the new trust relationship based upon common privileges P1^P2.*
6. *If both A and B respond with a positive answer, C sends confirmations to both A and B. Both A and B produce corresponding new policies for the newly established trust relationship, and create a new trust group element to represent this trust relationship. If either A or B rejects the privileges represented by P1^P2, C asks A and B to establish a trust relationship directly by mutual identity verification or private information exchange. The new trust group element contains a name for the new relationship, a list of participants within this trust relationship (here A and B), and a trust level decided by each participant. So the copies of the trust group element kept by A and B contain the same name and participant list, but have their own trust levels associated with this trust relationship.*

This protocol resolves two problems introduced in the Introduction. First, it introduces a lightweight negotiation process into the indirect trust establishment (bridging extant trust relationships), which assumes that every participant has the right to make its own decisions. Second, it prevents privacy leakage by exchanging privileges granted by the common third party instead of exchanging private attributes.

The Common Third Party

How to find an appropriate common third party is also an issue. The participants A and B need to exchange partner information to find out which one is appropriate for a common third party. As with most companies, A and B's partner lists should be expected to represent private information. Companies are generally unwilling to reveal their business partner lists to companies who have not already established a trust relationship. To solve this problem, we propose a complimentary protocol below to find all common third parties.

1. A and B send requests to all their trusted partners respectively, to ask if they are willing to act as the common third party.
2. Only participants who are both A and B's trusted partner receive requests from both A and B. If a participant agrees to act as the common third party, it sends a message indicating its willingness to A and B.
3. From all the candidates who send back their willingness, A and B choose one as the common third party.

Either A or B can express its willingness to use one of the commonly trusted parties. The potential partner either agrees or starts a brief negotiation to establish the common third party before the trust establishment process with indirect privacy enhancement.

PRIVACY PROTECTION IN INDIRECT TRUST ESTABLISHMENT

Trust Establishment

Trust establishment is the initial step towards trust management over multiple security domains. Its importance has been addressed since 1994 (Beth, Borcherding, & Klein, 1994). To establish a trust relationship among multiple parties effectively and efficiently, a dynamic and flexible trust establishment mechanism must be developed (Reiter, 1999). Most existing trust management systems establish trust relationships via a trusted third party and they are based on public key certificates in which the trusted third party signs a specially formed message certifying the identity associated with a public key. The two best-known certificate systems are those of PGP and X.509. They attempt to solve part of the trust management problem of finding a suitably trustworthy copy of the public key of someone with whom one wants to communicate. This common trusted third party solution may result in a tree structure of all involved parties.

Some other existing systems try to weave a web of trust instead of a tree structure. In this case, the direct trust establishment mechanism becomes the most basic building block. The simplest way is to map one identity in a security domain to one identity in another security domain. Some other extant trust management systems also establish trust relationships with other domains using roles as a basis for mapping. And more recently, group-based mechanisms have been proposed to establish trust relationships. We will identify issues related to privacy protection in direct and indirect trust establishment processes, and provide solutions for privacy protection using web service enhancements in an interconnected and federated network environment.

Privacy Protection

Privacy protection technology consists of various tools for various applications, which includes cryptographically secured protocols for on-line critical information transmission, digital certificates, cookie management software, privacy policy languages, and so forth. Meanwhile, various anonymity protection techniques are being developed such as mix cascades and anonymous authentication for peer-to-peer networks. There are also privacy protection techniques proposed for data mining and database queries, as well as enterprise privacy auditing and enforcement tools for user privacy preferences. Privacy protection for trust establishment has particular relevance to the provision of owner control, pseudonymity or anonymity, and proof of knowledge.

- Owner control: Control over identity, credential, and private attribute information must be given to the owner, and users can modify or erase information if desired. Also, the security domain administrator does not endorse any disclosure of any piece of the user's private information. The choice of attributes to be disclosed is entirely under the owner's control.
- Pseudonyms and anonymity: The holder of the pseudonym can prove ownership by forming a digital signature using the corresponding private key. Such keys could be bound to attributes within digital certificates to form attribute certificates. One can also uses of a trusted third party to act as a mediator, vouching for the user or his/her computing device but removing any information identifying the user.
- *Proof of knowledge: Inspection of a user's credential can prove the holder's entitlement to the credential without revealing any persistent information.*

Active and Passive Models

For all the information exchange in trust establishment for federated networks, participating entities in that interconnected environment may be either passive or active. Passive entities wait to be contacted by active entities, while active entities are capable of initiating interactions or communications with passive entities or active entities. The distinction between active entities and passive entities allows for flexible deployment of entities for trust related activities over multiple security domains, such that some entities may only actively interact with a trusted set of partners, while others may passively listen for one or more interactions or service requests from strangers. With these different types of entities, several privacy control models are possible.

- Active model: From the point of view of active entities, the requirements of privacy control can be 'pushed' or 'advertised' first, and then potential partners can follow corresponding policies for subsequent trust related activities.

- Passive model: From the point of view of passive entities, private information and attributes are always kept as secrets; privacy control is applied to those kinds of information when that information has to be disclosed or released.
- *Hybrid model: Active and passive entities keep their roles while they work together to provide privacy enhancement in the interactions for trusted related activities.*

We will apply a hybrid model that takes advantage of both active entities and passive entities to enhance privacy protection for trust establishment using web service enhancements. Flexible but effective privacy protection is possible by designing and implementing a comprehensive framework that incorporates all these different models and facilities.

Privacy Protection in Indirect Trust Establishment

Owner control is embedded in our proposed bridging protocol for indirect trust establishment. Any participating parties can decide which subset of privileges granted by the common third party will be used as the foundation for new trust relationships. Anonymity and proof of knowledge are also achieved by subsets of privileges, because different subsets of privileges imply different levels of trustworthiness for different new relationships. In indirect trust establishment, A and B choose active model, and C (the common third party) chooses the passive model in order to keep A's and B's privileges as secrets until they are willing to disclose them as the foundations for new trust relationship. Privacy is also protected by using commonly granted privileges instead of private attributes to negotiate new relationship is the bridging protocol.

USING WEB SERVICE ENHANCEMENTS TO PROTECT PRIVACY FOR DIRECT TRUST ESTABLISHMENT

Since privacy protection is also a concern in direct trust establishment, which involves direct negotiation between participating parties, we propose another privacy protection mechanism for direct trust establishment. We introduce two new elements, trust primitive and trust group, to facilitate privacy protection in this process.

Trust Primitive and Owner Control

Definition 1: *A trust primitive is defined as the minimal subset of attributes in a pre-packaged digital credential, which has a complete semantic meaning according to a set of policy requirements. A trust primitive is signed by the credential holder, which is either an individual user or a security domain.*

In the proposed privacy protection of direct trust establishment for web service environments, a trust primitive is represented as a subset of attributes in the attribute service and conveyed as an XML element when it is exchanged between securities domains. As illustrated in figure 2, trust primitive 1 corresponds to an electronic library access rule with three required attributes (4, 6, and 7). The holder of the digital credential can form this trust primitive element and will be allowed to use the electronic library if the server can verify that its three attributes are valid (that is, if the token has been issued by an acceptable authority, the token is not expired, and the holder operates in the role of student, faculty or staff). Since neither the name nor ID number is part of the set of attributes associated with this trust primitive, anonymous access is permitted. Trust primitive 2 is for library checkout. It contains the same three attributes in trust primitive 1 plus attribute 2, ID number, which is needed for library accounting. Trust primitive 3, consisting of attributes 3, 6 and 7, might be used to verify same-sex gender before granting entrance to a dorm floor, or by substituting attributes 1 or 2 for attribute 3 it might be used to verify a specific individual's residence on the dorm floor before granting entrance.

A credential holder forms a trust primitive voluntarily. Every request of retrieving a trust primitive from an outside domain will be verified by the attribute service to selectively disclose the subset of attributes associated with the trust primitive. Another merit of trust primitives is that they prevent initiation of selective disclosure from anyone except the credential holder. Thus owner control is achieved. Figure 3 shows the workflow of the proposed protocol working with trust primitives for privacy protection in direct trust establishment.

In this protocol, a service requester and a service provider (two principals) from different security domains are assumed. Before the beginning of this privacy protection enhanced trust establishment workflow, the attributes of the service requester are stored at the attribute service of the service requester's security domain. The requester is also assumed to hold a security token containing its identity and other security related information. The workflow is initiated by a need for the requester to disclose some of its attributes to the service provider for negotiation.

Figure 2. Attributes and three possible trust primitives

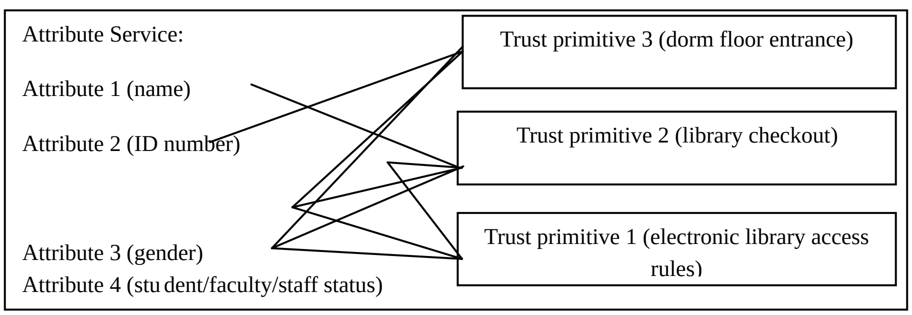
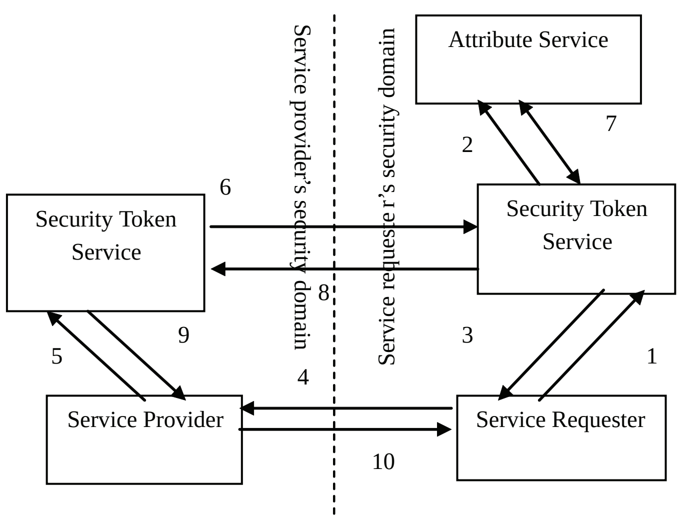


Figure 3. Workflow of privacy protection protocol for direct trust establishment



This need could be triggered by the access policy of the service provider's security domain, which is publicly accessible via a policy document, or by the service provider's direct request for the requester to provide one or more attributes. Thus a hybrid model is accommodated in this privacy protection protocol. Then the workflow executes from step one through step ten to finish a round of information exchange for trust establishment process. Step numbers below correspond to the numbered arrows in figure 3.

1. After reading the service provider's access policy or receiving a request for proving some attributes to obtain access, the requester initiates a service request, which contains the trust primitive corresponding to the policy or request for attributes from the service provider, to the STS in its own security domain. The security token held by the requester is also embedded in this service request message for proof of identity.
2. When the STS receives the service request, the STS extracts the security token from the message, verifies the legitimacy of the security token, and registers the trust primitive at the attribute service in its own security domain.
3. Then the STS adds this trust primitive to the requester's security token, re-signs the security token, and sends it back to the requester.
4. When the requester gets the newly signed security token, the requester embeds the security token in the access request and sends the request.
5. After receiving the requester's access request, the service provider asks its own STS to check whether the request complies with the service's access policy.
6. The STS of the service provider sends a request for attribute verification to the STS of the requester. The newly signed security token of the requester is sent together with the request.
7. The STS of the requester extracts the trust primitive, which corresponds to the access policy of the service provider, and uses this trust primitive as the query keyword to search the attributes disclosed by the requester at its attribute service.
8. The STS of the requester returns the attributes retrieved from the attribute service.
9. The STS of the service provider performs verification and sends its decision regarding the access request (granted/denied) to the service provider.
10. If access is granted, the service provider performs the requested operation and returns information to the requester; otherwise the provider issues a denial.

Sometimes the requester also needs to verify some of the service provider's attributes to negotiate a trust relationship, so the roles of the requester and the service provider can be interchanged. Several rounds of exchange form the negotiation needed to build a trust relationship between the two principals.

Trust Group and Partial Disclosure

Definition 2: *A trust group represents a group of partners who comply with the same set of policy requirements. The partners here are entities who have direct trust relationships with the policyholder.*

A trust group name is associated with a set of policy requirements. For example, if the service provider creates a set of policy requirements in the form of a WS-Policy file for negotiation of a trust relationship, the trust group name will be attached at the end of the file. Every partner complying with this set of policy requirements will use this trust group name to represent the corresponding trust relationship. A WS-Policy file containing a trust group looks like this.

```
<?xml version="1.0" encoding="utf-8" ?>
<policyDocument xmlns="http://schemas.microsoft.com/
    wse/2003/06/Policy">
<policies
xmlns:wssp=»http://schemas.xmlsoap.org/ws/2002/12/secext»
    xmlns:wsp=»http://schemas.xmlsoap.org/ws/2004/09/
    policy»>
<wsp:Policy wsu:Id=»trustlevelsec-token»>
<wssp:SecurityToken wsp:Usage=»wsp:Required»>
<wssp:TokenType>http://www.contoso.com/tokens/
    customXml#TrustLevelSecToken
</wssp:TokenType>
<wssp:TokenIssuer>http://www.cs.virginia.edu/TrustLev-
    elSTS.ashx</wssp:TokenIssuer>
</wssp:SecurityToken>
<wssp:SecurityToken wsp:Usage=»wsp:Required»></
    wssp:SecurityToken>
</wsp:Policy>
</policies>
<trustGroup>TG001</trustGroup>
</policyDocument>
```

After negotiation, a new trust group element is added to the requester's security token, and the security token is signed by the requester's STS again. Alternatively, the security token is replaced by a new security token issued by the service provider's STS, which contains the corresponding trust group element representing the new trust relationship. A trust group element is represented by three XML tags (trust group name and two domain/individual identities). Every policy holder also needs to record all the trust group elements that the holder is involved, and so do the partners. The trust group element looks like this.

```
<trustGroup>
<trustGroupName>TG001</trustGroupName>
<domain1>http://abc.com/localSTS.asmx</domain1>
<domain2>http://def.com/localSTS.asmx</domain2>
</trustGroup>
```

Access requests after the negotiation are granted by a verification of the trust group element in security tokens. If the access requirement for a trust group element changes in a service provider's policy, then the service provider needs to revalidate the previous trust relationship by invalidating the old trust group element, asking for the requester's trust primitives again, and then verifying whether the new trust primitives meet the requirements of the changed policy. If access is granted, a new trust group element will replace the old one for future use.

PROTOTYPE SYSTEM

Implementation of Indirect Trust Establishment with Privacy Protection

We have implemented the whole trust establishment system using the Microsoft .Net platform. The .Net framework together with the Web Service Enhancement (WSE) 2.0 SDK (Microsoft, 2006), which supports the WS-Security (Nadalin, 2004) and WS-Trust standard (Anderson, 2004), provides a complete platform for our system implementation. All the building blocks in our system architecture use web services as internal interfaces. Figure 4 illustrates the system architecture for indirect trust establishment with privacy protection. With the help of the proposed bridging protocol, trust domains can establish new trust relationships by extending current trust boundaries more smoothly and more securely. In this system architecture, the STSs are the main portals for interactions across trust domains. The STS is in charge of issuing and exchanging security tokens, which contain critical information such as identities, privileges and trust-related information. A policy repository is used to store and retrieve policy requirements and the corresponding privilege information used by the bridging protocol. The negotiation engine controls the whole process of information exchange and negotiation step by step. Each trust domain has an entire deployment of this system. But according to the bridging protocol and indirect trust establishment process, the functionalities used by participants A, B and the common third party C are different.

In indirect trust establishment mechanism, we use the trust group element to represent an

Figure 4. The architecture of the indirect trust establishment mechanism

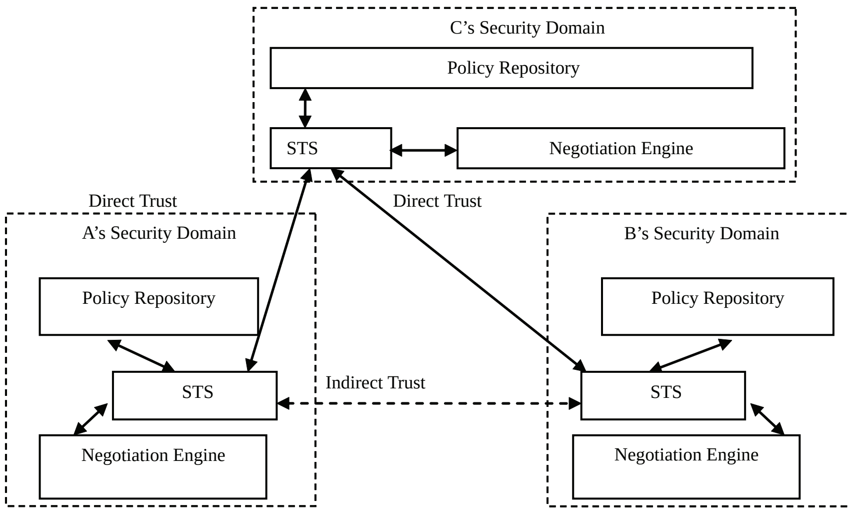


Figure 5. A trust group element in a security token and in the policy repository

```
<trustGroup>
<trustGroupName
>TG001
</trustGroupName
>
<domain1>
http://www.cs.virg
inia.edu/localSTS.
asmx
</domain1>
<domain2>
http://www.ee.virg
inia.edu/localSTS.
asmx
</domain2>
</trustGroup>
```

In a Security Token

```
<?xml version="1.0" encoding="utf-8" ?>
<policyDocument
xmlns="http://schemas.microsoft.com/wse/2003/06/Policy">
<policies
mlns:wssp="http://schemas.xmlsoap.org/ws/2002/12/secext"
xmlns:wsp="http://schemas.xmlsoap.org/ws/2004/09/policy">
<wsp:Policy wsu:Id="trustlevelsec-token">
<wssp:SecurityToken wsp:Usage="wsp:Required">
<wssp:TokenType>
http://www.contoso.com/tokens/customXml#TrustLevelSecToken
</wssp:TokenType>
<wssp:TokenIssuer>
http://www.cs.virginia.edu/TrustLevelSTS.ashx
</wssp:TokenIssuer>
</wssp:SecurityToken>
<wssp:SecurityToken wsp:Usage="wsp:Required">
</wssp:SecurityToken>
</wsp:Policy>
</policies>
<trustGroup>
TG001
</trustGroup>
<privileges>
read&write
</privileges>
</policyDocument>
```

In the Policy Repository

established trust relationship. In a security token, the trust group element contains a name for the relationship together with information from the two domains between which the trust relationship is established. In the policy repository, every trust group element is associated with a set of privileges granted by another party. Figure 5 illustrates the different formats of a trust group element in a security token and in the policy repository.

Implementation of Privacy Protection for Direct Trust Establishment

Meanwhile, this privacy protection enhanced trust establishment system is a subsystem of a federated cyber trust system (Weaver, Dwyer, Snyder, Van Dyke, Hu, Chen, & Mulholland, 2003). It provides the functionality of negotiating, building, and managing trust for a web service environment. When the federated cyber trust system is applied to a healthcare environment with hospital, pharmacy, insurance and billing security domains, the shaded boxes shown in Figure 6 are the modules involved in the trust establishment system. Boxes with dashed border lines represent different security domains; boxes with solid border lines represent modules in the hospital domain; arrows represent information flows or interactions. All the modules use web services as interfaces for their interactions.

We also have a graphic user interface to assist users to define and sign their trust primitives for privacy protection, which is shown in Figure 7.

Another concern is token formats. Among all the interactions between modules, there are two different types of interactions. One is the interaction between modules within a security domain. The other is the interaction between security domains. Two types of security tokens are used with different purposes to convey identities, credentials, trust group elements and other trust related information between service modules. We use username tokens to facilitate the interactions between service modules within the same security domain since it has little overhead and is easy to extend; we use SAML (Cantor, 2005) tokens to enable interactions between security domains since SAML is a recognized standard for interoperability among different platforms. Figure 8 gives examples of two token formats before being embedding into SOAP messages.

The architecture of our privacy protection enhanced direct trust establishment is illustrated

Figure 6. Federated cyber trust system for a healthcare application

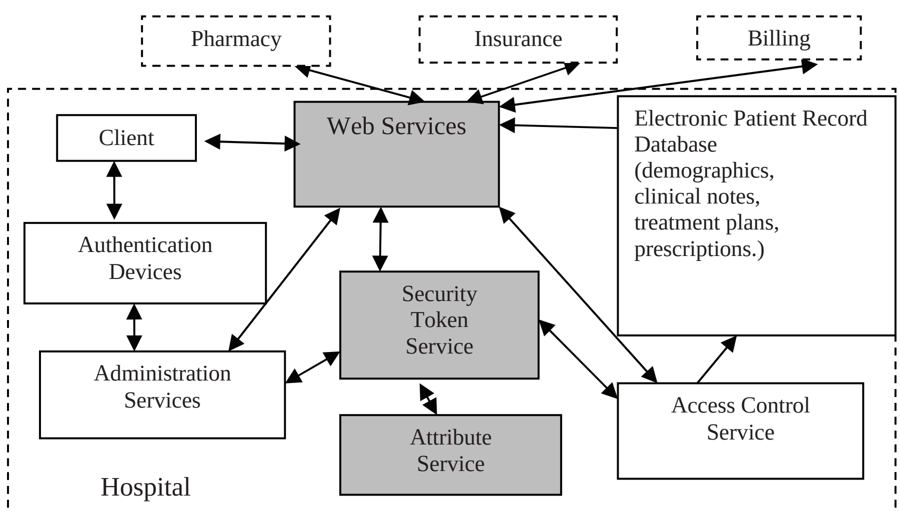


Figure 7. Graphic user interface for trust primitive definition

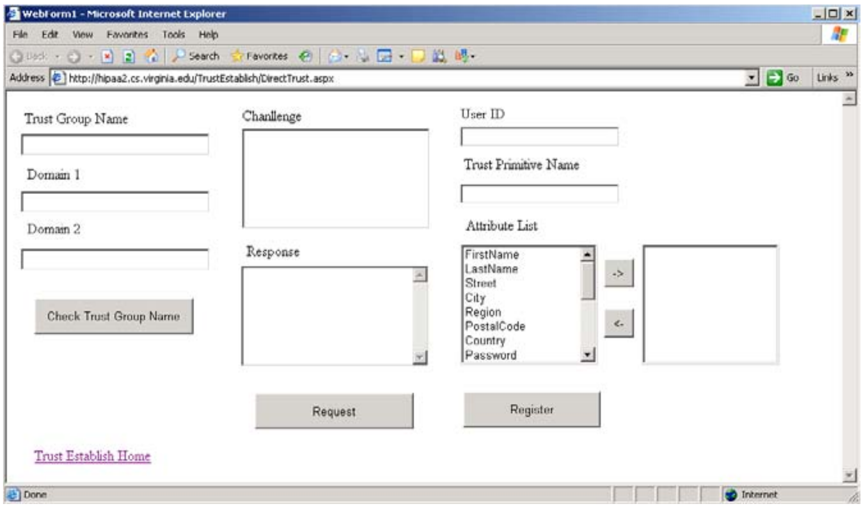


Figure 8. Token formats

```
<UsernameToken>
<CreateAt>04/20/2005 8:00:00 am
</CreateAt>
<ExpireAt>04/20/2005 5:00:00 pm
</ExpireAt>
<UserID>123</UserID>
<TokenIssuer>http://www.cs.virginia.edu/Trust
STS.asmx
</TokenIssuer>
<TrustGroup>
Name="TG001"
Domain1="http://www.cs.virginia.edu"
Domain2= "http://www.ee.virginia.edu"
</TrustGroup>
<TrustPrimitive>TP001</TrustPrimitive>
</UsernameToken>
```

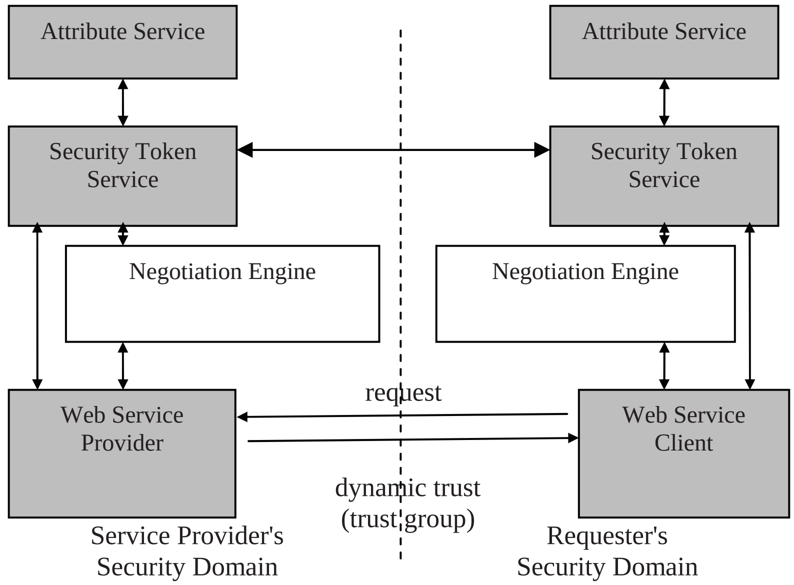
UserName Token

```
<saml:Assertion Version="2.0" ID ="ABC"
IssueInstant="timestamp">
<saml:Issuer>
http://abc.com/TrustSTS.asmx
</saml:Issuer>
<saml:Conditions
NotBefore="07/20/2006 8:00:00 am"
NotOnOrAfter="07/20/2006 5:00:00 pm">
<saml:Subject>
<saml:NameID>123</saml:NameID>
</saml:Subject>
<saml:Attribute>
Name="TrustGroup"
NameFormat="urn:oasis:names:tc:SAML:2.0:a
ttrname-format:basic"
<samlAttributeValue>
TG001
</samlAttributeValue>
Domain1="http://www.cs.virginia.edu"
Domain2="http://www.ee.virginia.edu"
</saml:Attribute>
</saml:Assertion>
```

SAML Token

in figure 9. The web service requester and web service provider build a dynamic trust relationship via negotiation engines and security token services in both security domains. The security token service in our implementation also includes a set of web services to interpret and exchange security tokens. At the same time, the security token service uses an attribute service to register trust primitives for the requester. Negotiation engines control the overall workflow to build trust relationships dynamically, which include the implementation of the protocol described in

Figure 9. The architecture of privacy protection enhanced direct trust establishment



the Using Web Service Enhancements To Protect Privacy For Direct Trust Establishment section for a single round of negotiation.

DISCUSSION

A Case Study for Healthcare Applications

As we know, paper-based operations are still dominant in healthcare industry, because trust relationships in healthcare applications are still based on paper contracts and certificates. And third-party-issued paper certificates can still lead to privacy leakage. With new legislations such as HIPAA and Sarbanes-Oxley, strict privacy protection is imposed on healthcare applications. We here describes a detailed case study to illustrate how the proposed mechanisms and architecture can be applied to real healthcare applications to establish trust relationships with privacy protection using web services. Figure 10 illustrates the real workflows to build a new trust relationship when a patient needs to fill a prescription at a new pharmacy using web services, and to bridge a trust relationship to another neighborhood pharmacy when the prescription cannot be filled at that pharmacy using existing between the neighborhood pharmacy and the hospital. First, a patient at the hospital tries to fill a prescription at a new pharmacy (pharmacy A). Then the direct trust establishment protocol is applied to allow the hospital domain verifying the required and the only required attributes such as pharmacy license from the new pharmacy to construct the new trust relationship for prescription filling. Then a part of the prescription cannot be filled by the new pharmacy. It forwards the partial prescription-filling request to a neighborhood pharmacy (pharmacy B). Assuming the neighborhood pharmacy has an existing trust relationship with the hospital, the two pharmacies can bridge a new trust relationship using the hospital as the anchor for partial prescription-filling request. Using our proposed direct and indirect trust establishment protocols can not only fulfill the required trust relationship constructions but also provide privacy protections in the trust relationship establishment processes.

Figure 10. A case study with interactions between a hospital and two pharmacies

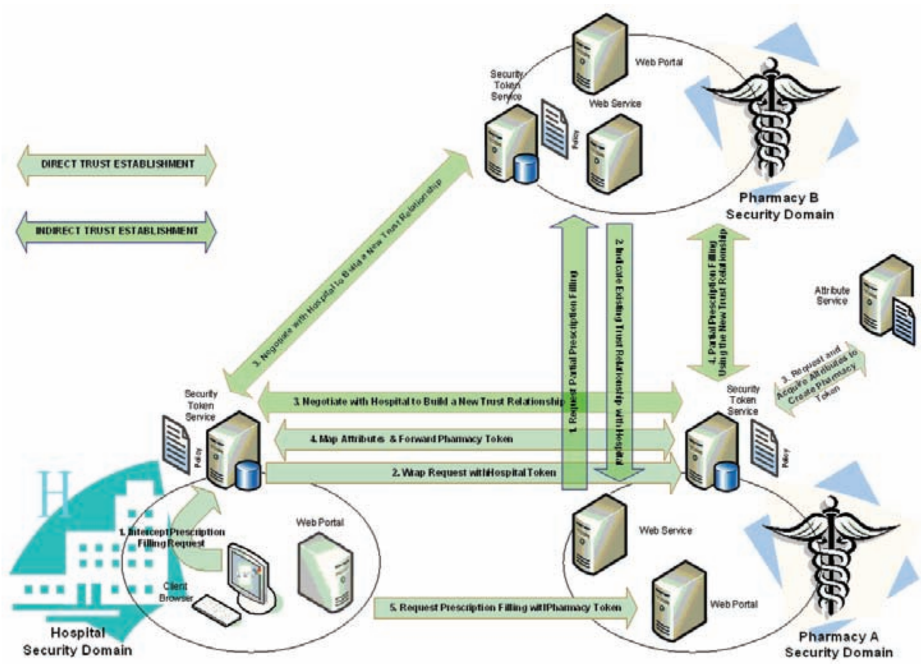
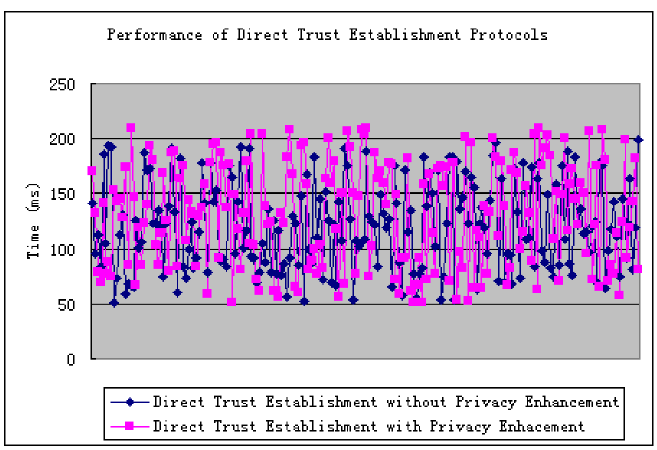


Figure 11. Comparison of direct trust establishment protocols' performance (with and without privacy enhancement)



Performance Evaluation

For federated co-operations, management of trust needs to handle privacy protections within these co-operations. We achieve privacy protection for trust establishment via two protocols using web service enhancements. The proposed privacy protection mechanisms do not change the structure of other daily operations, such as user enrollment, user authentication, and verification. And the privacy-enhanced indirect trust establishment is achieved via a whole new protocol design using existing trust relationships, which avoids direct sensitive information exchanges. The privacy-enhanced direct trust establishment only enhances a few additional steps in federated co-operations using available features in WS-Policy, security tokens, and attribute services. These mechanisms can also be easily embedded into other trust establishment protocols with certain modifications.

We compared the performance of the direct trust establishment without privacy enhancement, which shows all the credentials available in attributes upon request, and the proposed privacy-enhanced trust establishment protocol described in the section Trust Primitive and Owner Control. We randomly choose trust establishment processes conducted by patients and hospital staff when new trust relationships need to be built with a new domain. We measured the time used for two trust establishment processes (see Figure 11). The blue (black) series is the direct trust establishment with privacy enhancement; the red (grey) series is without privacy enhancement. Although the difference of the time used for a single direct trust establishment process varies from 50ms to 198ms, the mean value of the direct trust establishment time with privacy enhancement is 128.08ms and the mean value of the direct trust establishment process without privacy enhancement is 123.51ms (a 3.70% difference). We found the overhead in performance is almost unnoticeable.

CONCLUSION

In this article we described an indirect trust establishment augmented with lightweight negotiation to achieve owner control and privacy protection simultaneously. Our research motivation comes from the inadequacy of binary trust and the difficulty of using subjective formulas combining multiple (even conflicting) trust values. Our new privacy protection mechanism in indirect trust establishment is an alternative method with these advantages:

- It introduces the negotiation process into the indirect trust establishment (bridging extant trust relationships), which assures that every participant has owner control over the decision-making process for new trust relationships.
- It prevents privacy leakage by exchanging privileges granted by the common third party instead of exchanging private attributes.

Also we described a privacy protection mechanism for direct trust establishment that extends the extant trust establishment mechanisms for web services to gain many advantages from its privacy control and dynamic capabilities. Our research motivation comes from the complicated privacy requirements inherent to current healthcare data management and similar sensitive information management. Our new trust establishment mechanism is dynamic with these advantages:

- It allows only the requester to choose what attributes may be viewed by the service provider. Therefore, it is capable of enforcing the requester's privacy.
- It allows only the chosen attributes to be viewed by the service provider. Therefore, it is capable of disclosing private attributes selectively.
- It allows any trust relationships to be renewed whenever the service provider's policy is updated. Therefore, it is inherently dynamic.

Our future research will focus on the topological impact of privacy protection mechanisms and their applications to privilege delegation and enforcement of trust relationships.

ACKNOWLEDGMENT

This research effort was funded by Microsoft Research.

REFERENCES

A Joint White Paper from IBM Corporation and Microsoft Corporation (2002), Security in a Web Services World: A Proposed Architecture and Roadmap. Retrieved on XXX from, http://msdn.microsoft.com/library/en-us/dnwssecur/html/securitywhitepaper.asp

Abadi, M. (2003). Logic in Access Control, *Proc. of the 18th Annual IEEE Symposium on Logic in Computer Science*, pp. 228-233.

Anderson, S., et al. (2004). Web Services Trust Language (WS-Trust), Retrieved on XXX from, http://msdn.microsoft.com/ws/2004/04/ws-trust/.

Atkinson, B. et al. (2003), UDDI Spec Technical Committee Specification. Retrieved on XXX from, http://uddi.org/pubs/uddi-v3.0.1-20031014.htm.

Bajaj, S., et al. (2004), Web Services Policy Framework (WS-Policy). Retrieved on XXX from, http://specs.xmlsoap.org/ws/2004/09/policy/ws-policy.pdf.

Beth, T., Borcherding, M., & Klein, B. (1994). Valuation of Trust in Open Networks, *Proc. of the 3rd European Symposium on Research in Computer Security*, pp. 3-18.

Box, D.,et al. (2000). Simple Object Access Protocol (SOAP) 1.1. Retrieved on XXX from, http://www.w3.org/TR/2000/NOTE-SOAP-20000508/

Cantor, S., et al. (2005). Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0. Retrieved on XXX from, http://docs.oasis-open.org/security/saml/v2.0/saml-core-2.0-os.pdf.

Chadwick, D., Otenko, A., & Ball, E. (2003). Implementing Role Based Access Controls Using X.509 Attribute Certificates, *IEEE Internet Computing*, pp. 62-69.

Christensen, E., et al. (2001). Web Services Description Language (WSDL) 1.1. Retrieved on XXX from, http://www.w3.org/TR/wsdl.

Damiani, E., Vimercati, S., Samarati, P., & Viviani, M. (2005). A WOWA-based Aggregation Technique on Trust Values Connected to Metadata, *Proc. of International Workshop on Security and Trust Management*, pp. 131-142.

Dimmock, N. (2004). Using Trust and Risk in Role-Based Access Control Policies, *Proc. 2004 Symposium on Access Control Models and Technologies*, pp. 156-162.

Ellison, C., et al. (1999). SPKI Certificate Theory, Internet Report, RFC 2693.

Freudenthal, E., Pesin, T., Port, L., Keenan, E., & Karamcheti, V. (2002). dRBAC: Distributed Role-based Access Control for Dynamic Coalition Environments, *Proc. of 22nd International Conference on Distributed Computing Systems*, pp. 411-420.

Hallam-Baker, P., et al. (2004). Web Services Security X.509 Certificate Token Profile. Retrieved on XXX from, http://docs.oasis-open.org/wss/2004/01/oasis-200401-wss-x509-token-profile-1.0.pdf.

Josang, A. (1996). The right type of trust for distributed systems, *Proc. 1996 New Security Paradigms Workshops*, pp. 119-131.

Kent, S. (1993). Privacy Enhancement for Internet Electronic Mail: Part II: Certificate-Based Key Management, Internet Report, RFC 1422.

Li, N., Mitchell, J., & Winsborough, W. (2002). Design of a Role-Based Trust-Management Framework, *Proc. of the 2002 IEEE Symposium on Security and Privacy*, pp. 114-130.

Microsoft (2006). Major Features of the Web Services Enhancements. Retrieved on XXX from, http://msdn.microsoft.com/library/en-us/wse/html/3c55ed70-6b66-4620-bba7-59c5b2f2191d.asp.

Nadalin, A., et al. (2004). Web Services Security: SOAP Message Security 1.0 (WS-Security 2004). Retrieved on XXX from, http://docs.oasis-open.org/wss/2004/01/oasis-200401-wss-soap-message-security-1.0.pdf.

Papazoglou, M. & Georgakopoulos, D. (2003). Service-oriented Computing, *Communications of the ACM*, 46(10), pp. 25-28.

Pimlott, A. & Kiselyov, O. (2006). Soutei, a Logic-Based Trust-Management System, *Proc. of 8th International Symposium on Functional and Logic Programming*, pp. 130-145.

Rangan, P. (1992). An Axiomatic Theory of Trust in Secure Communication Protocols, *Computers and Security*, 11, pp. 163 – 172.

Reiter, M. (1999). Authentication Metric Analysis and Design, *ACM Trans. Information and System Security*, 2(2), pp. 138-158.

Sandhu, R., Coyne, E., Feinstein, H., & Youman, C. (1996). Role-based access control models, *IEEE Computer*, 20(2), pp. 38-47.

Schneider, F. (2003). Least Privilege and More, *IEEE Security and Privacy*, 1(3), pp. 55-59.

Tarah, A. & Huitema, C. (1992). Associating Metrics to Certification Paths, *Proc. 2nd European Symposium on Research in Computer Security*, pp. 175-189.

Vandenwauver, M., Govaerts, R., & Vandewalle, J. (1997). Role based access control in distributed systems, *Communications and Multimedia Security*, 3, pp. 169-177.

Weaver, A., Dwyer, S., Snyder, A., Van Dyke, J., Hu, J., Chen, X., & Mulholland, T. (2003). Federated, Secure Trust Networks for Distributed Healthcare IT Services, *Proc. of IEEE International Conference on Industrial Informatics*, pp. 162-169.

Zimmermann, P. (1995). The Official PGP User's Guide, MIT Press.

This work was previously published in International Journal of Web Services Research, Vol. 6, Issue 1, edited by L.-J. Zhang, pp. 49-68, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 8.13
The Interactive Computing of Web Knowledge Flow: From Web to Knowledge Web

Xiangfeng Luo
Shanghai University, P.R. China

Jie Yu
Shanghai University, P.R. China

ABSTRACT

Web Knowledge Flow provides a technique and theoretical support for the effective discovery of knowledge innovation, intelligent browsing, personalized recommendation, cooperative team work, and the semantic analysis of resources on Internet, which is a key issue of Web services and Knowledge Grid/Web(Zhuge, 2007; Zhuge, 2005). In this chapter, first the authors introduce some basic concepts related to Web Knowledge Flow. Next they illustrate the concepts of interactive computing, including the Web interaction model, the implementation of interactive computing and the generation of Web Knowledge Flow. Finally, the applications of Web Knowledge Flow will be given.

1. WEB KNOWLEDGE FLOW

1.1 Concept of Web Knowledge Flow

Many efforts have been done on the knowledge flow area. Some researchers study knowledge flow based on the organization of workflow, which is about the knowledge demand of the logic relationship and the role between the tasks of workflow. It can be realized by the way of pushing (Zhao, 2001; Wolverton, 1997). Taxonomy model aims at providing an overall picture of grid workflow verification and validation (Chen, 2007). Chen and Yang develop a novel checkpoint selection strategy that can adaptively select not only necessary but also sufficient checkpoints (Chen, 2007). Spiral model proposed by Nonaka describes the knowledge flow from epistemology to ontology. In the epistemology, knowledge flows from implicit knowledge to explicit knowledge, then from explicit knowledge to implicit knowledge. In the ontology, knowledge flows from person to group, and then from group to person (Nonaka, 1994; Nonaka, 1995). Based on Spiral model, Knowledge Flow Dynamic Model (*KFDM*) proposed by Nissen makes the knowledge flow over time explicitly. It can support a multi-dimensional representation that enables a new approach to analyze and visualize diverse knowledge flow patterns in enterprises (Nissen, 2002). Knowledge energy model proposed by Zhuge et al. takes knowledge energy as the driving cause to form an autonomous knowledge flow and explores the hidden principles (Zhuge,2005). The principles of knowledge flow engaged in cooperative cognition are explored by Dou from the perspective of learning and cognition evolution (Dou, 2006). Textual knowledge flow proposed by Luo et al. aims to provide an effective technique tool and theoretical support analysis for the discovery and cooperation of knowledge innovation, intelligent browsing, and personalized recommendation in Web services and e-Science Knowledge Grid (Zhuge, 2002). Other knowledge flow models are peer-to-peer team knowledge sharing and management based model (Luo, 2008), agent based model (Nissen, 2004), and the trust based model (Guo, 2005), etc.

Definition 1(Web Knowledge Flow, WKF):

Web Knowledge Flow(WKF) is a sequential link between topics with rich semantics, which is activated by user's demands and changes with the demands.

Compared with other multiple types of knowledge flow (Nissen, 2002; Zhuge, 2005), WKF has some special characteristics as follows.

- WKF reflects the flow of knowledge between topics on Internet;
- WKF contains rich semantics between topics, which leads to similar WKF, associated WKF, and causal WKF etc;
- WKF is activated by user's demands, in other words, WKF is a kind of service on demands;
- WKF changes with the change of user's demands.

When a user browses topics, a WKF is a browsing path of topics recommended to the user. When some browsing paths of topics are activated at the same time, one of them should be chosen and recommended according to user's demand.

As can be seen, Web knowledge flow provides a technique and theoretical support for the effective discovery of knowledge innovation, intelligent browsing, personalized recommendation, cooperative team work, and the semantic analysis of resources on Internet, which is a key issue of Web services and Knowledge Grid.

According to different relationship between nodes in Web Knowledge Flow, WKF can be classified into Association Web Knowledge Flow (Luo, 2008) and Similarity Web Knowledge Flow (Luo, 2008).

Definition 2:(Association Web Knowledge Flow, *AWKF*)

Associated knowledge flow is a sequential link with rich semantics between associated topics, which is activated by user's demands and changes with the demands.

Compared with other knowledge flows, *AWKF* has the following distinguished characteristics.

- *AWKF* can reflect the flow of knowledge between topics;
- *AWKF* can contain associated relation between topics;
- *AWKF* can be activated by users' demands, i.e., *AWKF* is a kind of service on-demand;
- *AWKF* can updates with users' demands.

When a certain user browses topics in the Web or an e-Science environment, *AWKF* is a browsing path of topics. When several browsing paths of topics are activated at the same time, they should be evaluated according to the user's demand and their contents, and then one of them should be chosen and recommended to the user.

Definition 3:(Similarity Web Knowledge Flow, *SWKF*)

Similar knowledge flow is a sequential link whose only difference from *AWKF* is that the relationship between topics in SKF is similarity relationship.

Therefore, if the user wants to browse similar topics on the Internet, an SWKF is helpful.

We know that web resources have two features in common:

- **Various Types of Representation:** Web resources exist in many kinds of representation: video, audio, text, and so on. First, it is impossible that the huge amounts of web resources are represented in one form. Second, different users have different requirement, which make the variety necessary. Even for one user, he/she must have different requirements in different time;
- **Out-of-order Organization Statement:** Web resources are distributed in out-of-order statement, which generates some challenges to the user's web activity, such as web search. First, out-of-order statement makes resource searching cost lots of time, which affects the search quality. Secondly, out-of-order statement can not lead to high accuracy of search.

On the other hand, requirement of users is not limited to one form of media. For example, the user may want to browse the web resources whose contents meet his requirement, whatever its representation form is.

Therefore, in order to facilitate high qualified web activity it is indispensable to do the following work:

- Unify the web resources into one representation;
- Organize all the web resources in order;

These are the basis for the generation of WKF. Element Fuzzy Cognitive Map(E-FCM) is used to represent web resource, whatever it is a video, audio, or text file. Semantic Link Network(SLN) is used to organize the web resources in order by introducing certain relationship. Next sections will include the contents of these two aspects.

1.2 Representation of Web Resource

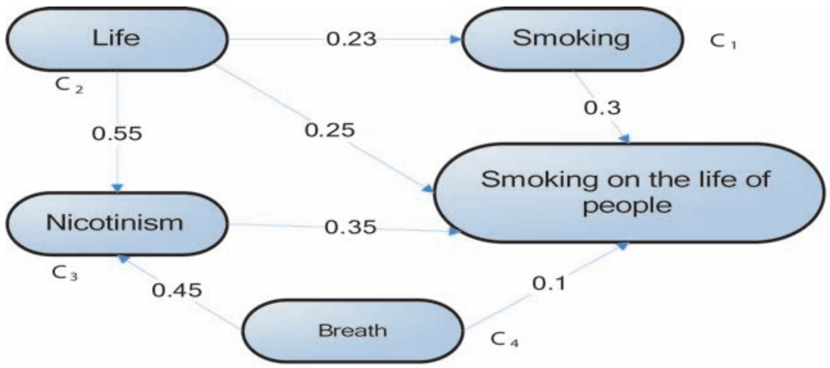
Element Fuzzy Cognitive Maps (E-FCMs) have a good capability to represent Web resource with rich semantics and can be understood by machine easily (Kardaras, 2006; Perusich, 2006). Therefore, E-FCM is proposed to represent Web resources.

Term 1:(Element Fuzzy Cognitive Map, E-FCM) Element Fuzzy Cognitive Map (E-FCM) is a fuzzy cognitive map, whose element concepts are represented by keywords; state values of element concepts are computed by the function of its frequency, position and font size in paragraphs or a section; theme concept is represented by the implied semantics of co-occurrence keywords appearing in a topic; the relations between concepts and their weights are represented by the relations between keywords and topic as well as their weights.

Figure 1 gives an example of E-FCM generated by the algorithm in (Luo, 2008), the vector of element concepts is {Smoking, Life, Nicotinism, Breath}

The knowledge of topic is represented by E-FCMs, which reflects the keywords' relations and their weights as well as the state values of co-occurrence keywords in a topic. Semantic information of co-occurrence keywords expressed by E-FCM is richer than a set of separate keywords

Figure 1. "Smoking on the life of people" represented by E-FCM (denoted as Topic 1)



because E-FCM stores topic information instead of the separate keywords of topic.

According to the capability of tagging topics, concepts can be classified into three categories (Luo, 2008): general concept, functional concept and seldom concept. Different type of concept has different contribution to the computing of web knowledge flow.

Term 2: (General Concept, *GC*) General concept is an element concept of E-FCM, which is a ubiquitous element concept appearing in an E-FCMs/topics library.

General concept has low capability of tagging topics for it has general semantics in an E-FCMs/ topics library. For example, in the domain of semantic Web, the keyword "semantic" has no specific function to tag topics because it is a ubiquitous concept in the topics library.

Term 3: (Functional Concept, *FC*) Functional concept is an element concept of E-FCM, which has specific semantics in an E-FCMs/ topics library.

Functional concept has high capability to tag topics for it has specific semantics in a domain. For example, in the domain of semantic Web, the keyword "OWL", "WSDL" have a specific function to tag topics because these keywords have particular meanings in the presented topics library.

Term 4: (Seldom Concept, *SC*) Seldom concept is an element concept of E-FCM, which has a low frequency in a domain E-FCMs /topics library.

Seldom concept has low capability to tag topics for it has a low frequency in a domain. For example, in the domain of semantic Web, the keyword "Knowledge Grid" has low frequency in the presented topics library.

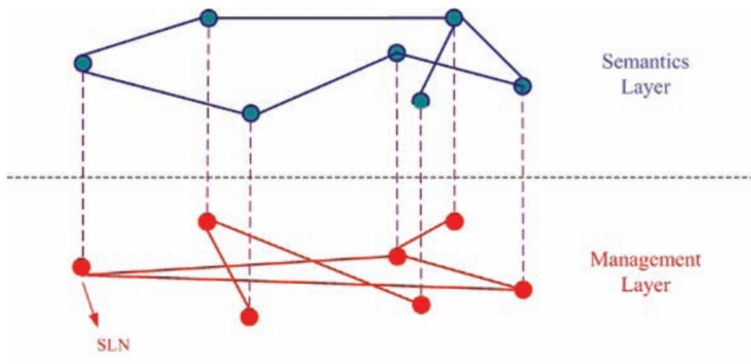
The semantic value of a seldom concept is bigger than a general concept. So when calculating semantic similarity degrees between topics, the common seldom concepts will bring bigger influences on the semantic similarity degrees than the common general concepts.

1.3 Ordered Organization of Web Resources

Semantic Link Network (SLN) (Zhuge, 2007) can link not only textual topics but also multimedia topics (e.g. relation of topics between text and video). Therefore SLN is introduced to link these Web resources based on the discovery of associated/similar topics, which can guide users' intelligent browsing of topics.

Term 5: (Semantic Link Network, *SLN*) A semantic link network consists of semantic nodes and semantic links (relations) between nodes. A semantic node can be a semantic community, a schema, a concept, a feature, an entity or an identity (Zhuge, 2007).

Figure 2. 2-layer Hierarchy Organization of Web Resources



The semantic link network (*SLN*) is designed to establish semantic relationships among various resources (data, image and various documents) aiming at extending the hyperlink network World Wide Web to a semantic-rich network (Zhuge, 2007).

Herein, the semantic nodes are the topics (e.g. multimedia topics or textual topics) and the semantic relations are the associated relations in a semantic link network. The knowledge flow based on *SLN* guides users to browse the cross media information[1] that makes the Web browsing more vivid, diversiform, and visual.

The web resources are organized in a 2-layer hierarchy way which is illustrated in Figure 2. The first layer is called semantics layer that the resources are organized by content and relationship with each other. The second layer is called management layer which is based on P2P structure (Dejan, 2002). Each node in these two layers is a SLN and represented by feature set. When a user's requirement is presented, first with semantics computing technique, the node which matches the requirement is located in semantics layer. Then mapping to management layer, the physical position this resource is got by P2P computing. Finally, similarity and association relationship computing is used to get the E-FCM which represents the specific topic and recommended to the user. After analyzing user's activities, analysis result is used in finding the node which has some relationship with user's preference. Then again, the node is mapped to management layer to get the physical position. In this way, a WKF is generated.

As can be seen, in this hierarchy structure semantics layer is used for high accuracy of web activity. Nodes on this layer are organized by semantics relationship which makes finding specific topic more accurately. Management layer is used for high efficiency of web activity. It has the special use of managing and locating resources. With P2P technique, the physical position can be got efficiently. In addition, by dividing all the web resources into several SLN, this hierarchy can deal with the great number of web resources well.

2. INTERACTIVE COMPUTING OF WEB KNOWLEDGE FLOW

When a user is browsing on the Internet, the topics he wants to read have some common characteristics in most cases. Maybe they all limited to one domain, or focus on one event. In other words, the topics have some relationship with each other. These topics that have the same relationship with each other are the nodes of WKF. How to find these topics according to a user's requirement is an important task. The key technique is interactive computing. It is a method which can get quantified information by analyzing user's requirement and behavior.

The key issues of generation of the WKF are the discovery of certain relations and the construction of the semantic links between topics. The discovery methods of association rules include Apriori (Agrawal, 1993), sequential pattern mining (Agrawal, 1995); episodes mining (Mannila, 1997); space association rules mining (Koperski, 1995); ring based association rules mining (Ozden, 1998); negative association rules mining (Savasere, 1998); association rules mining between affairs (Savasere, 1998); and calendar shopping basket analysis (Lu, 1998), etc. This section will introduce how to get the nodes of WKF and generate WKF by interactive computing.

2.1 Web Interaction Model

We know that Interaction Machine (Wegaer, 1997; Wegner, 1998; Goldin, 1998) is the extension of Turing Machine (Goldin, 1999).

Term 6: (Interaction Machines, IM) IMs extend TMs by adding dynamic input/output (read/write) actions that interact directly with an external environment (Wegner, 1998).

Interaction machines may have single or multiple input streams and synchronous or asynchronous communication, and can differ along many other dimensions, but all IMs are open systems that express dynamic external behavior beyond that computable by algorithms.

Term 7: (Interaction Histories): Observable behavior of IMs is specified by interaction histories.

As can be seen, the process of Web exploration can be modeled by interaction machine. The prediction of the user's next behavior is not only related to the user's present behavior and the history of his/her exploration. The behavior is represented by the semantics of the topic he/she browses. To describe the user's web activity and build the web knowledge flow, Web Interaction Model is given.

Definition 4: (Web Interaction Model, WIM) Web Interaction Model is used to build Web Knowledge Flow. It is a triad $M = (S, I, F)$, in which

--S is an enumerable set of states which describe the user's browsing state. Browsing state reflects the topics that the user have browsed and the interest that the user is interested in;

--I is an enumerable set of input states. Input state describes the user's behavior in browsing.

--$F : S \times I \to S \times O$ is a computable function which is called semantic computing of Web Knowledge Flow.

In WIM, the same input states may correspond to different output states, because of the history of browsing. With WIM, we can predict the user's interest and recommend corresponding pages to him/her.

The properties of WIM are as following:

- **Dynamic binding of inputs:** the input state may depend on the previous output. The user may react differently to different recommended topics which affect the input state;
- **Semantics dependence:** interactive computing is processing on the semantics layer of the organization of web resources. How to get the physical position of resources is the duty of management layer. In the process of computing, the semantics of topics are taken into account which makes the recommendation of topics based on content;
- **History dependence:** output can depend on previous history of browsing. The history of browsing is the basis for interactive computing;
- **Hidden information:** The user's interest is hidden and can be attained by semantics interactive computing with browsing history.

2.2 Implementation of Interactive Computing[2]

Semantics interactive computing is the key essential in WIM. It performs the task of semantics computing, and can be implemented by Markov Chain.

Assumption 1: Assuming that the process of user's browsing is in accord with homogeneous discrete Markov Chain. Therefore, the user's browsing process can be represented by a sequence which is composed of the web pages that he/she browsed.

This Markov Chain is represented by a triad $C=(X,P,\lambda)$, in which

- *X* is a discrete random variable. It is represented by an E-FCM which represents the topic user has browsed.
- *P* is the transition probability matrix. Each element p_{ij} of this matrix is the probability that topic_i transits to topic_j, i.e. after a user browses topic_i the probability that he/she browses topic_j.

$$P=\begin{vmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2n} \\ & & \cdots & & \\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn} \end{vmatrix},$$

  n denotes the number of the total topics.
- λ is the initial state distribution of the topics. $\lambda=(p_1,p_2,\dots,p_n)$

In this Markov Chain, how to decide *P* is a key task. Before discussing the implementation of this model, some structures used here will be introduced.

Definition 5: History Vector, VH(t)

For each user, we use a vector $VH(t)=\left(h_1,h_2,\dots,h_n\right)$ to represent the history of his/her browsing history till the moment *t*. If the user browses the *ith* topic, then the corresponding element value v_i will be added by 1. The initial the value of *V* is $VH(t)=\left(0,0,\dots,0\right)$.

Definition 6: Visiting State Vector, (VS) Visiting State Vector $VS=\left(vs_1,vs_2,\dots,vs_n\right)$ is a vector which labels the topics that has been visited. If topic_i have been visited, then the vs_i should be set 1.

Definition 7: User State Vector, US(t) User state vector $US(t)=(us_1,us_2,\dots,us_n)$ is a vector which indicates the user's state which is related to the content of the topics that he/she browses. When a user browses topic_i, the topics which are similar to topic_i may also be interests for the user. Therefore, the user state *US(t)* should be computed by the following formula:

$$us_i=\frac{\sum_{j=1}^{n} v_j\cdot sdm_{ij}}{\sqrt{\sum_{j=1}^{n} v_j^2}\cdot\sqrt{\sum_{j=1}^{n} sdm_{ij}^2}}$$

On the other hand, the next topic the user may browse is related to not only other users' browsing history, but also the semantic relationship. The topic that has association relationship with this topic might be the one the user wants to browse next time. Therefore, the association relationship between topics should be taken into account.

$$P=\alpha\cdot ADM+\beta\cdot S$$

$$s_{ij}=\frac{a_{ij}}{\sum_{j=1}^{n} a_{ij}}$$

α and β are the weight of *ADM* and *S*, and $\alpha+\beta=1$. $A=(a)_{ij}$ denotes the number of visiting pair *i-j* in all the users' visiting history. Therefore, for each user, the user's state in the next moment *NS(t)* can be computed by the following formula:

$$NS(t)=US(t-1)\times A$$

If all the topics have not been visited, then the element which has the biggest value In *NS(t)* will be the topic that the user will probably interested in. Obviously, the user doesn't want to visit the topic that he/she has already visited again. In this paper, an AND operation is used to filter this kind of topics by the following formula:

$$NS(t)' = NS(t) \quad and \quad VS$$

Therefore, the biggest element in *NP(t)'* may correspond to the topic which will be the one that the user will be most interested in. And in the application, this topic can be recommended to the user.

As can be seen, in the process of semantics interactive computing, not only all the user visiting histories but also semantics relationship between topics are related to the state of topic. In other words, the content of the topics is taken into account in this paper which improves the accuracy of state description.

2.3 Generation of Web Knowledge Flow

Based on the discussion above, the main steps of construction of Web Knowledge Flow should be the following:

1. According to the user's requirement, generate its corresponding feature vector;
2. With similarity relationship, get the first node of WKF;
3. According to the user's behavior, update the history vector and visiting state vector.
4. With similarity SLN, attain the state of the user which is represented by *US(t)*;
5. With association SLN and the visiting history of all the users, attain the transition probability matrix *P*;
6. With visiting state *VS*, get the new *NS(t)'*, and take the element which has the biggest value as the next node in WKF, and update the visiting history vector by $h_i = h_i + 1$;
7. If all the value in *NS(t)'* are less than the threshold, then stop and the WKF is generated. Otherwise, go to 5).

In this way, the Web Knowledge Flow is generated. As can be seen that due to the semantics relationship including similarity and association relationship between topics are involved in the process of generation, the generated WKF can well reflect the user's interest and meet his/her requirement.

3. APPLICATIONS

In order to acquire many new customers and keep the existing ones, Internet-base business provides more objective information and better service. This section will show how to apply WKF to one important field, Web personalization.

With the help of user's browsing activities, interactive computing can be used to recommend some topics that the user may be interested to the user. For example, a user's browsing activities are traced, and some information are shown in table 1.

Then the corresponding history vector *HV(t)* is generated: $hv_{83}=hv_{79}=hv_{88}=hv_{228}=hv_{339}=hv_{443}=hv_{506}=hv_{542}=hv_{575}=hv_{704}=hv_{708}=hv_{913}=hv_{916}=hv_{925}=hv_{968}=1$; $\mathrm{hv}_{847}=\mathrm{hv}_{898}=2$.

We also generate the SLN with similarity relationship and SLN with association relationship of 1000 Web pages which belong to the field of environment. The corresponding SLN are shown in figure 3 and figure 4, respectively.

Then the corresponding WKF which is composed of five nodes(shown in figure 5) is generated. From the user's browsing history, we can see that the user interested in the topics about climate change. And the nodes of generated WKF are also focus on this topic. Therefore, WKF can well grasp the user's interest and recommend corresponding

Table 1. Records of Browsing Activity

Topic	Times of Browsing
Asia will bear brunt of climate change-linked deaths WHO	1
Aussies Kyoto should have been ratified	1
Australia, New Zealand to cooperate on climate change	1
Canada blasted by own environmental watchdog	1
Environmentalist pleads not guilty in OR	1
EU to adopt climate fight plan despite differences	1
FACTBOX What is the Kyoto Protocol	1
France's Besson plans film focusing on environment	1
Hewlett Foundation plans climate change grants	1
House bill lets California restrict car emissions	1
Nations seek compromise in climate change talks	1
Minorities the forgotten victims of climate change	2
Lawyers say considering court challenge to Korea	2
Nature and man jointly cook Arctic	1
New Australian PM seals Kyoto ratification at climate meet	1
Obama, Clinton top McCain on environment votes report	1
At E.U. summit, climate change billed as major security risk	1

Figure 3. SLN with Similarity Relationship

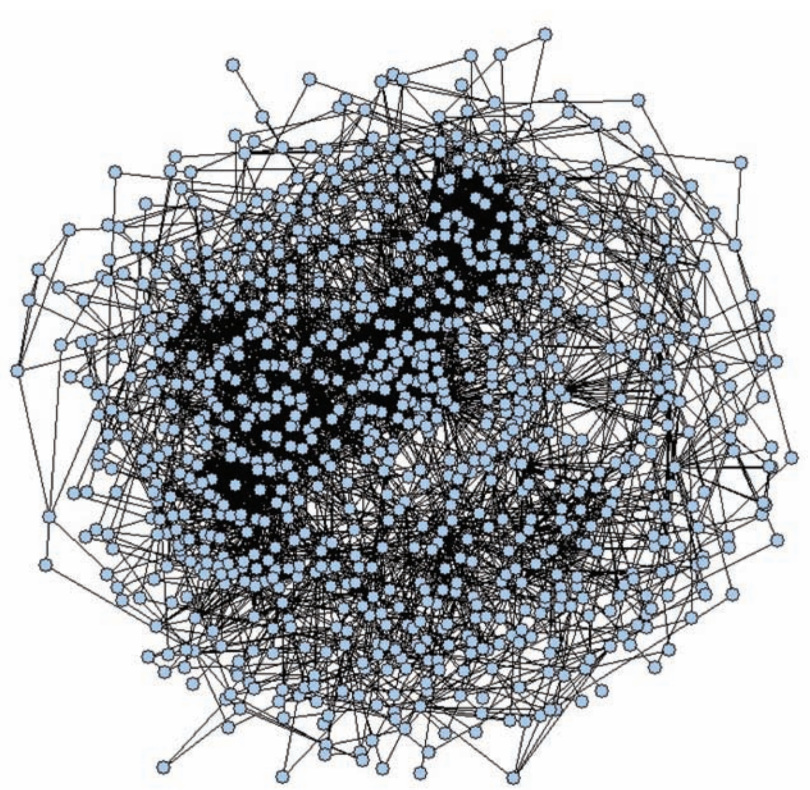


Figure 4. SLN with Association Relationship

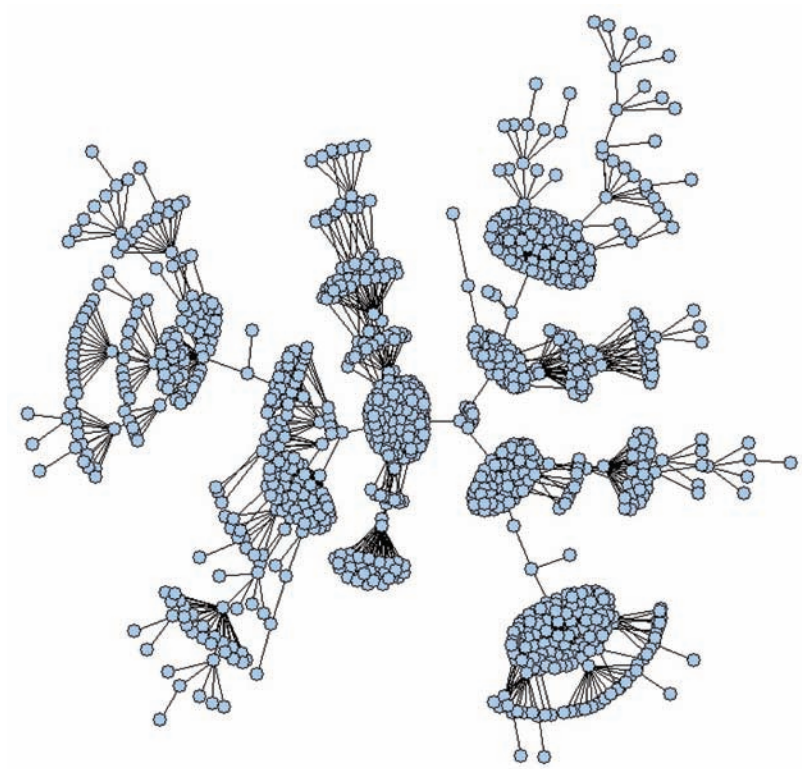
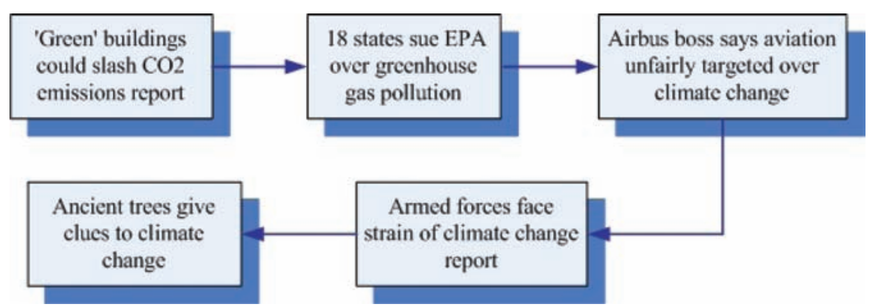


Figure 5. Generated WKF



pages to him/her. In addition, it can be seen that the nodes in WKF are dispersed in SLN with similarity and SLN with association relationship. It accords with the dynamics features of WKF. With this generated WKF, the user can get his/her interested topics and doesn't need to search them by himself/herself on the Internet.

REFERENCES

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (pp. 207-216). ACM Press.

Agrawal, R., & Srikant, R. (1995). Mining sequential Patterns. *Proceedings of the 11th International Conference on Data Engineering* (pp. 3-14). IEEE Computer Society Press.

Ashok, S., Omiecinski, E., & Shamkant, N. (1998). Mining for Strong Negative Associations in a Large Database of Customer Transactions. *In Proceedings of the 14th International Conference on Data Engineering* (pp. 494-502). IEEE Computer Society Press.

Chen, J., & Yang, Y. (2007). Adaptive Selection of Necessary and Sufficient Checkpoints for Dynamic Verification of Temporal Constraints in Grid Workflow Systems. *ACM Transactions on Autonomous and Adaptive Systems*, *2*(2), Article No. 6.

Chen, J., & Yang, Y. (2008). A Taxonomy of Grid Workflow Verification and Validation. *Concurrency and Computation*, *20*(4), 347–360. doi:10.1002/cpe.1220

Dejan, M. S., Kalogeraki, V., & Lukose, R. (2002). *Peer-to-Peer Computing* (Tech. Rep. HPL-2002-57). HP Lab.

Dou, W., Chen, G., Cheung, S. C., & Cai, S. (2006). Cooperative Cognition and Its Implementation under Web Environment. *Proceedings of the Second International Conference on Semantics, Knowledge and Grid* (pp. 46). Washington, DC: IEEE Computer Society.

Goldin, D., & Wegner, P. (1998). *Persistence as a form of interaction* (Tech Rep: CS-98-07). Brown University.

Goldin, D., & Wegner, P. (1999). *Behavior and expressiveness of persistent Turing machines* (Tech Rep: CS-99-14). Brown University.

Guo, W. Shi, X., Cao, L., & Yang, K. (2005). Trust in Knowledge Flow Networks. *Proceedings of International Conference on Semantics, Knowledge and Grid* (pp. 405 -412).

Kardaras, D., & Karakostas, B. (2006). E-Service adaptation using fuzzy cognitive maps. *The 3rd International IEEE Conference on Intelligent Systems* (pp. 227-230).

Koperski, K., & Han, J. (1995). Discovery of Spatial Association Rules in Geographic Information Databases. *Proceedings of the 4th International Symposium on Large Spatial Databases* (pp. 47-66), Springer Publisher.

Lu, H., Han, J., & Feng, L. (1998). Stock movement prediction and n-dimensional inter-transaction association rules. *Proceedings of the 3rd ACM-SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery* (pp. 121-127). ACM Press.

Luo, X., & Fang, N. (2008). Semantic Representation of Scientific Documents for the e-Science Knowledge Grid. *Concurrency and Computation*, *20*(7), 839–862. doi:10.1002/cpe.1271

Luo, X., & Hu, Q. (2008). Similar Knowledge Flow based Intelligent Browsing of Topics. *In Proceedings of 3rd International Workshop on Workflow Management and Application in Grid Environments.*

Luo, X., Hu, Q., & Xu, W. (2008). Discovery of Textual Knowledge Flow based on the Management of Knowledge Maps. *Concurrency and Computation*, ▪▪▪, 1791–1806. doi:10.1002/cpe.1319

Luo, X., Xu, Z., & Yu, J. (2008). Discovery of Associated Topics for the Intelligent Browsing. *The First IEEE International Conference on Ubimedia Computing* (pp. 119-125). Digital Object Identifier: 10.1109/UMEDIA.2008.4570876.

Mannila, H., Toivonen, H., & Verkamo, A. I. (1997). Discovery of Frequent Episodes in Event Sequences. *Data Mining and Knowledge Discovery*, (3): 259–289. doi:10.1023/A:1009748302351

Nissen, M. E. (2002). An Extended Model of Knowledge Flow Dynamics. *Communications of the Association for Information Systems*, *8*, 251–266.

Nissen, M. E., & Levitt, R. E. (2004). Agent-Based Modeling of Knowledge Flows: Illustration from the Domain of Information Systems Design. *Proceedings of the 37th Hawaii International Conference on System Sciences.*

Nonaka, I. (1994). A Dynamic Theory of Organizational Knowledge Creation. *Organization Science*, *5*(1), 14–37. doi:10.1287/orsc.5.1.14

Nonaka, I., & Takeouchi, H. (1995). *The Knowledge Creating Company*. NY: Oxford University Press.

Ozden, B., Ramaswamy, S., & Silberschatz, A. (1998). Cyclic Association Rules. *In Proceedings of the 14th International Conference on Data Engineering* (pp. 412-421). IEEE Computer Society Press.

Perusich, K., & Mcneese, M. D. (2006). Using Fuzzy Cognitive Maps for Knowledge Management in a Conflict Environment. *IEEE Transactions on Systems, Man and Cybernetics. Part C, Applications and Reviews*, *36*(6), 810–821. doi:10.1109/TSMCC.2005.855509

Ramaswamy, S., Mahajan, S., & Silberschatz, A. (1998). On the Discovery of Interesting Patterns in Association Rules. *Proceedings of the 24th International Conference on Very Large Databases* (pp. 368-379). Morgan Kaufmann.

Wegaer, P. (1998). Interactive foundations of computing. *Theoretical Computer Science*, *192*(2), 315–351. doi:10.1016/S0304-3975(97)00154-0

Wegner, P. (1997). Why interaction is more powerful than algorithms. *Communications of the ACM*, *40*(5), 80–91. doi:10.1145/253769.253801

Wolverton, M. (1997). Exploiting enterprise models for the automatic distribution of corporate information. *In Proceedings of the 6th International Conference on Information and Knowledge Management* (pp. 341-347). New York: ACM Press.

Zhao, J., Kumar, A., & Stohr, E. A. (2001). Workflow-Centric information distribution through email. *Journal of Management Information Systems*, *17*(3), 45–72.

Zhuge, H. (2002). A Knowledge flow model for peer-to-peer team knowledge sharing and management. *Expert Systems with Applications*, *23*(1), 23–30. doi:10.1016/S0957-4174(02)00024-6

Zhuge, H. (2005). Semantic Grid: Scientific Issues, Infrastructure, and Methodology. *Communications of the ACM*, *48*(4), 117–119. doi:10.1145/1053291.1053325

Zhuge, H. (2007). Autonomous semantic link network model for the Knowledge Grid. *Concurrency and Computation*, *7*(19), 1065–1085. doi:10.1002/cpe.1097

Zhuge, H., Guo, W., Li, X., & Ding, L. (2005). Knowledge Energy in Knowledge Flow Networks. *Proceedings of the First International Conference on Semantics, Knowledge and Grid* (pp. 3). Washington, DC: IEEE Computer Society.

ENDNOTES

1. For example, if you browse a text, next you may browse a video or an image which have an associated relation with the text.
2. Interactive computing is used on semantics layer. So this section will focus on the semantics layer and overlook the P2P location algorithm.

This work was previously published in Quantitative Quality of Service for Grid Computing: Applications for Heterogeneity, Large-Scale Distribution, and Dynamic Environments, edited by L. Wang, pp. 204-218, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 8.14
Knowledge Producing Megamachines:
The Biggest Web 2.0 Communities of the Future

Laszlo Z. Karvalics
University of Szeged, Hungary

ABSTRACT

In the present study, the authors point of departure is the control crisis of science whose resolution requires radical social innovation. The author then shows that the only possible way for achieving this is the partial fusion of certain portions of scientific activity with the system of public education, by means of organizing scientists, teachers, as well as middle and high-school students into hybrid, knowledge producing mega-machines. The author shall subsequently argue that doing so will at the same time bring about a pragmatic shift in public education, for which professionals in the field of pedagogy have long been ready in principle and in theory. As a final result we shall see the emergence of science and public instruction tailored to the global system level, within the framework of the information society.

DOI: 10.4018/978-1-60566-272-5.ch002

INTRODUCTION

The best way to predict the future is to invent it.
(Alan Kay)

The two perhaps most important sub-systems of the Information Society, Science and Public Education, are confronting a social innovation process of staggering force. Even though the research workshops of particular countries produce sensational results day by day, and the national systems of public education undergo continuing renewal, nonetheless in terms of their interest structures, their institutional mechanisms, and their financing, both science and public education have up to the present day continued to carry the imprint of the industrial era. For that reason, their functional disturbances can be managed ever less effectively by short lived pseudo-reforms of purely transitional impact.

The information technology background systems of modern sciences produce an incredible quantity of output signals. For many of the sciences (primarily genetics, oceanography, meteorology/climatology, environmental sciences, nuclear physics, pharmacology, archeology, and, first of all, astronomy) it is more and more problematic to manage the content of their permanently swelling background stores Beside financial resources the "human agent", human infrastructure, is becoming one of the bottlenecks. If we need brains in a "pre-digestive" process, it can easily find them where the task is exactly to make these brains able to do (even) scientific work: in the school benches. With the pupils socialized in the adequate community scope, involving resources and learning basic knowledge to satisfy their sateless desire to know and with their teachers an alliance may be created, and the biggest human GRID (the biggest Web 2.0. community) will be composed from these hybrid online clusters – the new type of knowledge producing and learning communities.

This process will, however, not run its course automatically. It requires efforts aligned with the same orientation, over several decades, by scientists from the various specialized fields, coordinators of instruction, political decision makers, teachers, social researchers, and information technological system developers. The final result guided by a vision, and the broad sweep of the project that builds the path leading to that goal, make for social innovation of a scope and importance whose like has never before been formulated either by the sociology of science, by the philosophy of education, or by research in futurology.

At this point our assertions concerning the future are hypotheses. Our aim is to elaborate scenarios ripened in a series of debates and work toward consensus-based conceptual structures, all of which will make it possible to initiate effective and soundly based social action and coordination, if and when the vision gains acceptence.

THE CONTROL CRISIS OF SCIENCE AND THE INEVITABLE CONTROL REVOLUTION

In the relevant literature there is a general acceptance of statements such as that the globalization of science has accelerated, that modes of knowledge production are emerging which follow new patterns, or that the rapid build-out of the new cyber-infrastructure of science introduces radical changes in methodologies of numerous scientific fields. There is, however, a considerable divergence of opinions concerning the depth of the challenge facing science and what the most comprehensive framework might be for interpreting the respective changes.

Beniger

On my part, I consider the model introduced by James Beniger in his epochal work, The Control Revolution (Beniger, 1986), to be the most fertile theoretical approximation. I hold so because the current situation of science can be elegantly interpreted using Beniger's category of a control crisis while also convincingly revealing the defining features of the incubating control revolution.

Shortly after the publication of his book, Beniger himself attempted to summarize in an independent study how his model might be extended to global science (Beniger, 1988).

The control revolution was the successful answer given to the lightning-fast process of industrialization which evolved during the century following the 1830's. Beside reining in speed and energy, adequate answers were successfully found to governance and enterprise management through technological innovations supporting the flow and elaboration of information, together with the social innovation of modern bureaucracy. Beniger's attention is drawn early to the double role in this process played by telematics, the increasingly interwoven world of information and communication systems. With its innovations, telematics supports the broad establishment of new and effective control structures. Yet in so far as the very processes whereby information is interpreted and evaluated for control purposes, are not successfully subjected to regulation over and over again by use of adequate methods, the feedback weakens and the system runs into new forms of control crisis.

When Beniger applies this to science, as a system constructed par excellence from the streaming of information flows, he perceives almost everywhere the indications of a growing control crisis. He finds the primary threat in the large-scale startup of new systems of telematics which disturb, or with their excessive radicalism—because they abandon a paper-based world—even disorganize, the accustomed flow patterns of already produced knowledge. Thereby they further weaken the functioning of the most important feedback mechanism, the citation system.

It is strange how completely Beniger (1988:26) is mistaken when he has fears for scientific reports, the publishing of specialized journals, or the publication of conference proceedings in their capacity as feedback mechanisms, on account of their exposure to information challenges. *"Telematics threatens global science … [with] … a crisis of control. Many involved with the computerization of information systems have predicted—some gleefully—a decline in the formal scientific paper, a blurring of the distinction between research notes and papers and between papers and the response to them by others, an increase in multiple authorship by scores or even hundreds who participate in a telematic discussion, and the decline of formal journals, editors, and the gatekeeping function more generally."* (Commenting on Beniger's work, Gerhard Fröhlich shares this anxiety (Fröchlich, 1996)

Surely, the last two short decades have brought a control revolution precisely from this point of view, with the creation of the new environment of digitalization, web-browsing, or full-text search capability. (For a summary of this point of view, see (Nentwich, 2005). And yet, at the same time Beniger's general model is more relevant than ever. The current control crisis of science cannot be found along the obstacles of accessibility to results; it has shifted to the scenes of new knowledge creation. By to-day the lack of feedback can already be attributed primarily and principally to system-level constraints on the interpretation of knowledge and of raw information and data.

Revolution in Sign Production, Shortage in Brains

Modern sciences, with their up-to-date information technology parks, are producing output data in quantities already so staggering as to make these incapable of being overviewed in a properly interpretive manner by the scientific community—which, to make things worse, is continually perfecting its capacity of producing and storing even more new information and data. *"New data – whole new types of data – are accumulating faster than researchers can make sense of them. The result is something like an optical illusion"* (Hugh Kieffer, cited by Norton, 2004). Those sciences which have high levels of demand for raw data and their interpretation—genetics, oceanography, meteorology/climatology, environmental sciences, atomic physics, pharmacology and above all astronomy—are currently undergoing the cyber-infrastructure revolution with their GRIDs, their enormous capacities for calculation, simulation and visualization, their more and more intelligent agents and work-flow devices.

Yet the scientists are aware of the control crisis. They all have the bitter experience that their efforts of building new models and coming up with pioneering connections and hypotheses are constrained by the small capacity of the analytic personnel available for handling lower-level, supportive transformational tasks. These tasks include surveys of measurement data, of elementary objects, or of relevant singular events; the testing of map structures; or the confirmation/ verification of masses of elementary correlations. Any successes achieved in automating or computerizing the analysis of the raw data will only reproduce at the next higher analytic level the experience of support personnel being unable to cope with the mass of transformational tasks.

In the past, scientists had met this experience only when surveying the specialized literature and running into the limits of the library services or the reference/abstract/search systems. Yet by now the capacity limit shows up in relation to the output of each scientist's own data—a control crisis that cannot be managed by traditional approaches. This is because until now the preferred tool of control revolutions was the automation/computerization of the kind of human intellectual effort that could be translated into appropriate algorithms, just as the computer itself had replaced human computations done by pencil and paper. (Grier, 2005)

Wherever this algorithmic translation can be continued—as in the case of robot librarians or the "robot scientist" used in gene sequencing—the impact of the control crisis can be moderated. As A. M. Weinberg says, *"extreme automation may be appropriate for those activities that are time constrained, it may lead to clogged information channels for those scientific activities for which time is less important than depth of understanding."* (Weinberg, 1989). Thus the scientific community has come step by step to the recognition that the bottleneck is in the areas of knowledge and insight that cannot be reduced to algorithms; the process of knowledge production is constrained by human brains capable of interpreting, placing in context, and thereby counterbalancing the sheer mass of raw data being generated. Therefore the new scientific control revolution can only arise from the human infrastructure, it can only be a human revolution—and as such calls not for technological but for social innovation.

PRELUDE TO THE KNOWLEDGE PRODUCING MEGA-MACHINES

During the past three decades the science establishment has tried to ease the intensifying pressure with three parallel minor innovations:

1. Organizational and institutional solutions facilitating the optimal use of the available capacities and numbers of human brains (*intensification*).
2. Interconnection of existing research staffs into virtual communities of ever growing size (*concentration*).
3. Attempted massive mobilization of new brains capable of being involved in the solution of scientific problems (*extensive growth*).

Each of these solutions yielded some temporary and partial results—but these results paradoxically ended up by reinforcing the basic problem, because of the increased demand for data production stimulated precisely by the successful feedback. We shall show below, one by one, why neither of the three approaches can be expected to promise further advance. Yet each of these three attempts has made a significant contribution to identifying the feasible path toward the real control revolution.

Intensification has time as its equivalent. When the staff of research assistants suddenly increases, this also means that we have to discover a way of reassigning the precious time of leading scientists to activities that yield higher added value. Sophisticated bibliographic software is meant to serve the same objective. And scientists gather into problem-centered invisible colleges which are arising in parallel with the hierarchies of traditional authority, in order to make sure that the time required to arrive at new insights will be shortened by means of an intensive exchange of knowledge. The fundamental constraint on such ways of intensification arises from the ever smaller part of the aggregate knowledge in a given area of science that has come to be represented as the personal knowledge of any one individual. As formulated close to forty years ago by an outstanding Hungarian economist, Ferenc Jánossy, "*Precisely this is why we have to beware of tugging by sheer force at the net of individual knowledge, trying to cover all of the increased area, until the meshes are torn and only large holes remain instead of the dense coverage of the net.*" (Jánossy, 1975). This recognition opens the way to the search for more decentralized and democratic knowledge production models.

Concentration is embodied in ever larger cooperating communities of researchers. The pioneering Human Genome Project has been followed by several similar research undertakings whose common element is the allocation of human resources required by such extensive research tasks, on the basis of novel principles.

Precursors

At the end of 2006 more than 100 thousand scientists from more than 175 countries have taken part in the **Innocentive initiative.** The website (http://www.innocentive.com) organizes the cooperation of world-level researchers as *problem solvers*, and companies wishing to solve their development questions focusing on science, as *seekers*. It is effective primarily in the search for innovative answers given to complex challenges, mainly in the fields of pharmacology, biotechnology, chemistry, food industry and plastics industry, with financial awards often exceeding $ 100,000.

The **Academici** website offers a search framework constructed for facilitating any emerging scientific cooperative initiatives and has thereby made it possible for scientists and researchers anywhere in the world to share their experiences at any time without limits and restrictions or to discuss any questions, proposals, suggestions, or problems. On www.academici.com those interested can search according to several criteria, such as research field or scientific interest. In this way educational or scientific institutions located at the greatest world distances from each other can mutually get in touch in a matter of minutes.

The **DILIGENT project** of the European Union develops safe, coordinated, dynamic and cost effective test beds for virtual scientific communities, to facilitate the sharing of knowledge and cooperation. Involving the website (http://www.diligentproject.org/), experiments are proceeding in two real-time application fields, an environmental sustainability project and a cultural-heritage preservation topic, by combining the Grid and the DL (digital library) technologies. It is readily apparent that from the point of view of control structures, the size constraint is given by the maximal number of the scientists who can be meaningfully interconnected. Meanwhile, the cooperative forms, workflow solutions, software systems, and online cooperative cultures created for several tens of thousands of participants are already paving the way for making possible the operations and organization of research communities of several-million-members which are expected to emerge in the course of the control revolution.

The Advent of Participatory Citizen Science

Extensive growth comes from drawing into science some social groups that had earlier, to some extent consciously, been excluded. The orientations of citizen science and participatory research (Irwin, 1995, Park et al., 1993) do not fight for the renaissance of amateur science but for the integration of knowledge created outside the scientific establishment, into the scientific problem solving processes. Typical examples, often with the spirit of movements, come into being and evolve in order to support the scientific emancipation of indigenous knowledge, that is, knowledge concerning the local environment, weather, therapies, animals or plants.

There exists, however an even more comprehensive context, the aspect of "open science" which fights, on theoretical and historical grounds, for breaking down the walls between Science and Non-science (Dasgupta and David, 1994). Open science, in the spirit of a new ethos (actually, one harking back to the 17th century) and fired by powerful economic arguments, carries the banner for a program of broadly spreading out such knowledge as already exists, and hopes from this — among other things — for a growth in the number of persons involving themselves in the cultivation of science. The attraction of new brains has been started in several ways, based not upon theoretical but rather very practical considerations, by drawing participants primarily into problems of the environment and sustainability, because of their personal exposure and ready competency. The projects with the broadest participation are, however, connected to space research where millions of amateur astronomers have even earlier been able to get into the scientific establishment with their observations and results. In the highly computation-intensive SETI@home project only the redundant computer time was lent by more than four million people, but in the Stardust@home project voluntary brains were already also being needed, in order to act as virtual microscopes in the identification of collected micro-particles. (The *Johnson Space Center,* with the help of the *Planetary Society* and the *University of California at Berkeley,* has performed a high-resolution scan of the so called aerogel used for capturing spatial dust. The scanned micro-images are available in 700,000 movie-length segments via the Internet, containing less than 50 expected micro-particles scattered like needles in a vast haystack. Results obtained in interpreting the images by voluntary scientific participants joining in from the public, if proven reliable by tests, will help in finding and isolating individual micro-particles. These will be named by their discoverers (http://planetary.org/programs/projects/stardustathome/).
This monotonous and tiring work required what amounts almost to a hunt for volunteers, who were permitted to join the search after successfully passing a test.

Practices of bringing new brains into science, undertaken so far, clearly demonstrate that in selected cases this may lead to success; it does, however, not provide a system-level solution to the problem of the missing human infrastructure. Voluntarism and contingency cannot be reconciled with a systemic revolution of control. And yet, we owe a lot to these projects, because they are proving day by day they that it is possible both to imagine and to operate problem solving communities of never seen size, and to boot, with outsiders. In the public discourse, they strengthen narratives relating to the decentralization, democratization and massive expansion of science, and thereby they open up the door to the real control revolution—the creation of knowledge producing mega-machines.

PLAN OF A MEGA-MACHINE

The mega-machine is Lewis Mumford's noted category (Mumford, 1967, 1970). Studies in the history of technology often refer to it as the term for a large and hierarchically organized task community—from the pyramid builders to the many armies of history. Mumford intended, however, to convey more than a simple metaphor. He was in part also interested in how the human components of a mechanism cooperate with the tool components, the mega-technics; how a few large cities which attain a central role in communications, labeled megalopolis, become the scenes of these events; how social control of technology emerges; and how communities of this size can be described with the terminology of cybernetics. Mumford (1967:191-192) was interested in how "*… to turn a random collection of human beings... into a mechanized group that could be manipulated at command. The secret of mechanical control was to have a single mind with a well-defined aim* *at the head of the organization, and a method of passing messages... (to) the smallest unit."*.

We can even consider the information society itself, in its most comprehensive sense, as a mega-machine (May, 2000). Yet we find the expression especially useful when we are looking for a name to designate the kind of knowledge-producing communities that will be able to exceed the largest present scales of scientific cooperation perhaps a hundredfold. And since we have seen the short-comings of current efforts aimed at increasing the number of brains participating in scientific research tasks, the conditions outlining the criteria that future scientific mega-machines will have to fulfill have become quite clear:

- Guarantee of staffs of sufficient size.
- Guarantee of staff availability.
- Professional-methodological quality assurance of activities.
- Accountability, dedicated areas of responsibility, deadlines met and capable of being met, clear project scenarios, and professional management.
- Organized transfer of or instruction in the knowledge minimally necessary to join a project.
- Facilitation of tasks, continuous mediation among individual participants of the system.
- Flexible, many-sided online system management service, oriented to facilitating horizontal as well as vertical interactions.

But where could the many tens of millions of brains be found for this? How could communities of this size be made manageable and operable in line with the above expectations?

According to a rough but conservative extrapolation based on UNESCO data for the year 2004 (http://gmr.uis.unesco.org/ViewTable.aspx), by 2007 there were approximately *800 million students* in the 12-18 age cohort, supervised in an orderly and structured manner by some *40 million teachers*. Public education is a human mega-machine that is already at hand, striving, to boot, toward a fully online existence at astonishing speed. It appears to be perfectly suitable, while radically renewing its current functioning, to become at the same time also a mega-machine for research, thereby solving the control crisis of science. Or put in a different way, other than the hundreds of millions who are now articulated into the national systems of public education, we can find no social group on earth or in heaven that would so completely satisfy our prior expectations concerning size, availability and organization.

Why is it then that this statement, when we first come upon it, should nevertheless feel so startling, bizarre, and utopian? Indeed, once we realize that it is necessary, we only have to convince ourselves that it is possible. If we accept the goal — to plan and put in motion the new mega-machines built on public education — then we can take our time pondering the details of the path that will lead us there. So the challenge does not lie in listing all the steps that are indispensable for implementing the initiative but rather, in figuring out how to make the vision itself realistic. Can it possibly be sustained as a hypothesis? Is a mega-machine conceptualized in this way at all imaginable? Is it outlandish in relation to the institutional and functional order of public education or to the contrary, is it wonderfully adaptable to that order? But above all, why do we find it so hard to throw ourselves enthusiastically into listing the fantastic functional consequences that we hope will flow from the birth of the new mega-machine?

Perhaps it sounds strange at first but it is impossible to find a single serious and objective counter-argument. Yet acceptance of the vision is all the more impeded by hardened views, mistakes, preconceptions, misunderstandings and optical illusions, which must be brought into the open, exhibited, and demolished before we can once again take a searching look at the vision itself. As it usually happens, when seeking to define the outlines of the Future, we run into the attitudinal roadblocks of the Past.

EDUCATION: THE SPIRIT OF THE INDUSTRIAL ERA, OR THE DELAYED CENTURY OF THE CHILD

Ellen Karolina Sofia Key wrote her emblematic book entitled The Century of the Child in 1900 (Key, 1900), a polemic discourse in which she attempted a showdown with the shadows of the 19^{th} Century. Ellen Key envisaged a radically new 20^{th} Century built on a positive child image, in which the schools would no longer deliberately freeze but rather set free the intellectual and spiritual energies hiding in the child. Sadly, the Swedish author was mistaken. Despite all the emancipatory, promising initiatives of that century, the real turnaround is still waiting in the wings. The school as an institution has shown an appalling lack of change ever since the end of the 19^{th} Century. Its objective function pursues social and labor market integration, its image of the child is paternalistic, and its pattern of knowledge flow is determinedly one-way. The school decants the necessary abilities, skills and knowledge into the child's head with the help of the teachers and the material of instruction, while keeping the child's time budget, comportment, and knowledge organization under strict control, standardizing the required minimal levels of knowledge. The reason why the tired commonplaces of the so-called world crisis of education never mature into an interpretation as a control crisis, is simply that the school is actually experiencing no crisis at all, it is fulfilling its task, the feedbacks are working — just so long as the fundamental assumptions and criteria of success continue to be those of the industrial era. Likewise, the science of the industrial era is not in crisis, either, because the processes of selection for the replacement of its cadre of scientists are continuing smoothly without any hitch. (And of course the alarm bells are immediately rung as soon as the process begins to falter somewhere, with particular disciplines attracting insufficient interest on the part of the relevant age groups.)

But let us now replace the out-of-date objective function with that of the information era. From the point of view of up-to-date knowledge asset management, is the school getting a suitable return from the brains committed to its care? Does it deal with the knowledge carried by the students on the basis of an appropriate valuation, does it regard such knowledge as a capital asset, does it make an effort to integrate that knowledge in a truly transformative fashion as much as possible with the total mass of knowledge?

If we measure the school's quality of functioning by how well it creates a foundation for the holistic image of the human essence as one of life-long learning, then it is immediately apparent that we are in deep trouble. The crucial question therefore becomes whether a totally and essentially industrial-era mega-machine might be capable of renewing itself by assimilation to the demands of the information age, or whether the imperative of change must necessarily aim at dismembering the mega-machine itself. Ivan Illich's answer was clear, definite and scandalizing: the school as an institution was itself the chief obstacle to renewal.

The current search for new educational funnels must be reversed into the search for their institutional inverse: educational webs which heighten the opportunity for each one to transform each moment of his living into one of learning, sharing, and caring. (Illich, 1970)

In my personal opinion, a process of creative destruction leads to far more difficulties than the re-planning and re-engineering of the mega-machine of public education. The school has changed much, to its advantage. For a long time, pedagogical theory has also been knocking on the door with a set of claims that it delineates normatively in connection with the school, the teaching/learning processes, and the children of the information age. Specifically:

- Education should move away from authoritarian orthodoxy toward a world of horizontal communications and of cooperating, problem solving teen-age student communities.
- Let us teach each child from an early age to think critically, thereby supporting the development of a facility that allows pondering problems from different perspectives and, even independently, with a thoroughness maturing toward the scientific level.
- The teacher should be an animator, not a device for the recitation of instruction material. The teacher should smartly direct the independent knowledge operations of the students that also make use of library and Internet resources, intervening only at critical points.
- Wherever possible, the principle of learning by doing and getting the children to make their own discoveries should replace mindless cramming.
- Education should build on children as self-confident and smart users of the world of information-technology devices and systems who are already on the brink of absorbing this experience and knowledge at the same time as their mother tongue.

If we want to build a pragmatic pedagogical program on the above theoretical expectations, we could nowhere find a better option than to involve teen-age students and their teachers by the millions into the suitably prepared and well-organized problem-solving processes, involving network communications and cooperative work, that are emerging in a number of scientific areas.

The spirit of the out-of-date industrial era is left with just one position it can take to save it from retreat. That is to call into question the suitability of the 12-18 age group for active and creative participation in scientific group work.

Mini-Einsteins versus Intellectual Immigrants

It may seem gratifying to collect a colorful posy of the latest acts of student geniuses. How a Hungarian high-school student living in Calgary, with two weeks' work, succeeded in proving Albrecht Dürer's five-hundred-year old geometric conjecture which had earlier resisted the mathematical community's best efforts? How a fifteen-year-old boy, jointly with his world famous co-author, published an excitingly original astrophysical hypothesis? How teen-age girls were involved in choosing the optimal landing point for the Mars terrain exploration robot?

Yet, what needs proof is not that talented teen-age students are at times capable of producing results comparable to those of grey-haired professors. We lose our way if we try to look for economy-size scientists among 12-18 year old children; the mega-machine demands something entirely different. The students, with the help of their teachers, must be capable of and suitable for appropriately undertaking massive partial tasks, typically of a low level of abstraction, yet still calling for natural intelligence. Let us replace one single research assistant with a hundred children and four teachers, and we will begin approaching realistic proportions. And ten thousand children with four hundred teachers may well provide us with as much problem-solving power for a scientific program as a research staff of one hundred. Of course, what we should expect from a student is, accordingly, one hundredth of what we may expect from an adult scientific researcher.

All this is well known by all those who had tried their hand at on-site scientific work jointly with massive groups of students. Where research programs, because of their very nature or a lack of support, have to make do without sufficient human resources, they increasingly turn to teenagers under the whip of necessity. Our space limitations preclude displaying the variety of color and the richness of form of currently running efforts of this kind. Tens of thousands of students are becoming articulated into research programs of great vitality, in the course of which they undertake full-valued ornithological observations, ecological surveys and measurements, or data collection efforts pertaining to local history. The exacting database of Estonian trees has been compiled by the student participants of the Tiger Leap program of the Estonian Schoolnet. The SG@Schools project in Singapore is being planned with reliance, in part, on the efforts of students in gene sequencing, animation tasks, and complex financial calculations.

The screenagers, representatives of today's generation that had grown up on television and the computer, are adapting to the new cognitive environment —according to Douglas Rushkoff's apt remark (Rushkoff, 1996),— as naturally and flexibly as the children of immigrants learn the language of the recipient country, faster and more effectively than their parents. And it is hard to imagine that the digital kids, who effortlessly cope with directing the traffic of the Chicago airport on a simulation game, who are daily managing extended online contact networks, and who can smartly transform information from one complex format into another, were not to be destined to accomplish even more. Would they not be suitable for administering questions in sociological surveys — perhaps concerning precisely their own age group? Or to read original sources and abstract their contents? Or to translate professional texts with their teachers' help? Or to answer questions, even by the thousand, put to them by scientists about what they find digging deep into their individually assigned little fields of inquiry? Or to try aggregating their partial bits of knowledge within flexible ad hoc communities? Of course they would. It would take no more than looking at them and seeing them in this particular way.

Still, many think that scientific activities are not for teenagers, that it is better to keep science at the level of illustrations meant just to awaken their interest, in order to make them disposed to learn. Their situation is therefore powerfully illuminated by the results of comparative educational surveys which show that if we integrate a broad collection of scientific problem solving methods, approaches, and philosophies into educational practice from an early age, we find an abrupt increase in children's disposition to learn as well as in their learning performance. Pedagogical action research has also shown that just as the acquisition of computer skills creates no particular problems, in the same way neither are there cultural, linguistic, religious or social obstacles to the acquisition of skills in undertaking activities of a scientific type. By transforming the public education system in a way that integrates information and science literacy into education from a very early age, we can promote an equality of opportunities more effectively than in any other way.

The children stand ready to become parts of the knowledge producing mega-machines. It is only the scientists, education-policy leaders, parents, and teachers who will have to come to believe this.

The Architecture of the Mega-Machine and its Impacts

In the coming years or decades, a series of research projects and experimental programs will put together and give precise shape to the basic structure and to the thousands of small details of the operation of knowledge producing mega-machines. In the discussion below I wish to survey, without any pretense to completeness, some of the fundamental aspects and characteristics that will almost certainly be relevant to the future system:

- Mega-machines will organize as hybrids of researchers, teachers and students which are multiply articulated in depth and in which assignments are distinguished as a function of the time requirement, profundity, and scope of each task.

- The age, experience, interests and earlier project involvements of each teen-age student will point to many differing roles. The representatives of science are also bound to arrive at specific divisions of labor amongst themselves. The ones facing, however, the largest shift in their identities are the teachers. At an earlier stage of their careers they had to choose between scientific work and the teaching profession which offers a lower social prestige. Yet as knowedge brokers, directors, and coordinators of bona fide research sub-projects, they will move into a totally new, inspiring professional and motivational environment, in their role as essential cogwheels in the functioning of the mega-machines. Beside this, the continuing role of exposition will also remain part of their responsibility — the awakening of interest, the nurturing of essential basic skills, and the incubation of the ability of students for participation in scientific group work.
- The disciplines currently experiencing the deepest control crises are the ones most likely to take the lead in building up their problem-solving clusters. These include genetics, biotechnology, and the environmental sciences as one set of fields; space research and astronomy as another; and finally, from among the social sciences, history and archeology.
- The inclusion of students in building project organizations can begin above the ages of ten or twelve. It seems to be obvious that there will be a particular system level that includes all students as parts of large, long-term undertakings aimed at structuring vast masses of raw information. This will involve the continuous follow-up of modular objects tailored to the capabilities of individuals (or of elementary research communities) and the steady enrichment of the knowledge pertaining to these objects (heavenly bodies, particular gene sequences, archaeological specimens, or historical sources), together with the mobilization, as required, of such knowledge. From here, not everyone's way necessarily leads to problem solutions at higher system levels, but all will remain members of this elementary problem community where one can continue doing one's modest partial task as long as attending school or even beyond, life-long.
- Within the mega-machine, dynamic cluster formation will necessarily get going along the typical attributes of the scientific problems under study. And since the unit topics, if they are well chosen, are likely to be multi-dimensional and complex, the students, in their capacity as individual owner/managers of particular dedicated problem objects, can be members of three, four, or even more problem communities organizing around particular attributes.
- The representatives of science are present in the system in part as clients who, on the basis of jointly agreed criteria, define the tasks to be undertaken. The tasks then tend to spread out within the mega-machine, possibly with benign, multiple overlaps. The scientists are also present within the problem field continuously and interactively, with responsibility for quality assurance, the creation of professional-methodological standards, and enforcement of the latter. The system harbors the possibility of competition and choice; the potential topics compete for resources, and individual clusters are relatively autonomous in choosing the focus of their activities. The results obtained thanks to the mega-machines will be the shared treasure of humankind.

- Other than maintaining the national language in scientific work, the force field guiding the operation of the mega-machine is basically a global one. The clusters are kept in motion by horizontal contacts among students of diverse nationalities, languages, cultures and religions, and through such an interconnectivity a new, global consciousness emerges in a meaningful way.
- This is closely connected with global problems that have so far called for consciousness raising in part as a challenge to education. The school of the future—in James Martin's words—will also be a civilization school, since it demands that the basic knowledge necessary for the survival of civilization be introduced into the curricula urgently and in their full range. "Education for survivability—the most important subject we can teach." (Martin, 2006). A work of similar outlook and message but with richer development of the topic is (Adams and Carfagna, 2006).
- Yet all this will be perceived as more organic and credible once everyone gets a share in laying a scientific foundation for the solution.
- For several reasons, the necessary lingua franca of the mega machine can only be English. (It is not impossible, though, that over the long run the torch will be taken over by an artificial language developed especially for this purpose.)
- Even though the research activities of the students are tied to practical reality by the objects of study individually assigned to them, the online work management systems nevertheless are taking on a key role in the process. It is an important question whether the open platform of the Internet is suitable for handling interactive efforts by several million participants, requiring work management systems of never seen dimensions. In any case, it is worth paying increased attention to software developments based on pioneering principles that promise solutions of entirely new levels of effectiveness. (We consider the Croquet project, www.opencroquet.org, as one of these.)
- Finally, it is important to emphasize that being a part of the scientific mega-machine will only be one side or one function of school life; there will be no total change of the guard compared to earlier solutions of instruction and subject matter, only an internal reorganization of proportions.

POLICY IMPLICATIONS

James Beniger notes correctly that in the case of large-scale social innovations it is always politics that has the decisive word: *" … the information society does not spring spontaneously from advanced industrialization. Technological possibilities for control present societal choices, which are themselves subject to political control"* (Beniger, 1988:22). And if innovation is global, as is the case with the future hybrid mega-machine of science and public education, then as an inevitable complication, two political system levels get in each other's way. Beside national education and science policies, the international organizations representing the global level are also very much concerned, since the mega-machine can only be coordinated in a transnational space. And although UNESCO's profile (science, education, infocommunication) would predestine the organization to lead such an important innovation program, looking at its current planning and operational system it is hard to imagine that it would be able to fulfill the task. It is more probable that the professional world organization of one of the scientific disciplines might come to a consensus about building a smaller pilot prototype of the mega-machine. And similarly, within the strategic-political decision space of the nation state it is hard to expect that a vision, no matter how realistic it may be, will gain priority, let alone support, solely on the basis of its projection of a future that is highly attractive in many ways. It is much more probable that the growing control crisis — whose spectacularly visible aspect is the increasing lack of financial and human resources compared with what would really be needed—will sooner or later inherently enforce the search for solutions. When national science policies see no other way, they will out of sheer necessity reach for the possibilities offered by the mega-machines.

It is especially difficult to realize and accept that in a global social innovation of such scope the expected benefits will not accrue in the form of traditional competitive advantages.

Why should the United States wish to turn in this direction, given that its leading position attained in science is sure to be safe for a long time to come, even without the upheaval of the mega-machines? When does the moment of epiphany arrive, insisting that the value of new knowledge produced by the mega-machine is becoming clearly measurable even under conventional budgetary criteria? How will the truth get to the point where it can be grasped—once again from a budgetary or national security perspective—that common actions organized in favor of common scientific aims reduce to insignificance or overwrite the traditional forms of international conflict? Let us not forget, it is part of the vision that the educational systems of the world's zones of conflict will also be integrated into the mega-machine. In the 21st century which offers little chance for religious, cultural or political convergence, could precisely science not become one kind of common language?

And what about the smaller countries? Is there any point for them to get going with their own mini-scale mega-machines? For example, on the basis of their recent educational and science-policy accomplishments, mentioned earlier, Finland, Estonia, and Singapore would without further ado be suitable and mature enough to experiment with sciences, themes, clusters—mobilizing in their entirety their teenage student cadres of hundreds of thousands. And what stands in the way of the possibility of these countries starting to build a mega-machine in cooperation with an ambitious country like Chile, with one of the Chinese or Indian provinces, or with one of the smaller American federal states?

Even the longest journey begins with the first step. Are we standing ready to initiate a discourse?

ACKNOWLEDGMENT

For the help I have received in preparing the final verson of this chapter, I wish to express my sincere thanks eminently to *Thomas Vietorisz* (Cornell University) and to *Adam Tolnay* (Stanford University), *Nicholas Vonortas* (George Washington University), *Gabor Agoston* (Georgetown University).

REFERENCES

Adams, J. M., & Carfagna, A. (2006). *Coming of Age in a Globalized World: The Next Generation Kumarian Press* (See especially the section titled: Global Education: Schooling World Citizen).

Beniger, J. R. (1986). *The Control Revolution. Technological and Economic Origins of the Information Society*. Harvard University Press

Beniger, J. R. (1988). Information Society and Global Science. In Telescience: Scientific Communication in the Information Age. *Annals of the American Academy of Political and Social Science*, *1*, 14-28. The study was republished in Dunlop and Kling (1991)

Dasgupta, P., & David, P. A. (1994). Toward a New Economy of Science. *Research Policy*, *23*, 487–521. doi:10.1016/0048-7333(94)01002-1

Dunlop, C., & Kling, R. (Eds.). (1991). *Computerization and Controversy: Value Conflicts and Social Choices.* San Diego: Academic Press

Fröchlich, G. (1996). The (Surplus) Values of Scientific Communication. *Review of Information Science*, *2*.http://www.inf-wiss.uni-konstanz.de/RIS/1996iss02_01/articles01/02.html

Grier, D. A. (2005). *When computers were human*. Princeton University Press

Illich, I. (1970). *Deschooling Society.* New York: Harper and Row

Irwin, A. *(1995)* Citizen Science: A Study of People, Expertise and Sustainable Development (Environment and Society).

Jánossy, F. (1975). *A gazdasági fejlödés trendvonaláról* (On the trendline of economic development) (2. enlarged edition, Magvető, Budapest. (English version: The end of the economic miracle; Appearance and reality in economic development International Arts and Sciences Press (1971)

Key, E. (1900). *The Century of the Child*. In English it was first published in 1909, following the German translation of 1902.

Martin, J. (2006). *The Meaning of the 21st Century. A vital blueprint for ensuring our future*. Riverhead Books

May, C. (2000). The Information Society as Mega-Machine. The Continuing Relevance of Lewis Mumford. *Information Communication and Society*, *3*, 241–265. doi:10.1080/13691180050123721

Mumford, L. (1967). *The Myth of the Machine, 1: Technics and Human Development*. New York: Harcourt Brace Jovanovich. The Myth of the Machine, vol. 2: The Pentagon of Power. New York: Harcourt Brace Jovanovich, 1970.

Nentwich, M. (2005). Cyberscience: modelling ICT-induced changes of the scholarly communication system. *Information Communication and Society*, *8*, 542–560. doi:10.1080/13691180500418451

Norton, O. (2004). Planet Ice. Mars revisited. *National Geographic*, *1*(9).

Park, P., Brydon-Miller, M., Hall, B., & Jackson, T. (1993) *Voices of Change: Participatory Research in the United States and Canada*.

Rushkoff, D. (1996). *Playing the Future. What We Can Learn from Digital Kids*. Riverhead Books, New York

Weinberg, A. M. (1989). Science, government, and information: 1988 perspective. *Bulletin of the Medical Library Association*, *1*, 1–7.

This work was previously published in Social Web Evolution: Integrating Semantic Applications and Web 2.0 Technologies, edited by M. D. Lytras; P. Ordonez de Pablos, pp. 17-30, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 8.15
Utilizing Past Web for Knowledge Discovery

Adam Jatowt
Kyoto University, Japan

Yukiko Kawai
Kyoto Sangyo University, Japan

Katsumi Tanaka
Kyoto University, Japan

ABSTRACT

The Web is a useful data source for knowledge extraction, as it provides diverse content virtually on any possible topic. Hence, a lot of research has been recently done for improving mining in the Web. However, relatively little research has been done taking directly into account the temporal aspects of the Web. In this chapter, we analyze data stored in Web archives, which preserve content of the Web, and investigate the methodology required for successful knowledge discovery from this data. We call the collection of such Web archives past Web; a temporal structure composed of the past copies of Web pages. First, we discuss the character of the data and explain some concepts related to utilizing the past Web, such as data collection, analysis and processing. Next, we introduce examples of two applications, temporal summarization and a browser for the past Web.

DOI: 10.4018/978-1-59904-576-4.ch017

INTRODUCTION

As the Web changes continuously, it is necessary to preserve the past content of pages for a future reuse. The Internet Archive[1] is the best-known and largest public Web archive containing data crawled since 1996. Other Web archives exist, for example, ones containing Web pages from particular countries (e.g., Arvidson, Persson, & Mannerheim, 2000; Hallgrimsson & Bang, 2003). Besides, there are also numerous repositories of past copies of pages such as caches, site archives, personal page repositories or search engine caches.

Web archives provide a view on the history of the Web reflecting past societal states. Past content of pages can reveal the histories of underlying elements represented by these pages, such as institutions, companies, people or other entities. For example, one could approximately detect when a particular member left some laboratory by detecting the time point at which her or his name was removed from the list of laboratory's personnel. In general, the use of Web archives can greatly benefit researchers and practitioners in many areas, such as history, sociology or marketing.

Furthermore, analyzing information from the past can help not only in better understanding the history of our society but also understanding its present state. This is because Web archives can provide contextual information about Web pages and the objects or concepts discussed on them as well as their inter-relations. For example, we can analyze information from Web archives concerning a given company in order to use it as a context for better understanding the present information about this company. In general, mining past Web content has a potential to stimulate and improve the traditional Web mining process in the sense that it provides contextual information and sheds new light on present data.

Past Web is considered here as a part of the WWW space where pages no longer have any change potential; they are "frozen" past snapshots of pages. The live Web, on the other hand, is the present Web, containing pages that we can currently view online. These pages may be changed or updated and they usually provide full interaction capabilities.

In the past Web each page has its history and lifetime. Links between the old content of pages can be reactivated again. In this way, a temporal structure can be obtained reflecting connectivity between pages in the past. Another aspect of the past Web is missing data. A given content after its deletion from a page may never be reproduced if it has not been preserved in any repository. Besides, due to the rapid growth of the Web, selective type archiving often needs to be done.

In this chapter, we approach the problem of discovering knowledge from the past Web. First, we discuss the character of data that is used and methods for acquiring and processing it. We propose techniques for analyzing and selecting candidate Web pages for mining. This approach is based on analyzing long-term characteristics of pages with a special focus on their content changes as they are most interesting from the viewpoint of pages' evolution. Next, we introduce temporal summarization, which is an adaptation of a traditional text mining task into the past Web scenario. We propose summarizing histories of Web pages to generate abstraction of events and salient concepts described in selected portions of the past Web. We also discuss the possibility of discovering object histories in past content of Web documents. Finally, we describe an application for browsing and navigating the past Web. We show an implementation that is similar to those of traditional browsers for the live Web and of video players.

The rest of this chapter is organized as follows. In the next section, we discuss the related research and attempt to place this work in the wider context of text and Web mining. The following two sections describe the data accumulation, preparation and analysis. In the next section we discuss temporal summarization and investigate the possibility of object history detection from the past Web. The next section describes a browser for the past Web, while the last section concludes the chapter with a brief summary.

RELATED RESEARCH

Web Dynamics

The dynamics of the Web has been measured in many experiments (Brewington & Cybenko, 2000; Cho & Garcia-Molina, 2000; Fetterly, Manasse, Najork, & Wiener, 2003; Ntoulas, Cho, & Olston, 2004) which demonstrated that the content and link structure of the Web continuously change. Although many pages on the Web are short-lived, meaning they are deleted shortly after being created (Ntoulas et al., 2004), many important Web documents persist over time. Popular and main, or top-ranked, pages usually belong to this category as it often takes a long time for a page or site to gain popularity and accumulate a high number of in-links.

The results of Web dynamics research indicate the level of volatility of the Web as a whole. On the other hand, the study of update patterns of individual pages has been carried out for prediction of their future changes (Cho & Garcia-Molina, 2000, 2003; Ntoulas et al., 2004). The frequencies and degree of changes are the most often used measures to set up crawling schedules for maintaining fresh indexes of search engines. In practice, however, it is usually difficult to predict content changes in pages although some Web documents, for example, newswire sources, change in a more or less periodical fashion. In this research, we go beyond the simple analysis of change statistics as we focus on the distribution of content and its context over time.

Text Mining

Text mining is defined as a nontrivial extraction of implicit, previously unknown and potentially useful information from textual data. Text mining evolved from data mining and is a promising field as much information nowadays is stored in the form of electronic text. We consider our approach to be similar to temporal text mining, because, to a certain extent it resembles efforts that were taken in analyzing and mining streams of text data. Generally, mining news articles or other text streams along the time dimension has been studied well (Allan, Gupta, & Khandelwal, 2001; Allan, 2002; Kleinberg, 2003; Li, Wang, Li, & Ma, 2005; Mei & Zhai, 2005; Papka, 1999; Swan & Allan, 2000; Wang & McCallum, 2006). For example, the well-known TDT (Topic Detection and Tracking) research initiative (Allan, 2002) was aimed at detecting, classifying, and tracking events in news corpora. Recently, Wang and McCallum (2006) identified topics persisting over dynamic collections of documents. Another work showed the development of topic patterns in news articles over time (Mei & Zhai, 2005). Li et al. (2005) proposed a probabilistic model for retrospective event detection in news corpora. An approach toward temporal summarization of news events was proposed in Allan (2001) where novelty and usefulness of sentences retrieved from newswire streams were calculated for the construction of a final summary. Another related work called TimeMines (Swan & Allan, 2000) was proposed for finding and grouping significant features in historical document collections based on applying chi-square test.

While news articles and, in general, any text streams are usually represented as transient text snapshots, the content of pages often persists over time. Duration of content has certain relation to its semantics and relative importance in a page. Thus, in contrast to typical text data streams, one has to consider three types of content in pages at every time point: static (persisting over time), deleted, and added. Additionally, pages have certain inherent topics that determine the context of their transitory content and that can enhance the mining process.

Web Mining

Web mining is often described as the application of data mining techniques for extracting knowledge from the Web. It is traditionally divided into usage, structure and content mining. Web usage mining identifies the behavior patterns of users visiting Web pages for the purpose of optimizing Web sites. It is usually based on historical data, which is collected during certain time periods for its subsequent analysis (Cooley, Srivastava, & Mobasher, 1997; Kosala & Blockeel, 2000). Web usage mining can show how the users' access to Web sites changes over time. Web structure mining focuses on the link structure and graphical representation of the live Web. There have been, however, several approaches proposed to analyze the evolution of links over time (Amitay, Carmel, Herscovici, Lempel, & Soffer, 2004; Chi et al., 1998; Toyoda & Kitsuregawa, 2003). For example, temporal link analysis was used for detecting trends in page collections (Amitay et al., 2004) or for visualizing evolutions of Web communities (Chi et al., 1998; Toyoda & Kitsuregawa, 2003).

Web content mining uses the content of Web pages for knowledge extraction. Blog related research is probably the most prominent example of Web content mining in which the temporal aspect of pages is considered (Gruhl, Guha, Liben-Nowell, & Tompkins, 2004; Kumar, Novak, Raghavan, & Tomkins, 2003). Blogs help to detect and analyze social structures and social relations as well as provide information on society opinions, hot topics or recent trends. Blogs, however, are a unique media type as they usually contain complete versions of their past content with explicit timestamps provided as well as they are highly personalized and subjective. We believe that a general framework for mining any page types in the past Web is required.

Although most approaches to Web content mining generally neglected the temporal dimension of pages (Cooley et al., 1997; Kosala & Blockeel, 2000), there were, however, several works that investigated the usefulness of data on page histories for knowledge discovery (Arms et al., 2006; Aschenbrenner & Rauber, 2006; Jatowt & Tanaka, 2007; Rauber, Ascenbrenner, & Witvoet, 2002; Yamamoto, Tezuka, Jatowt, & Tanaka, 2007). Rauber et al. (2002) discussed the possibility of analyzing past Web data for identifying changes in Web-related technologies, particularly in the features and characteristics of Web pages, such as a file format, language, size, and so forth. The objective was to create statistics describing Web changes over time. Aschenbrenner and Rauber (2006) surveyed the work that has been done toward mining large portions of Web content with consideration of its temporal aspect. They also provided a general outlook on the potential of mining Web archives. Arms et al. (2006) have reported on building a research library for facilitating study of the Web evolution. This is an ongoing project aiming to build an infrastructure for analysis of massive portions of the data that is stored in Internet Archive. Practical usage of the past Web has been recently demonstrated by Yamamoto et al. (2007), who have proposed an application similar to question answering systems for extracting and combining knowledge from the Web and Web archives. It uses Web archive data for detecting changes in opinions and user knowledge over time.

Mining the content of the past Web is different from the usual Web content mining in several aspects. First, the temporal dimension of content and links in page histories poses new challenges and opportunities for understanding their roles and interrelations in contrast to traditional Web content mining. Second, pages and Web sites should be treated as dynamic objects having certain age, histories, trends, patterns, and so forth. Thus, the notions of a page and its content need to be separated in a way in which the latter one is considered as a transient component occurring in a higher level object, that is, a page. Content has then its own duration of occurrence while the page history is considered as the composition of different content occurring throughout the page's lifetime. Third, there is an issue of missing and incomplete data. In order to obtain satisfactory results, multiple snapshots of the past content of pages have to be found and acquired as well as approximation methods need to be applied for an optimal page history reconstruction.

DATA ACQUISITION AND PREPARATION

Data acquisition and preparation are important steps in the knowledge discovery process. In the mining of the past content of the Web these steps mean the retrieval of data from Web archives and the reconstruction of Web document histories (Jatowt & Tanaka, 2007). The following issues are involved here. First, it is by definition an ex post facto process, as the data is the past content of pages. If one could predict beforehand which Web pages are going to be used, one could simply set up a crawler with a suitable crawling frequency so that page evolution would be captured with a desired precision. However, it is assumed that the user is unable to make such a prediction, and rather that she or he wishes to acquire knowledge in real time using the available, preserved data. Hence, past snapshots of Web pages are gathered in real time from available resources with the aim of reconstructing the past with the highest possible precision. Thus, when talking about crawling in the context of the past Web, we mean querying past Web repositories for the data they contain. Second, because data is scattered in different repositories, it has to be searched for and identified before being used. Therefore, it is necessary to use efficient search and download techniques to locate and gather multiple snapshots of past content with a minimal cost. Due to the large size of data, in practice, usually, only its small portion can be fetched and analyzed locally. Therefore, the focus of this research is on the analysis of the limited amount of data rather than on building a framework for examining the past Web from a macroscopic viewpoint. In addition, there is an issue of the trustworthiness of past content, which is directly related to the trustworthiness of past Web repositories. For example, data obtained from a personal Web repository would normally be less trustworthy than the data collected from a large Web archive containing millions of pages and having a professional maintenance and control. Finally, only fragmentary data can be obtained due to the unpredictable change pattern of the Web and limited resources of archival systems. This calls for employment of efficient techniques for estimation of actual content that pages had in the past.

Collecting Snapshots

***Definition 1*:** Past page snapshot is a copy of page content that was published in the Web at a given time point in the past. The timestamp of the snapshot indicates the date when it was captured.

As mentioned above, because of resource limitations, Web archives contain only fragmentary past data. As a general attempt to alleviate this problem a kind of meta-archive approach (Jatowt, Kawai, Nakamura, Kidawara, & Tanaka, 2006) can be used to maximize past Web coverage and consequently to increase the precision of history reconstruction. This approach presumes communication with several past Web repositories at the same time. An intermediary module is required between these repositories and the local system to translate queries into the format required for each repository. After receiving a request for a page history, the module queries the repositories about their data. The repositories should then send a list of stored page snapshots with their metadata so that a fetching policy can be determined.

The optimal strategy would be first to check the signatures (checksums) of snapshots, if they are provided, in order to detect the ones that actually contain content changes from among all data provided by the cooperating repositories. This would prevent downloading identical page snapshots from different repositories, thereby maximizing fetching efficiency[2]. However, currently, Web archives do not provide such information. Instead, some repositories, such as the Internet Archive, provide lists of page snapshots that have any changed content when compared to the neighboring snapshots. By utilizing this information, only the snapshots with content changes inside archives would be fetched. In general, the efficiency of the data collection would depend on the metadata that is provided in past Web repositories.

Such a meta-archive approach would provide a unified interface to the history of the Web, making the data acquisition process less dependent on the resources of single Web archives. However, as Web archive interfaces are diverse, different data acquisition methods would have to be used. In addition, we make an assumption here that the URLs of pages remain the same over time, although, in practice, they may change even though the content of pages remains almost the same.

McCown and Nelson (2006) and McCown Smith, and Nelson (2006) have recently measured the persistence and availability of page copies inside the repositories of major search engines and the Internet Archive. The objective was to estimate the possibility and to provide methodology for reproducing the latest versions of Web sites in case of the loss of Web data.

Reconstruction of Page Histories

***Definition 2*:** Page history reconstruction is the process of reproducing the past content of a page using available snapshots for obtaining the continuous representation of page history.

***Definition 3*:** Optimal page history reconstruction is a reconstruction which accurately reproduces page history; that is, the errors resulting from such a reconstruction are equal to zero. Having determined an optimal page history, it is possible to recreate page content for any time point in the past that shows the actual content the page had at that time.

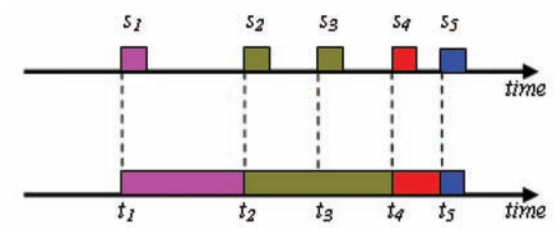
However, unless the page was unchanging, it has been crawled continuously or the implicit information about its past changes is provided, there will be usually some error involved in the history reconstruction. Only for certain types of pages, for example wikis, complete past data is available as the preservation of their versions is usually automatically done. In case of such pages, the reconstruction error would be equal to zero as all past changes can be derived from available page versions. In addition, some pages may contain temporal annotations in their present content that can be used to enhance the history reconstruction. For example, blogs often provide timestamps of content insertion. Nevertheless, for the majority of hypertexts, usually, neither implicit version management nor temporal annotations are provided.

We propose a simple approach for the page history reconstruction (Jatowt & Tanaka, 2007). First, collected snapshots are chronologically ordered according to their timestamps. If past snapshots are not associated with any temporal metadata then they cannot be directly included in the ordered sequence of past snapshots without a prior determination of their timestamps. For example, Yahoo! search engine provides cached snapshots of Web pages but it does not attach any timestamps to them. Estimating a timestamp of a snapshot could be possibly done by comparing similarities between its content and the content of other snapshots with known timestamps.

Second, every previous page snapshot is considered to represent the actual state of page content for the time period until the next page snapshot in the sequence. For example, suppose that five snapshots have been collected, s_1, s_2, s_3, s_4 and s_5, with timestamps, t_1, t_2, t_3, t_4 and t_5, where $t_1<t_2<t_3<t_4<t_5$ (Figure 1). Let us also suppose that snapshots s_2 and s_3 are exactly same. After the simple approximation, the page content is assumed to be the same as that in s_1 during the period $[t_1, t_2)$, the same as that in s_2 during $[t_2, t_4)$ and equal to s_4 in $[t_4, t_5)$. The reconstructed page history is then represented as a minimal sequence of 2-tuples containing different page versions and their starting dates ($\{(s_1, t_1), (s_2, t_2), (s_4, t_4)\}$ in the above case).

Page history reconstruction could be improved by considering additional information, for example, by analyzing changes in other pages belonging to the same site. Also, using the results of the temporal analysis of pages, especially their updating patterns, could make the reconstruction

Figure 1. Example of page history reconstruction



more accurate. Finally, historical snapshots of mirror pages, if there are any, could be utilized.

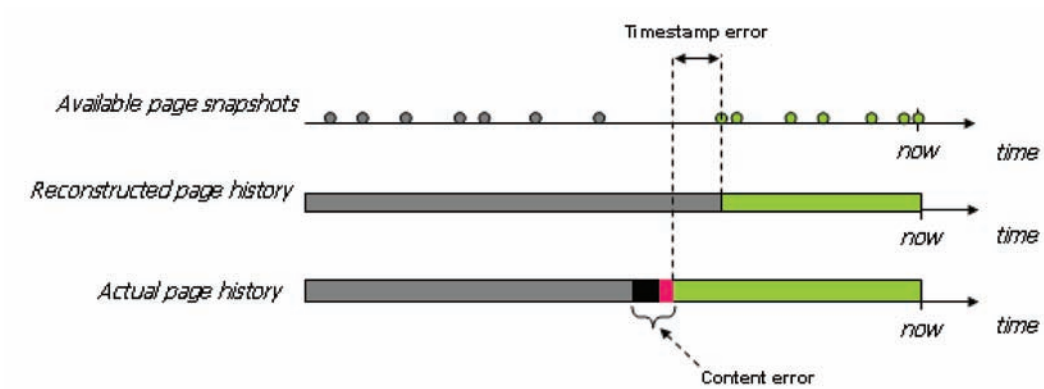
History Reconstruction Error

Usually, it is difficult to determine an accurate page history that would reflect the actual page content as it was at any arbitrary time point in the past unless the complete set of actual page versions is provided, for example, by a page author. Hence, mining the content of the past Web will typically be carried out using incomplete data with varying levels of precision and trust. It is thus necessary to consider the issue of missing data.

We can distinguish two types of errors in the page history reconstruction assuming that the page crawling was independent from the page update pattern (Jatowt & Tanaka, 2007). The first one, which we call a content error, is caused by uncertainty related to the content that appeared on a page in the past. Consider two retrieved past versions of the page (v_L and v_R) captured at time points t_L and t_R ($t_L<t_R$). The probability, $P(v_i)$, that there is any version v_i satisfying $t_L<t_i<t_R$ and containing any content different from that in v_L and v_R depends on many factors such as the length of the period from t_L to t_R, page type, content difference between v_L and v_R or the average change degree of the page. Intuitively, the shorter the time distance between the page snapshots and the more even their distribution over time are, the lower is the average probability of any transient, undetected content occurring in the page.

The second error type, which we call a timestamp error, is due to uncertainty in estimating the dates of content changes. The timestamp error, like the content one, depends also on the number of acquired past snapshots and their distribution over time. Figure 2 illustrates both error types. The top timeline shows available past snapshots of a page. For simplicity, let us assume that the page snapshots are empty (i.e., blank page) or they contain only one content element, be it a picture or a text snippet. Those snapshots that contain the element are marked by a green color, while the empty snapshots are marked by a grey color. After reconstructing the history of the page (the middle timeline) and comparing it with the bottom timeline that shows the actual page history,

Figure 2. Content and timestamp errors in the reconstructed page history



we can see that the reconstructed history contains both content and timestamp errors.

Site History Reconstruction

Pages usually belong to larger information units, or Web sites. Reconstructing histories of a Web site requires detecting the changes in site's topology over time and retrieving past content of pages that belonged to the site. As an input the starting page (e.g., the top page of a site), time frame *T*, and the depth *D* (i.e., the number of hops from the starting page) need to be specified.

The data accumulation system collects all available snapshots of the starting page that have timestamps within *T*. It then searches their content for any links to other pages on the site (i.e., pages having the same domain name). For each such a link, it collects available, previous snapshots of the linked page that have timestamps within *T*. These snapshots are then searched in the same way for links to other pages on the same site. The entire process is repeated until the specified depth *D*. In general, a page is considered to belong to the site's history if, during the time frame *T*, it was linked from another page belonging to the site at that time and if it was located a smaller number of hops from the starting page than the specified depth *D*. Intuitively, the number of page snapshots collected at the initial steps of the crawl (few hops from the starting page) has an influence on detecting pages at later steps. This is because pages may remain undiscovered if the links pointing to them occurred only in the undetected, transient content of other pages in the site. We call the error caused by the missing links a page error.

A site history is represented as a set of reconstructed page histories that belonged to the site in the past. The precision of the site reconstruction can be enhanced by utilizing topological information preserved in the past content of site-map pages if they existed. Many Web sites include site-map pages designed to help users navigate sites. Utilizing the site-map page history could help to detect transient pages that have not been discovered by the above crawling approach and thus minimize the page error, as well as it could help to more precisely determine the actual time points of page creation and deletion within sites (timestamp error for whole pages).

PAGE TEMPORAL ANALYSIS

Page temporal analysis is the study of page content over time. Its results should be particularly useful if pages are associated with specific objects such as companies, institutions, persons or other entities. Understanding the temporal characteristics of a page over a long time frame can shed light on the associated objects or on other information appearing on the page. For example, if certain content occurred for a long time on a page which was updated frequently and regularly, then we can treat the content in a different way or with a different level of trust than if it occurred on a page that was generally static or even obsolete. A similar idea applies to a page devoted to a specific topic vs. a page that deals with many varying topics throughout its history. In other words, page temporal analysis can be used to find temporal context for information from the past. Having determined the context, it is possible to better understand the connection between Web pages and their transient content as well as to identify pages most relevant to target objects.

When mining the histories of Web pages for real-world information, we must distinguish between the valid time and transaction time of events, both of which are often used in the database research. The valid time of an event is considered as the time at which the event occurred in the real world. The transaction time is the time at which the information about the event was stored in a database or, in our case, added to a certain Web page. It can be estimated by searching the page history for the earliest occurrence of the content related to the event (Jatowt, Kawai, & Tanaka, 2007). The valid time, on the other hand, can be detected from temporal expressions appearing in the content of past page versions. This would require using special taggers and resolvers of temporal expressions in text. In addition, techniques such as the one described by Bar-Yossef, Broder, Kumar, and Tomkins (2004) could be applied for classifying page content as current or obsolete.

Next, we present a simple framework for analyzing page histories. After page history reconstruction, HTML tags, scripting code, and multimedia objects are removed from available page versions. Vector representation is then created for textual content of the past versions using a weighting method such as a term frequency. Let $\boldsymbol{V}=(\boldsymbol{v}_1,\boldsymbol{v}_2,..,\boldsymbol{v}_n)$ denote the sequence of vectors of the consecutive page versions, where $\boldsymbol{v}_j$ is the vector of a page version at time point t_j ($t_1 \le t_j \le t_n$). Next, the contents of the neighboring versions are compared with each other using a change detection algorithm such as *diff*. Added content appearing in the page's history is thereby found. All changes in each version are then grouped together and represented as a change vector. Consequently, a sequence of change vectors is obtained, $\boldsymbol{C}=(\boldsymbol{c}_{(1,2)},\boldsymbol{c}_{(2,3)},..,\boldsymbol{c}_{(n-1,n)})$, where $\boldsymbol{c}_{(j,j+1)}$ is a vector for an added-type change obtained by comparing page versions $\boldsymbol{v}_j$ and $\boldsymbol{v}_{j+1}$.

The content of past versions can be compared against any query containing terms describing given topic of interest. In order to do so, at each selected time point, a query vector, $\boldsymbol{q}_j$, is constructed by assigning uniform weights to all query terms. The sequence of query vectors is denoted as $\boldsymbol{Q}=(\boldsymbol{q}_1,\boldsymbol{q}_2,..,\boldsymbol{q}_n)$. Different values can be assigned to $\boldsymbol{Q}$ at different time points to reflect changes in the chosen topic of interest. Otherwise, the query vector is made static by having the same content at all times. To measure the relationship of past page content to the query topic, the similarity between $\boldsymbol{V}$ and $\boldsymbol{Q}$ is calculated using a cosine similarity measure. In result, the sequence of similarities is obtained: $\boldsymbol{sim}(\boldsymbol{V},\boldsymbol{Q})=(sim(\boldsymbol{v}_1,\boldsymbol{q}_1),sim(\boldsymbol{v}_2,\boldsymbol{q}_2),..,sim(\boldsymbol{v}_n,\boldsymbol{q}_n))$, where $sim(\boldsymbol{v}_j,\boldsymbol{q}_j)$ is the cosine similarity between the vector of past version $\boldsymbol{v}_j$ and query vector $\boldsymbol{q}_j$ (Figure 3). Similarly, the sequence of similarities between the vectors of the changes and the query vectors is calculated, $\boldsymbol{sim}(\boldsymbol{C},\boldsymbol{Q})=(sim(\boldsymbol{c}_{(1,2)},\boldsymbol{q}_2),sim(\boldsymbol{c}_{(2,3)},\boldsymbol{q}_3),..,sim(\boldsymbol{c}_{(n-1,n)},\boldsymbol{q}_n))$, where $sim(\boldsymbol{c}_{(j,j+1)},\boldsymbol{q}_{j+1})$ is the cosine similarity between change vector $\boldsymbol{c}_{(j,j+1)}$ and $\boldsymbol{q}_{j+1}$ (Figure 4).

First, a change frequency can be defined (Equation 1).

$$CF = \frac{fc}{n} \tag{1}$$

Here, *fc* is the number of non-zero elements in $\boldsymbol{C}$. Another measure called a change degree indicates the average change size of a page (*size(a)* denotes the size of element *a*).

Figure 3. Similarity calculation between the sequence of version vectors and the sequence of query vectors

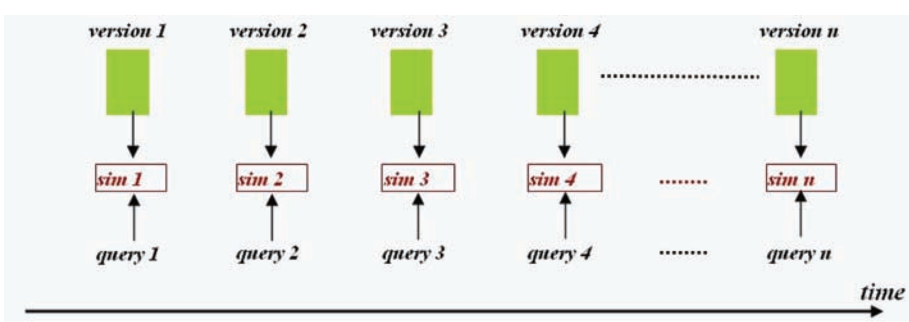
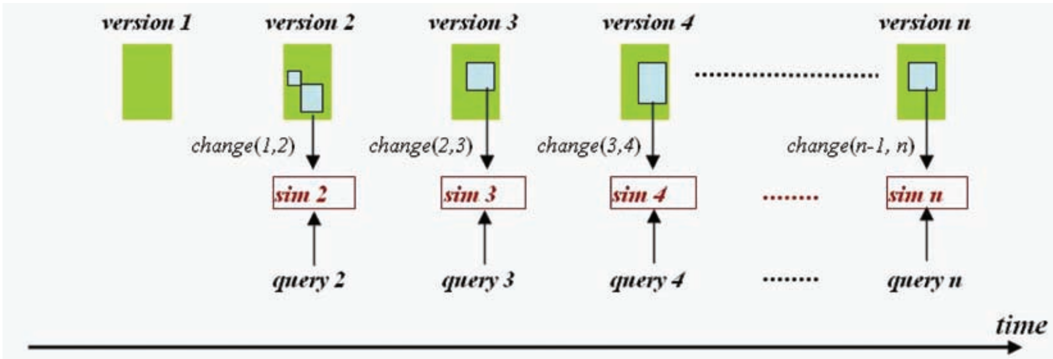


Figure 4. Similarity calculation between the sequences of change and query vectors; changes are depicted as small rectangles inside page versions



$$CD = \frac{\sum_{j=1}^{n} \frac{size\left(c_{(j,j+1)}\right)}{size\left(v_j\right)}}{n} \quad (2)$$

Besides these simple measures, the long-term relevance of a page to the query topic can be calculated. It is expressed as the weighted average of the elements of ***sim***(***V***,***Q***) by taking into account the duration of page content over time (Equation 3).

$$TR = \frac{1}{\sum_{j=1}^{n}\left(\alpha_{j+1} * \left(t_{j+1} - t_j\right)\right)} \sum_{j=1}^{n}\left[\alpha_{j+1} * sim\left(v_j, q_j\right) * \left(t_{j+1} - t_j\right)\right] \quad (3)$$

A page is considered relevant if its content overlaps with the sequence of query vectors during a large portion of a chosen time period. Using this approach, we can estimate the degree of page relevance to any topic within a given time frame. As the recent content is often likely to be more important, Equation 3 is adjusted by applying a weighting scheme depending on the age of page versions.

$$\alpha_j = e^{-\lambda\left(t_{now} - t_j\right)} \quad (4)$$

In addition, the long-term topic stability of a page can be computed by detecting the average similarity between consecutive past versions over time (Equation 5).

$$TS = \frac{1}{\sum_{j=1}^{n}\left(\alpha_{j+1} * \left(t_{j+1} - t_j\right)\right)} \sum_{j=1}^{n}\left[\alpha_{j+1} * sim\left(v_j, v_{j+1}\right) * \left(t_{j+1} - t_j\right)\right] \quad (5)$$

The long-term relevance and long-term topic stability are calculated considering the whole page content in the past, including the static content (the content that did not change between consecutive page versions). In contrast, we can compute a measure showing the degree of page updating based on the amount of changed content over time that is related to the query (Equation 6).

$$TRC = \frac{1}{\sum_{j=1}^{n}\left(\alpha_{j+1} * \left(t_{j+1} - t_j\right)\right)} \sum_{j=1}^{n}\left[\alpha_{j+1} * \frac{sim\left(c_{(j,j+1)}, q_{j+1}\right)}{t_{j+1} - t_j}\right] \quad (6)$$

A combination of different measures can also be used. For example, the measure of the temporal quality of a page is based both on the relevance of the changed content over time to the query topic and on the size of the changes:

$$TA = \frac{1}{\sum_{j=1}^{n} \alpha_{j+1}} \sum_{j=1}^{n} \left[\alpha_{j+1} * \frac{sim\left(c_{(j,j+1)}, q_{j+1}\right)}{\left(t_{j+1} - t_j\right)} * \frac{size\left(c_{(j,j+1)}\right)}{size\left(v_j\right)} \right] \tag{7}$$

According to this measure, a page is more attractive from the viewpoint of the query topic if its changes were relevant to that topic and if they were relatively large. Small changes are usually less likely to be attractive than large ones. Additionally, the temporal quality of the page is higher if the page was modified often in the past. In general, the greater the number and the larger the size of related changes that occurred within short time periods, the higher is the attractiveness of the page. The page temporal quality can be used to identify candidate pages for mining. Naturally, the precision of results depends on the amount and characteristics of the input data that is on the size of errors resulting from the history reconstruction process.

Finally, the trend of page relevance to query can be measured by fitting a regression line to the historical plot of the similarity between page content and query vectors. This allows for estimating the long-term change direction of the page relevance. A rising trend would mean that the page content becomes closer to the query topic.

TEMPORAL SUMMARIZATION

Document summarization is a well-known text mining task. Automatic summarization of Web pages aims at creating compact versions of Web documents that would contain only the most important content. Traditionally, summaries were constructed from static snapshots of Web pages (Berger & Mittal, 2000; Buyukkokten, Garcia-Molina, & Paepcke, 2001; Delort, Bouchon-Meunier, & Rifqi, 2003). However, as pages are dynamic, their content is often changed. In this section, we briefly describe the concept of temporal summarization which is the extension of the traditional summarization task into the time dimension (Jatowt & Ishizuka, 2004a, 2004b; Jatowt & Ishizuka, 2006). It is used to summarize temporal versions of Web documents in order to provide information on important content, hot topics or popular events described in pages over time. Web users are often overloaded with large amounts of data. Automatic temporal summarization would help them in discovering salient information from parts of the past Web such as histories of pages or their collections.

Following the classical division of document summarization research, two types of temporal summarization can be distinguished: single- and multi-page temporal summarization. Single-page temporal summarization attempts at capturing salient content that occurred on a page over a certain time period. The summary should thus reveal main page topics during a predefined time frame. On the other hand, in multi-page temporal summarization, multiple snapshots of a topical collection of pages are analyzed for changes over time. The summary should reveal important events or concepts that occurred in a given topical area over time. The key issue in this type of summarization is gathering pages which are up-to-date and related to the target topic so that a reliable and consistent topical collection can be synthesized. Below we discuss the multi-page temporal summarization in more detail.

Multi-page Temporal Summarization

Web collection for multi-page temporal summarization can be obtained in several ways; for example, it can be created from a user-provided set of related Web documents that she or he usually revisits for fresh information or it could be downloaded from existing Web directories. While Web directories group topically related Web documents, they provide only a limited number of categories. In a more flexible way, the collection could be synthesized by filtering search engine results based on the analysis of their temporal characteristics such as long-term relevance or temporal quality. Naturally, duplicate pages should be discarded in this process. After the initial set of topically related pages is ready, it is extended in time by reconstructing page histories for a chosen time period.

In the following step, textual data is extracted from the accumulated past versions. Then, an extractive type summarization algorithm is used to detect useful sentences for constructing a summary. First, so-called long-term scores are calculated for all terms by comparing terms' distributions in documents over time. These scores are later used to identify important sentences to be included in the summary. We propose two approaches for the long-term score calculation. One uses a sliding window that is sequentially moving through the temporal collection to search for bursts of terms in added or deleted content in the collection (Jatowt & Ishizuka, 2004a). Any terms that were added to or deleted from many pages in the collection at around the same time have high values of the long-term scores. Another approach to the calculation of long-term scores is based on the analysis of term frequency plots. The parameters of term frequency plots such as variance, slope of a regression line and intercept are calculated and compared for identification of salient terms (Jatowt & Ishizuka, 2004b). The terms with outstanding features, such as the ones with upward trends or high variance would be then scored highly. More details on the both term scoring methods can be found in Jatowt and Ishizuka (2004a, 2004b) and Jatowt and Ishizuka (2006).

After the long-term scores of terms are computed, the summarization system searches for sentences suitable for constructing the summary. Sentence selection is based on analyzing plots of the terms that have the highest long-term scores. The plots are examined to identify intervals with the closest match to the shape of an ideal plot. For example, the system may search for a time period where the frequency plot of a term has a shape that most resembles the ideal shape in which the plot suddenly increases and remains at a relatively high level over a long time. Such a plot shape may indicate the onset of an important event represented by the term. Thus, sentences containing the term are extracted from the collection within the selected time period. The system tries here also to maximize the number of different terms with top long-term scores in the selected sentences. Lastly, after the predefined number of sentences is extracted, the system orders them based on their timestamps and relative locations in their original page versions. Each sentence may also have a link to its page version added to be used in case users wish to obtain more details. Furthermore, a number of additional heuristics may be used to increase the coherence and readability of the final summary, for example, by inserting explanatory content or by modifying or reordering the selected sentences.

Discovering Object Histories

Related to temporal summarization is object history reconstruction. Objects are defined here as higher level concepts and abstractions that represent persons, institutions, ideas, organizations, and so forth. Objects can be represented by groups of related words or n-grams. Thus, object histories could be modeled using the histories of the representative terms and their inter-relationships. Time points of changes and the durations of terms' occurrences on pages would provide clues about the timing of events related to objects represented by these terms.

Object's history should be most accurate if it has been derived from a source that directly represents the object (e.g., company homepage, personal blog). The relationship between the page and objects discussed on this page can help in understanding the content related to the objects. In general, contextual information about objects can be derived from the characteristics and topical scopes of analysed pages. Furthermore, the co-occurrence of similar information among different resources increases its trustworthiness as well as helps to better determine the starting and ending points of events. The larger is the number of different data sources devoted to an object, the more reliable and accurate the discovered knowledge should be.

A possible example of object history reconstruction is an automatic creation of personal bibliographies or their parts. There is much personal data published on the Web. For example, employment data is sometimes reported on company or personal Web pages (e.g., on blogs), and other personal information can be found. This information could be collected and processed to construct biography parts.

By analyzing semantic and temporal clues derived from past Web content it could be possible to improve the detection process by employment of various heuristics. For example, the temporal information derived from the chronological ordering of events reported on past pages might help in understanding the events and may provide hints for a further search. One such possible heuristic is the detection of person's employment dates. Suppose that at some time point a person's name was removed from the page of some laboratory. Then, the system could search for the page of another institution that reported hiring the person at around that time. Note, however, that there might be certain latency between the actual events and their reports in the Web (i.e., valid and transaction times).

BROWSING PAST WEB

Apart from mining the content of the past Web, it is important to have a tool that allows for viewing data in detail, for example, in order to manually inspect the data from the viewpoint of discovered results. Such a tool should be intuitive, easy to use and possibly resemble similar applications used for the current Web. In this section, we describe the framework for a past Web browser (Jatowt et al., 2006) that supports browsing and navigation in the past Web. A browser built using this framework would be a client-side system that downloads, in a real time, past page snapshots from Web archives for their customized presentation. Such a browser would enable viewing the evolution of pages and browsing the past structures of the Web.

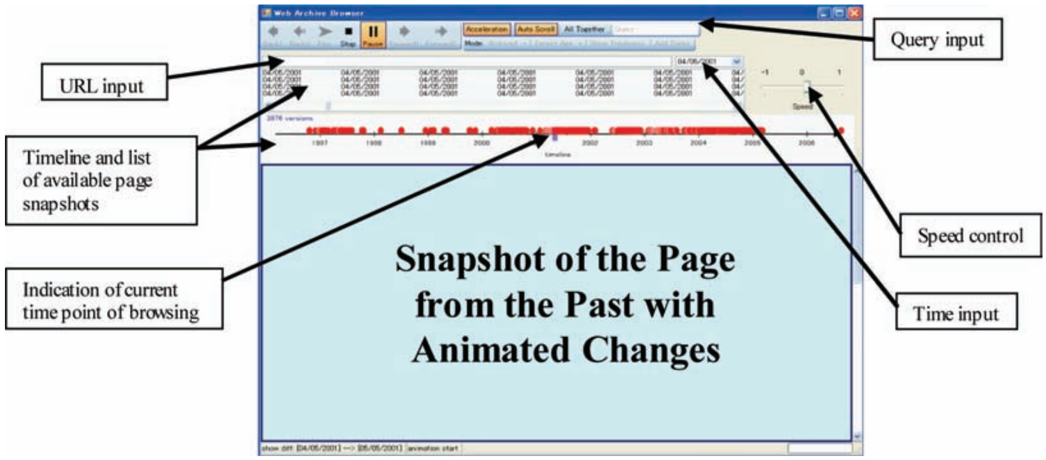
The proposed browser integrates histories of Web pages with their present versions and has a standard functionality of a traditional browser for the live Web. Consequently, browsing the live and past Web can be done almost at the same time. Thanks to this, users browsing the live Web can access the histories of viewed pages in case they need to find some content from the past, observe the page evolution or, simply, to access the latest page snapshot if the present page cannot be properly viewed due to any reasons such as a server failure.

Browsing

Two basic types of browsing are distinguished here: vertical and horizontal. The former means browsing different pages around a certain point of time by following links, while the latter means viewing past snapshots of a single page along the time direction, that is, browsing the past Web in a horizontal direction. A mixture of both kinds of browsing enables users to traverse the past Web both in time and space dimensions.

To start the horizontal browsing, the URL of a page and a point of time have to be provided. The browser fetches a page snapshot whose timestamp is closest to the user-provided time point. Next, the browser automatically downloads the following page snapshots and displays them in a passive manner. This type of viewing results in a minimum user interaction, because page snapshots are presented to the user one by one, like in a slideshow, with a certain delay predefined for each snapshot. As when watching a video, the user can pause or stop the motion, enabling the

Figure 5. Past Web browser



detailed examination of the currently presented snapshot or following its links. Besides, the user may enter a new date or a different URL to make a jump to another snapshot. In addition, a timeline is automatically constructed and displayed above the page content (Figure 5). It shows the distribution of page snapshots indicating the points of time for which snapshots are available. The currently viewed snapshot is indicated in the timeline by a blue rectangle. The information provided by the timeline prevents users from being lost in the hyperspace of the past Web by informing them about the current time point of browsing and the overall distribution of snapshots. At the same time, it is also a navigation tool thanks to which users can make a jump to any page snapshot simply by clicking on any point on the timeline. The timeline can be also zoomed to provide the more detailed view. Besides the timeline, the clickable list of all page snapshots together with their timestamps is also displayed (Figure 5).

Horizontal browsing is enhanced by a page presentation in which content changes are detected and emphasized. Keeping in mind the large size of the past Web, with lots of static, redundant data, the most effective method for horizontal browsing seems to be the one using change visualization. We think that changed data is the most important in page histories and that enhancing horizontal browsing with the change indication can portray page evolution and, in addition, help reduce the amount of browsing needed, especially in the case of static (unchanging) pages. Both content additions and deletions between neighboring page snapshots are then detected using a change detection algorithm and emphasized to indicate the content variance in pages. This enables users to spot not only the added content in consecutive page snapshots but also to identify the removed one. However, effectively showing both change types in a combined view on a single page would be difficult, especially in the case of large and overlapping changes. Thus, we propose using animation effects in order to efficiently show both change types. The change presentation algorithm displays the changes gradually, in the form of animation. Content that was deleted in the page history first blinks for a certain time period and then disappears, followed by the inserted content that first appears on the page and then blinks for a short time. Page snapshots are processed in this way line by line from left to right and from top to bottom. Content that was static between consecutive snapshots remains displayed on the page. After the page transition between two consecutive page snapshots is completed, the browser waits a predefined time period with the latter page snapshot displayed and then it proceeds to analyze the following page snapshot. The user can control the speed of the presentation using a slider provided in the top-right corner of the browser (Figure 5). Besides, as sometimes page snapshots may be too large to be shown at once, a user can choose between the automatic scrolling option and the option of displaying only the top part of page content.

Animation of changed content results in a smooth transition between sequential page snapshots. By animating changes user's attention is drawn to the changed content. In addition, changes are also highlighted by different colors to increase their visibility. However, for simplicity, in the case when the amount of change in a page snapshot is higher than the predefined threshold, no animation is done and changes are emphasized using only different background colors.

The user can stop the horizontal browsing at any time by pressing stop or pause buttons in a similar way to video players. Next, she or he can view the currently displayed page snapshot in detail or follow any link. Upon clicking on a link, the browser loads the snapshot of the linked page that is closest in time to the one being currently viewed and, after a short time period, it automatically starts the horizontal browsing on the new page.

The browser is also equipped with two back and two forward buttons to enable navigation in the space as well as in time dimensions. Besides, there is an additional navigation mechanism provided (automatic jumping facility). It enables the browser to skip periods in the page history during which the content did not change or did not change much. When this functionality is switched on, the browser displays only those page snapshots that contain more than a certain amount of change. This enables faster viewing of page evolution by omitting changeless periods.

Finally, a search option enables users to specify queries for filtering changes. If a query is issued, only the changes that contain the query terms are animated. Other changes are treated as static content and thus are not animated. This browsing style results in the filtered view of page history. Users can thus observe page histories from the viewpoint of topics that they are interested in. For example, a newswire page history could be browsed for information about "Iraq" or "presidential election" over selected time periods.

Related Research and Future Work

Visual Knowledge Builder (VKB) (Shipman & Hsieh, 2000) was an early proposal of an application that provides a mechanism for enabling history navigation in private hypertexts. The objective was to allow users to play back the history of a hypertext for witnessing the authoring of hypertexts, understanding the context of their creation and authors' writing styles. The browser interface had some similarity to VCR players.

WERA[3] (Web ARchive Access) and Wayback Machine[4] are applications for accessing Web archives. WERA supports time and URL input for specifying a particular page snapshot. There is a timeline provided showing the available page snapshots and indicating the currently browsed one. Users can view the consecutive page snapshots by clicking arrows in the timeline.

Wayback Machine is a Web-based interface to the Internet Archive. After a user inputs a URL, optionally with a time period specified, links to the available page snapshots are listed on the "directory" page. The user can then click on any snapshot to view its content or follow its links if the linked snapshots are also stored in the archive. The directory page indicates also page snapshots that contain changes by marking them with asterisks. Horizontal browsing using Wayback Machine is difficult, as users need to access the directory page each time if they wish to view other snapshots.

Both the Wayback Machine and WERA are server-side applications designed for single Web archives. Our proposed browser is a higher-level, client-side application that allows for the usage of multiple past Web repositories at the same time, thus, enabling browsing of the past Web rather than browsing single archives. Browsing the past Web is also facilitated by combining passive, automatic page viewing together with a change presentation. The framework has also functions that minimize the user effort and time required to find specific information in the past snapshots of pages. In addition, navigation mechanisms are provided to enable traversal of the link structure of the past Web. Testing the browser built on the proposed framework demonstrated its usefulness (Jatowt et al., 2006). Users were able to move freely in the past Web, find desired information and relatively easily obtain an overall view of pages' evolution.

In a multi-authoring area, an interesting application has been recently proposed for effective visualization of histories of wiki pages (Viégas, Wattenbeg, & Dave, 2004). It allows viewing contributions of different authors and their persistence over time as demonstrated on the example of Wikipedia pages[5].

There are several possible directions for expanding the proposed framework. For example, location-based browsing would allow a user to select a certain area on a page and then view its evolution over time, provided that the structure of the page did not change substantially. This would limit the presentation to only those changes that occurred in the selected area, for example, in the sports section of a newswire page. Next, links on visited snapshots could be annotated with timestamps of page snapshots that will be accessed when following these links. Thanks to it, a user would know how much time jump she or he is going to experience upon clicking on a certain link. Lastly, a comparative past web browser could enable comparison of histories of two or more pages highlighting their common or similar parts.

CONCLUSION

The Web has become nowadays a major means of communication and an important information repository. Due to its dynamic, ever evolving character, much of the content regularly disappears from the live Web and can only be accessed through Web archival repositories. Knowledge discovery from past Web is a challenging and promising research direction. Mining the content of the past Web differs from traditional Web content mining and thus requires a novel approach. In this chapter, we have described several issues related to mining data in Web archives. First, we provided the outlook on the data collection and preparation steps and emphasized their importance. Next, we demonstrated the methodology for determining page temporal characteristics as a source of contextual information for describing pages and their transient content. Then, data summarization and object history detection were described as examples of mining tasks on the past Web. Finally, we proposed the application for browsing and navigation in the past Web.

ACKNOWLEDGMENT

This research was supported by the MEXT Grant-in-Aid for Scientific Research in Priority Areas entitled: Content Fusion and Seamless Search for Information Explosion (#18049041, Representative Katsumi Tanaka), and by the Informatics Education and Research Center for Knowledge-Circulating Society (Project Leader: Katsumi Tanaka, MEXT Global COE Program, Kyoto University) as well as by the MEXT Grant-in-Aid for Young Scientists B entitled: Information Retrieval and Mining in Web Archives (#18700111).

REFERENCES

Allan, J. (Ed.). (2002). *Topic detection and tracking: Event-based information organization.* Norwell, MA, USA: Kluwer Academic.

Allan, J., Gupta, R., & Khandelwal, V. (2001). Temporal summaries of news topics. In *Proceedings of the 24th Annual Conference on Research and Development in Information Retrieval*, (pp. 10-18). New Orleans, LA, USA: ACM Press.

Amitay, E., Carmel, D., Herscovici, M., Lempel, R., & Soffer, A. (2004). Trend detection through temporal link analysis. *Journal of the American Society for Information Science and Technology*, *55*(14), 1270–1281. doi:10.1002/asi.20082

Arms, W. Y., Aya, S., Dmitriev, P., Kot, B. J., Mitchell, R., & Walle, L. (2006). Building a research library for the history of the Web. In *Proceedings of the Joint Conference on Digital Libraries*, (pp. 95-102). Chapel Hill, NC, USA: ACM Press.

Arvidson, A., Persson, K., & Mannerheim, J. (2000). The Kulturarw3 project—the Royal Swedish Web Archive—an example of "complete" collection of Web pages. In *Proceedings of the 66th IFLA Council and General Conference*, Jerusalem, Israel.

Aschenbrenner, A., & Rauber, A. (2006). Mining Web collections. In J. Masanes (Ed.), *Web archiving* (pp. 153-174). Berlin, Heidelberg, Germany: Springer-Verlag.

Bar-Yossef, Z., Broder, A. Z., Kumar, R., & Tomkins, A. (2004). Sic transit gloria telae: Towards an understanding of the Web's decay. In *Proceedings of the 13th International Conference on World Wide Web*, (pp. 328-337). New York: ACM Press.

Berger, A. L., & Mittal, V. O. (2000). OCELOT: A system for summarizing Web pages. In *Proceedings of the 23rd Conference on Research and Development in Information Retrieval*, (pp. 144-151). Athens, Greece: ACM Press.

Brewington, E. B., & Cybenko, G. (2000). How dynamic is the Web? In *Proceedings of the 9th International World Wide Web Conference*, (pp. 257-276). Amsterdam, the Netherlands: ACM Press.

Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001). Seeing the whole in parts: Text summarization for Web browsing on handheld devices. In *Proceedings of the 10th International World Wide Web Conference*, (pp. 652-662). Hong Kong, SAR, China: ACM Press.

Chi, E. H., Pitkow, J., Mackinlay, J., Pirolli, P., Gossweiler, R., & Card, S. K. (1998). Visualizing the evolution of Web ecologies. In *Proceedings of Conference on Human Factors in Computing Systems*, (pp. 400-407), Los Angeles: ACM Press.

Cho, J., & Garcia-Molina, H. (2000). The evolution of the Web and implications for an incremental crawler. In *Proceedings of the 26th International Conference on Very Large Databases*, (pp. 200-209). Cairo, Egypt: ACM Press.

Cho, J., & Garcia-Molina, H. (2003). Estimating frequency of change. *Transactions on Internet Technology*, *3*(3), 256–290. doi:10.1145/857166.857170

Cooley, R. Srivastava, J., & Mobasher, B. (1997). Web mining: Information and pattern discovery on the World Wide Web. In *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence*, (p. 558). IEEE Press.

Delort, J.-Y., Bouchon-Meunier, B., & Rifqi, M. (2003). Enhanced Web document summarization using hyperlinks. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*, (pp. 208-215). Nottingham, UK: ACM Press.

Fetterly, D., Manasse, M., Najork, M., & Wiener, J. (2003). A large-scale study of the evolution of Web pages. In *Proceedings of the 12th International World Wide Web Conference,* (pp. 669-678). Budapest, Hungary: ACM Press.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the 13th International World Wide Web Conference,* (pp. 491-501). New York: ACM Press.

Hallgrimsson, Þ., & Bang, S. (2003). Nordic Web Archive. In *Proceedings of the 3rd Workshop on Web Archives in conjunction with the 7th European Conference on Research and Advanced Technologies in Digital Archives*, Trondheim, Norway. Springer-Verlag.

Jatowt, A., & Ishizuka, M. (2004a). Summarization of dynamic content in Web collections. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases,* (pp. 245-254). Pisa, Italy: Springer-Verlag.

Jatowt, A., & Ishizuka, M. (2004b). Temporal Web page summarization. In *Proceedings of the 5th Web Information Systems Engineering Conference*, (pp. 303-312). Brisbane, Australia: Springer-Verlag.

Jatowt, A., & Ishizuka, M. (2006). Temporal multi-page summarization. [IOS Press.]. *Web Intelligence and Agent Systems: An International Journal*, *4*(2), 163–180.

Jatowt, A., Kawai, Y., Nakamura, S., Kidawara, Y., & Tanaka, K. (2006). Journey to the past: Proposal for a past Web browser. In *Proceedings of the 17th Conference on Hypertext and Hypermedia*, (pp. 134-144). Odense, Denmark: ACM Press.

Jatowt, A., Kawai, Y., & Tanaka, K. (2007). Detecting age of page content. In *Proceedings of the 9th ACM International Workshop on Web Information and Data Management*. Lisbon, Portugal: ACM Press.

Jatowt, A., & Tanaka, K. (2007). Towards mining past content of Web pages. *New Review of Hypermedia and Multimedia* [Taylor and Francis.]. *Special Issue on Web Archiving*, *13*(1), 77–86.

Kleinberg, J. M. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, *7*(4), 373–397. doi:10.1023/A:1024940629314

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations*, *2*(1), 1–15. doi:10.1145/360402.360406

Kumar, R., Novak, P., Raghavan, S., & Tomkins, A. (2003). On the bursty evolution of Blogspace. In *Proceedings of the 12th International World Wide Web Conference*. Budapest, Hungary: ACM Press.

Li, Z., Wang, B., Li, M., & Ma, W.-Y. (2005). A probabilistic model for retrospective news event detection. In *Proceedings of the 28th Annual International Conference on Research and Development in Information Retrieval*, (pp. 106-113). Salvador, Brazil: ACM Press.

McCown, F., & Nelson, M. (2006). Evaluation of crawling policies for a Web-repository crawler. In *Proceedings of the 17th Conference on Hypertext and Hypermedia*, (pp. 157-168). Odense, Denmark: ACM Press.

McCown, F., Smith, J. A., & Nelson, M. L. (2006). Lazy preservation: Reconstructing Web sites by crawling the crawlers. In *Proceedings of the 8th ACM International Workshop on Web Information and Data Management*, (pp. 67-74). Arlington, VA, USA: ACM Press.

Mei, Q., & Zhai, C.-X. (2005). Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proceedings of the 11th International Conference on Knowledge Discovery and Data Mining*, (pp. 198-207). New York: ACM Press.

Ntoulas, A., Cho, J., & Olston, C. (2004). What's new on the Web? The evolution of the Web from a search engine perspective. In *Proceedings of the 13th International World Wide Web Conference,* (pp. 1-12). New York: ACM Press.

Papka, R. (1999). *Online new event detection, clustering and tracking*. Unpublished doctoral dissertation, Department of Computer Science, University of Massachusetts, USA.

Rauber, A., Aschenbrenner, A., & Witvoet, O. (2002). Austrian online archive processing: Analyzing archives of the World Wide Web. In *Proceedings of the 6th European Conference on Digital Libraries*, (pp. 16-31). Rome, Italy: Springer-Verlag.

Shipman, F. M., & Hsieh, H. (2000). Navigable history: A reader's view of writer's time: Time-based hypermedia. [Taylor and Francis.]. *New Review of Hypermedia and Multimedia*, *6*, 147–167. doi:10.1080/13614560008914721

Swan, R., & Allan, J. (2000). Automatic generation of overview timelines. In *Proceedings of the 23rd Conference on Research and Development in Information Retrieval*, (pp. 49-56). Athens, Greece: ACM Press.

Toyoda, M., & Kitsuregawa, M. (2003). Extracting evolution of Web communities from a series of Web archives. In *Proceedings the 14th Conference on Hypertext and Hypermedia,* (pp. 28-37). Nottingham, UK: ACM Press.

Viégas, F., Wattenberg, M., & Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the CHI Conference*, (pp. 575-582). Vienna, Austria: ACM Press.

Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining,* (pp. 424-433), Philadelphia, PA, USA: ACM Press.

Yamamoto, Y., Tezuka, T., Jatowt, A., & Tanaka, K. (2007). Honto? Search: Estimating trustworthiness of Web information by search results aggregation and temporal analysis. In *Proceedings of the APWeb/WAIM 2007 Conference*, (pp. 253-264). Hunagshan, China: Springer-Verlag.

ENDNOTES

1. Internet Archive: http://www.archive.org
2. This efficiency is important in case when stream data is required.
3. WERA: http://archive-access.sourceforge.net/projects/wera
4. Wayback Machine: http://www.archive.org
5. Wikipedia, the free encyclopedia: http://en.wikipedia.org/wiki/Wiki

This work was previously published in Intelligence Integration in Distributed Knowledge Management, edited by D. Król & N. Nguyen, pp. 286-304, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 8.16

New Forms of Deep Learning on the Web:
Meeting the Challenge of Cognitive Load in Conditions of Unfettered Exploration in Online Multimedia Environments

Michael DeSchryver
Michigan State University, USA

Rand J. Spiro
Michigan State University, USA

ABSTRACT

We claim that the Web has the potential to be a quintessential multimedia environment for complex learning, particularly in ill-structured domains. This chapter explores the cognitive load considerations associated with several aspects of deep and extended learning on the Web. We also propose the need for a reconceptualization of Cognitive Load Theory for comprehension and learning in more ill-structured conceptual arenas. This reconceptualization emphasizes the need for learning approaches that promote flexible knowledge assembly through processes of organic, reciprocal, and deep Web learning.

DOI: 10.4018/978-1-60566-158-2.ch008

INTRODUCTION

The Web has the potential to become a quintessential multimedia learning environment. Both formal and informal learners are increasingly turning to the search engine as their primary source of information. For example, college students regularly use the Internet and commercial search engines (e.g., Google) before, or in lieu of, local library resources (Griffiths & Brophy, 2005; Jones, 2002; Thompson, 2003; Van Scoyoc, 2006). At the same time, the content provided to them on the Web is presented in multimedia form, often comprising various combinations of text, data, pictures, animation, audio, and video, of differing levels of interactivity. These myriad forms of information all battle for learner attention and consideration.

The second author argued that this migration away from traditional information resources to Web mediated multimedia learning environments is ushering in a revolution in thought, a New Gutenberg Revolution (2006a, b, c, d, e). He outlined how, given the dramatic increases in the speed with which information can be accessed, the increasing well directedness of search (due to more advanced search algorithms, data organization, and searcher skill), and the *de facto* assumption of "ambient findability" (Morville, 2005), the Web is becoming a more fertile knowledge landscape than man has ever known. Consequently, the Web is particularly well suited to support *deep* learning for subjects and concepts that are complex and ill-structured; the kind that we seem to be finding more and more of in the world everyday. These ill-structured domains of knowledge demonstrate an inherent *irregularity* of conceptual application across instances and contexts (for discussion of the special qualities of learning in ill-structured domains, see Spiro & DeSchryver, in press; Spiro, Vispoel, Schmitz, Samarapungavan, & Boerger, 1987; Spiro, Feltovich & Coulson, 1996).

In order to harvest information from this landscape in the most effective and meaningful ways, learners will need to explore the Web with a Post-Gutenberg Mind. This involves searching with *advanced Web exploration* techniques and an *opening mindset* that together result in advanced knowledge acquisition that goes well beyond the cursory and fact based searches most common to learning on the Web (Kuiper, Volman & Terwl, 2005). The knowledge structures acquired by a Post-Gutenberg Mind will be tailored assemblages that are contextualized, interconnected, and flexible, enabling the everyday creativity necessary to succeed in a world increasingly driven by complex and rapidly changing information.

Sounds great, doesn't it? Well, as it turns out, not everyone is prepared to use the Web to learn like this right now. In fact, we do not really know precisely how deep learning on the Web occurs. As such, it is imperative that we begin to thoroughly examine the phenomenon. We need to better understand the specific ways that Post-Gutenberg learning will manifest itself. What specific affordances and aspects will enable deep Web learning of ill-structured concepts? How will they differ from the characteristics of traditional learning? How will we best prepare learners to maximize the benefits of deep learning on the Web? What will be the cognitive load considerations for learning that is of such depth and complexity?

With these questions in mind, we recently embarked on an inquiry to investigate the emergent aspects of Post-Gutenberg learning for an advanced Web learner in an ill-structured domain. In this study, we documented the decision-making, knowledge construction, and general learning reflections of an advanced Web searcher who examined the subject of climate change through deep and extended Web learning. The data were collected from a baseline mind-map of existing knowledge, extensive notes taken by the learner during each session, the parallel use of Clipmarks (an online tool that allows for portions of Web resources to be saved, tagged, annotated, updated, and retrieved in multiple ways), and corresponding updates to the mind-map. A detailed analysis of the data collected from this study has provided a better picture of what Post-Gutenberg learning may look like. And, while we do not claim any generalizability from the results of this demonstration case study (i.e., we sought to demonstrate what *is possible* not *what is*), several interesting phenomena emerged that are worthy of continued scholarly examination. (For a more detailed presentation of the theoretical implications of this study, see Spiro and DeSchryver, in preparation; for a more detailed accounting of a full list of deep Web learning aspects and their derivation, see DeSchryver and Spiro, in preparation.)

One of the primary considerations that surfaced was whether this form of advanced knowledge acquisition would be, in general, too challenging for the learner. Cognitive Load Theory has provided an excellent framework with which to address these questions (e.g., see Sweller, van Merrienboer & Paas, 1998; van Merrienboer & Sweller, 2005). Therefore, this chapter focuses on the cognitive load implications of deep Web learning in ill-structured domains that were identified during our investigation. First, we present an overview of deep Web learning. Second, we review the basic assertions of Cognitive Load Theory and suggest the need to reconceptualize cognitive load as it relates to the goals of learning in ill-structured domains and the skills needed for the requisite flexible knowledge acquisition. Finally, we discuss the cognitive load considerations for specific aspects of deep learning in ill-structured domains on the Web, the adjunct online tools that support them, and related motivational developments.

DEEP WEB LEARNING AND THE NEW GUTENBERG REVOLUTION

Post-Gutenberg learning holds the promise for increasingly complex, but cognitively manageable knowledge acquisition, through learning and instruction that better fits the contours of the world and knowledge about that world. Among other things, it may provide answers to many of the greatest challenges to learning complex and ill-structured topics by supporting the development of the type of adaptive knowledge application that is increasingly important in the world today.

The foundation of Post-Gutenberg learning is the dramatically increasing high speed of connection and access to effectively unlimited Web resources. The former has resulted from technical advancements in hardware and software (e.g., search engine and data-mining algorithms that result in more precise results); enhanced techniques for information organization, such as tagging *en masse*; and, increased integration that advanced Web learners are demonstrating with information search, access, and evaluation (DeSchryver & Spiro, in preparation; Tremayne & Dunwoody, 2001). The seemingly unlimited nature of information available on the Web is described by Morville's (2005) notion of ambient findability, or the crossroads of ubiquitous computing and the Web, in which we can find anyone or anything from anywhere at anytime. This breadth and depth of knowledge provides a multiplicity of perspective, context, interconnectedness, and points of entry that are essential to learning complex, ill-structured concepts.

Most current Web-based learning is rather cursory and fact based (Kuiper, Volman, & Terwel, 2005), therefore new skills and approaches are needed to in order for Post-Gutenberg learning to develop. Foremost among these are the ability to search using *advanced Web exploration* techniques and with an *opening mindset* that promote more wide ranging searches that go well beyond finding facts or "answers" by using just simple Google results or embedded hyperlinks. These wide ranging searches unfold using learner-initiated, complex, reciprocally adaptive (LICRA) techniques that capitalize on the Web's affordances for deep learning by permitting maximally unfettered and externally oriented nonlinear traversals of knowledge spaces. We use "externally oriented" in the sense that the learner's next steps can be strongly influenced by the content that is encountered in Web explorations, rather than operating with more internally driven or "top-down" expectations and learner guidance, such as that provided by embedded and precompiled hyperlinks or simple sequential review of Google result lists. The learner creates their own search phrases based on information they encounter on the Web, either through employing specific ideas from the current page as new search phrases, or by conceptualizing novel search phrases based on recent activity, past experience, and the related momentum Web learning affords. These techniques result in learner controlled non-linear movement through the conceptual landscape, where the learner creates "undefined" connections between resources. In our recent investigation, this LICRA type searching led to several more conceptual breakthroughs than did the use of existing embedded hyperlinks (DeSchryver & Spiro, in preparation).

While performing these searches, learners much be prepared for discovery, complexity, change, and creativity (Spiro, 2006a) and see the Web as *open* to their exploration. When reflected in LICRA searches, openness allows the learners to see the key phrases or ideas in the current document (text, animation, audio, or video) that merit a new branching inquiry. An opening mindset also includes being open to using the Web for more than "basic" Web searches. At Google alone, searches through images, videos, blogs, scholarly materials, books, and news are separately available. Searching with these individual engines provides several unique knowledge landscapes, variant representations, and differing perspectives based on the identical search phrase. Finally, an open mindset is essential to taking advantage of the myriad opportunities for unexpected and *serendipitous* learning on the Web.

Together, these foundational components to deep Web learning undergird *virtual simultaneity*, defined as the condition "in which many things are [simultaneously] being considered in the context of each other and in which *conceptual wholes greater than the sum of the parts* can form" (Spiro, 2006e, p. 2). The simultaneity is virtual, not temporal, in the sense that it is within a functioning cognitive space. This is a core aspect for Post-Gutenberg learning, which, in concert with cognitive spreading activation, allows connections to be noticed that would not be noticed otherwise. It facilitates multiple conceptual comparisons and contrasts, allows for increasingly complex but cognitively manageable learning, provides an acceleration of the acquisition of experience, and develops open knowledge structures that can be tailored to new contexts. The speed increases outlined above have resulted in a learning environment where virtually simultaneity is not only possible, but may be common.

While very little empirical evidence exists for this emergent learning phenomenon, our recent demonstration case study provided evidence of the type of learning we describe above. For instance, the concept of virtual simultaneity was exemplified well in an early session of the study. While learning about climate change, the subject navigated quickly through successive articles in the NY Times global warming section, scanning articles about the New Hampshire ski industry, autobahn speed limit considerations, carbon sequestration, and Pacific Islanders' concerns over rising ocean water. Though seemingly unrelated, this progression of articles one after another lead directly to the learner recognizing and documenting the importance of "selfishness" and "local recognition" as two important concepts in his understanding of climate change. We submit that the speed with which information can be accessed and the precise targetability provided by modern search engines on the Web directly facilitated such rapid conceptual development in a way that no other medium could. In this case, there was clearly a *loose* connection made among what were otherwise *heterogeneous* resources.

Two specific benefits arise from such loose associations and resultant conceptual development. First, the learner begins to appreciate the *conceptual variability* inherent to ill-structured domains. It becomes clear that though similarities exist across the disparate examples, they are not exactly the same. The nuances of New Hampshire residents' concerns with its ski industry cannot be mistaken for the concerns of Pacific Islanders'; these examples are not interchangeable. Fast nonreductive conceptual induction safeguards the conceptual rough edges that are desirable when learning concepts in ill-structured domains. These rough edges directly facilitate the second benefit: *flexible use.* When faced with situation specific need for the concept of selfishness or local recognition, the learner will be well prepared to select an appropriate prototype example (or examples) for application, based on both the similarities and differences that have been preserved among the candidates.

The development of several even more complex knowledge structures related to climate change were also apparent in our study. Among them, consider the concept of carbon markets, about which the subject had minimal prior knowledge at the onset of the study. However, during early research sessions, a complex understanding of carbon markets developed quickly, including the varying effects of mandatory and volunteer cap and trade systems; the Chicago Climate Exchange; related government inclinations (from the United States, European, and Chinese governments); and corporate interests. The subject considered the benefits and drawbacks of carbon markets, from the innovations that were directly credited to Clean Development Mechanism applications, to the inherent waste that carbon market critics argued. Ideas about why (typically anti-regulatory) US business interests are actually calling for more federal regulation in this arena were considered, and proposals that markets may not be able to cure the problem, given some assertions that climate change is the "greatest market failure" ever, were examined.

The implications of this specific example for deep learning in an ill-structured domain are two-fold. First, imagine trying to gather the resources to "instruct" a learner about the topics outlined above, and then completing this task in just a few hours. Any reasonable analogous approximation of the time necessary to replicate this information with traditional resources would be *many* times greater. Examination of a knowledge structure this complex and ill-structured cannot otherwise be accomplished without a learning environment like the Web, its specific affordances, and the skills and mindset outlined above. However, even more important was the evidence that this knowledge was later *flexibly reassembled*. In one of the final research sessions, in the context of how to change individual behavior to address climate change, the learner recombined several of the ideas about carbon markets with concepts from elsewhere (e.g., capitalism and economics) to propose the use of cap and trade systems for individual energy consumption. When considering the impact higher demand, lower supply energy may have on home energy use, the subject adapted the concept of lower effective per-unit carbon costs from carbon markets (and how mandated versus volunteer caps impact the level of incentive provided), with the profit considerations of utility companies, to propose how the adoption of a "personal" cap and trade system for home energy would differ from those currently in place in industrial carbon marketplaces. This thought experiment served to both strengthen his understanding of industrial cap and trade systems (through revisitation of the related ideas) and provided key insights into some of the economic considerations needed to better understand issues related to his inquiry into individual climate change behavior. A more detailed accounting of both learning experiences is provided in DeSchryver & Spiro (in preparation). In sum, this demonstration case study suggested a clear relationship between the way that the material about carbon markets was learned and how it was later flexibly applied.

The level of complexity varied widely for the ideas encountered by the subject over the course of our study. However, he examined several more broadly ill-structured concepts, identified their inter-relationships and integrated them into an evolving and complex knowledge structure of climate change. These included:

- The role of capitalism/consumerism in climate change
- The mixed messages of corporate climate change agendas (e.g., Wal-Mart)
- The ever-changing nature of religion as it relates to climate change, and how "religious" perspectives often differ from the theological and philosophical considerations

- The educational implications of changing climate related behavior
- The complex relationship between climate change mitigation and adaptation
- Technology as both a climate change "cause" and "cure"
- The nascent eco-biz industry and related alternative fuel considerations
- The role and impact of international agencies (e.g., IPCC) and studies (e.g., Stern Report) in climate change
- Why the US and Europe have decidedly different approaches to climate change (both individual and governmental)
- The relationship among supercomputers, modeling, and climate change
- The role of media as it relates to climate change

It is important to note the extent to which each of these ideas interconnected with the others. For instance, the concept of religion was encountered several times. The subject developed an understanding of how both personal and institutional religious considerations shape issues related to climate change in powerful ways, and cannot be disconnected from policy and personal behavioral decisions. At the same time, these ideas demonstrated irregularity across instances, a hallmark of ill-structured concepts that differentiate them from complex, but well-structured concepts. For instance, early in the study, it was determined that several key religious institutions had begun to support policies to aggressively combat climate change. After having seen multiple examples of this positive support, the *temptation to generalize* on the part of the subject could have been very high (e.g., he could have generalized the concept that religious organizations support aggressive policies to mitigate climate change). However, a few sessions later, information about how one key religious institution had reversed its position was encountered. The ability to recognize and appreciate the irregularities that exist for concepts applied in ill-structured domains is critical, and directly facilitated by deep Web learning; learning in this way disables the temptation to overgeneralize by inculcating a mindset that "it's not that simple."

COGNITIVE LOAD THEORY

While we find the above learning outcomes impressive, it may well be that this method of learning is just *hard*, or cognitively demanding. How, then, can we ensure that more people will be able to achieve these outcomes without overwhelming cognitive demands? We will address this question through the lens of Cognitive Load Theory (CLT), a framework often utilized for just such concerns in learning.

CLT has been a leading framework for the study of human learning (e.g., see Sweller, van Merrienboer & Paas, 1998; van Merrienboer & Sweller, 2005). At the heart of CLT is a limited working memory for novel information and procedures, which demonstrates no such limitation when information is retrieved from long-term memory and related schemas (Ericsson & Kintsch, 1995). CLT assumes that long-term memory has unlimited capacity and that learning only occurs when changes in long-term memory have occurred (Kirschner, Sweller, & Clark, 2006). Cognitive load is calculated by the additive relationship among *intrinsic*, *extraneous*, and *germane* cognitive load (Paas, Renkl & Sweller, 2003; Sweller, van Merrienboer & Paas, 1998; van Merrienboer & Sweller, 2005). Intrinsic cognitive load is based on the fundamental characteristics of the material to be learned, most importantly, the number of elements that have to be considered simultaneously for successful learning (called *element interactivity*). Extraneous cognitive load is the *ineffective* load imposed by poor decision making related to the organization and presentation of information to be learned. Germane cognitive load is *effective*, in that it requires the "mindful engagement" of the learner with what is most essential to enhancing learning. Both extraneous and germane cognitive load are considered in CLT to be under the control of instructional designers. The ideas of CLT have been examined in a number of experiments that have confirmed the theory's predicted relationship between kind of load and success in learning for different instructional designs.

In recent years, these basic ideas have been developed and extended to incorporate the demands of "real-life" tasks and the complex learning they often represent (van Merrienboer & Sweller, 2005). For example, the role of "chunking" by experts in reducing the number of interacting elements, which results in lower effective intrinsic load, is increasingly of interest. CLT researchers have also begun to examine the effect of motivation on cognitive load in real-life learning and training contexts (e.g., see Pass, Tuovinen, van Merrienboer & Darabi, 2005).

Learning of the depth and complexity we describe in the section above has several cognitive load implications. Often, the total cognitive load is high. But, this heightened load should not always be avoided, given the potentially deleterious effects to learning from oversimplifying complex material (Feltovich, Coulson & Spiro, 2001; Spiro, Feltovich, Coulson & Anderson, 1989). It does no good to ignore *necessary* learning difficulty. When high cognitive loads for learning in ill-structured domains are encountered, they primarily derive from the nature of the material itself, or its intrinsic cognitive load. We have no choice but to acknowledge that this is the case, and work with it. Whereas, CLT has recently expanded and developed to address complex learning by "artificially" reducing element interactivity to reduce intrinsic cognitive load, since "understanding complex information may not be necessary or even possible in the early stages of learning" (van Merrienboer & Sweller, 2005, p. 157), we do not advocate this approach for ill-structured domains.

If intrinsic cognitive load in an ill-structured domain is high, artificially reducing the complexity in early learning is dangerous in two ways. First, with respect to the local concept, early simplifications interfere with the later acquisition of complexity (Feltovich et al., 1989, 1997, 2001; Spiro, Coulson, Feltovich, & Anderson, 1988; Spiro, Feltovich, Coulson, & Anderson, 1989). Second, such simplifications inculcate a reductive mindset, in general. In stead of simplifying complex, ill-structured material, we need to learn how to deal with it. Learning in and about ill-structured domains is hard and requires mental effort. Boiling these ideas down for simple presentation makes the ideas easier to learn (and easier to teach and assess), but does not facilitate the situation specific, interconnected, context dependent, flexible assemblies of knowledge that are necessary for successful learning outcomes.

In other situations, cognitive load may appear to be higher than it actually is. We present two reasons why this is case. First, in the following section, we revisit the concept of cognitive load and explore how it must be reconceptualized in order to better apply to ill-structured concepts, deep Web learning, and the Post-Gutenberg Mind. This reconceptualization mitigates some of the cognitive load concerns by arguing that common sources of extraneous cognitive load in well-structured domains actually represent germane load in ill-structured domains. Then, in the section thereafter, we discuss how specific aspects of deep learning on the Web may further ameliorate cognitive load concerns.

COGNITIVE LOAD THEORY REVISITED

We find the evolution of CLT to be impressive in its consideration of expertise, real-life tasks, and complex learning. However, given that in its application, CLT has primarily been concerned with "relatively well-structured procedural and conceptual domains" (van Merrienboer & Sweller, 2005, p. 156), we propose that a reconceptualization is necessary in order for CLT to apply to real-life learning tasks in ill-structured domains, especially in deep Web learning. (See also Gerjets and Scheiter (2003) for their useful suggestions about the relationship between cognitive load, learner goals, and individual processing strategies in hypertext learning environments.) We present this reconceptualization with four distinct arguments. First, we highlight how ill-structured domains and well-structured domains are dissimilar and how the goals for learning in each must also necessarily differ. Second, we outline a reconceptualized germane cognitive load that is more appropriate for learning in ill-structured domains. Third, we discuss how this reconceptualization underscores our position that *what is considered extraneous cognitive load in well-structured domains is often germane in ill-structured domains*. Finally, we argue that the Web represents a quintessential environment in which to optimize the ratio of extraneous load to germane load for learning in ill-structured domains.

Learning Goals in Ill-Structured versus Well-Structured Domains

The learning goals in well-structured and ill-structured domains are not the same. This stems from the structural differences that exist between them. Well-structured material, even that which demonstrates complexity, exhibits orderliness and regularity that underscore the goals for learning. It is always possible to present essential information or known procedures for well-structured material. Answers exist to questions posed in well-structured domains. Therefore, the primary goal for learning in well-structured domains is the construction and automation of schemas. Traditional CLT, primarily working within well-structured domains, therefore defines germane cognitive load to comprise the mental effort devoted to such schema construction and automation (van Merrienboer & Sweller, 2005).

However, the goal for learning in ill-structured domains is very different, focusing on the construction of open and flexible knowledge structures for situation specific application. This is primarily due to the *irregularity* that concepts and phenomena demonstrate across instances and applications, such that pre-specifying the conditions under which knowledge will be used is not possible. Because of this irregularity, emphasizing only the acquisition and automation of prepackaged schemas does not provide the flexible knowledge structures required; schemas alone cannot prepare learners for the wide scope of application needed, since essential information and known procedures do not exist. Since there are not specific answers with universal application to questions posed in ill-structured domains, preparing the learner for situation-sensitive development of "schemas-of-the-moment" is of utmost importance. (For a more detailed presentation of the special qualities of learning in ill-structured domains, see Spiro & DeSchryver, in press; Spiro, Vispoel, Schmitz, Samarapungavan, & Boerger, 1987; Spiro, Feltovich & Coulson, 1996.)

Germane Cognitive Load for Ill-Structured Domains

The concept of germane cognitive load must therefore be reconceptualized in order to be consistent with the learning goals for ill-structured domains. In general, we encourage mental efforts that resist oversimplification and resist reductive thinking. In this way, we advocate an emphasis on learning activities that support new constructions of indefinite numbers of situation sensitive assemblages and not retrieval of precompiled schemas or templates from long-term memory. These activities comprise a more appropriate conceptualization of germane cognitive load for ill-structured domains, and should include:

- Recognizing interconnections
- Criss-crossing the knowledge landscape
- Experiencing multiple perspectives and representations
- Testing candidate generalizations with a presumption of "it's not that simple" until proven otherwise
- Seeing patterns of context dependence
- Identifying how surprising similarities and surprising differences unfold

Recognizing that Extraneous Cognitive Load Activities in WSDs Often Appear as Germane Cognitive Load in ISDs

Based on traditional conceptions of CLT, the activities we have outlined that are crucial to learning in ill-structured domains represent extraneous cognitive load when learning in well-structured domains. As well they should. When essential information is known, or a correct procedure available (as is typically the case for well-structured material), it should be presented to learners in a way that maximizes their ability to acquire, update, and automate schemas. Criss-crossing the knowledge landscape to find essential information would increase extraneous load dramatically, wastefully, and unnecessarily.

The reverse is also true. Schema construction and automation (considered germane to learning in well-structured domains) is extraneous to the goal of flexible learning in ill-structured domains. It is equally as wasteful to construct and automate schemas in ill-structured domains as it is to criss-cross well-structured domains looking for known information. Since tailored "in-the-moment" assemblages are most germane to learning in ill-structure domains, intact schema retrieval is minimized, and any time spent constructing intact schemas is essentially wasted. In fact, use of intact schemas tends to compartmentalize knowledge in a way that works against the goals of flexible knowledge, by blocking the ability for the learner to properly recognize the inherent interconnectedness in ill-structured domains (Spiro et al., 1991). In this way, schema construction not only represents extraneous load for learning in ill-structured domains, but also directly undermines the primary process of flexible knowledge construction.

The Web is a Quintessential Learning Environment in Which to Optimize Germane Cognitive Load for Learning in Ill-Structured Domains

Our conceptualization of germane cognitive load for ill-structured domains will also appear to represent extraneous cognitive load when considered for traditional linear media. However, with hypermedia and especially now the Web, we finally have learning environments that accommodate advanced knowledge acquisition in ill-structured domains and work to make our proposed mental efforts germane. In fact, the affordances of the Web keep extraneous load *much lower* for learning activities that are specifically relevant to ill-structured domains than they would otherwise be with traditional media.

For instance, if you accept that criss-crossing non-linear irregularly interconnected knowledge is essential to learning in ill-structured domains, what better medium exists than the Web? Consider again how long it would take for a learner to find print media in a library even closely approximating the expansive and up-to-the-minute resources outlined above for climate change. Then, imagine how long it would take to begin to recognize the intricate interconnectedness among these myriad resources. The extraneous load requirements would be overwhelming with traditional media. However, on the Web, interconnectedness is inherent. What is the optimal use of cognitive load resources to recognize interconnections among disparate information resources, card catalogs and indexes, or the instant-access, fully searchable, non-linear Web? There is no comparison. The Web is ideal for learning in this way.

COGNITIVE LOAD CONSIDERATIONS FOR SPECIFIC ASPECTS OF DEEP WEB LEARNING IN ILL-STRUCTURED DOMAINS

As we have noted, many aspects of deep Web learning in ill-structured domains that are germane might, in well-structured domains or with traditional linear media, be considered extraneous, and vice-versa. In this section, we discuss several aspects of deep Web learning in ill-structured domains and how they may help to maximize the ratio of germane cognitive load to extraneous cognitive load. Again, we present these ideas as *what is possible* for deep learning on the Web not *what is* the case for everyone. Continued research that examines these phenomena more specifically will help identify the conditions under which they apply more broadly, and how to best prepare novice learners for deep Web learning in ill-structured domains.

LICRA Searches

Learner-initiated, complex, reciprocally adaptive (LICRA) searches are a core feature to successful deep Web learning in ill-structured domains. These searches promote active involvement from the learner, require heightened attention, and emphasize the need for value determination and judgment. Given this additional effort required to generate successful LICRA searches, it should be expected that associated cognitive loads are rather high. Indeed, traditional CLT often associates searching of any kind-- for information needed to complete a learning task, for solutions, or for referents in an explanation-- with extraneous cognitive load (Paas, Renkl, & Sweller, 2003; van Merrienboer Sweller, 2005).

However, consistent with the above account of how cognitive load applies to ill-structured domains, this effort represents *germane* cognitive load. LICRA searches are part of the learning process and cannot be viewed separately. When learners use a LICRA phrase, they create implicit interconnections among resources. The iterative nature of these searches also facilitates criss-crossing the conceptual landscape in order to construct a better sense of the whole. As a result, when compared to following embedded hyperlinks or scanning through a single list of Google results, these iterative search techniques promise an even greater payoff.

Extraneous cognitive load concerns about hypertext learning are somewhat ameliorated by LICRA searching, as well. The decision-making that is "required" by the visual cue of an embedded hyperlink may serve to increase extraneous load, especially if multiple links appear on the same page (DeStefano & LeFevre, 2007). This "interruptive" quality does not exist for LICRA searches, because they result from an internal, somewhat spontaneous, assignment of importance on the part of the learner. There is also some concern that following hyperlinks, especially for a long time or to "semantically distant information" (DeStefano & LeFevre, 2007, p. 1620) increases cognitive load and reduces comprehension. However, criss-crossing, or following links, either embedded or learner initiated, is a core feature of deep Web learning when learning in ill-structured domains. And, many forms of extraneous load associated with the criss-crossing supported by LICRA searches should be increasingly mitigated by the growing body of adjunct learning aids (see below) that help to manage and make explicit the choices and relationships that emerge while exploring the knowledge landscape. We also envision that as learners become more effective at LICRA searching, the "after-the-fact" ratio of germane to extraneous load will also optimize (i.e., more LICRA phrases will yield useful information and less will be "dead-ends").

It should be noted that there are well-structured skills that are necessary in order for learners to perform successful LICRA searches. For instance, using quotation marks to target exact phrases, employing Boolean logic, searching a specific site, using a tilde to return related items in a search, and using the "link:" phrase to determine backlinks, are all useful skills that should be taught to and practiced by Web users. We support the notion that the automation of schema for such skills is essential to good LICRA searching and should be done with the lowest possible extraneous cognitive load.

Serendipitous Learning

Serendipitous learning comprises an accidental encounter with seemingly unrelated or useless information that then becomes relevant. Such serendipitous learning is automatic, unconscious, and spontaneous in nature, inherently requires no conscious mental effort, and thus demonstrates a very low cognitive load requirement. While the value of making serendipitous connections between unrelated ideas is not new to hypertext research (e.g., Bernstein, Bolter, Joyce, and Mylonas, 1991), it has often been a form of "quasi-random" knowledge association, given the finite amount of information in closed hypertext systems or when just following links provided on a web page. However, in deep Web searches utilizing LICRA methods, serendipitous information can, and does, emerge at any time. The likelihood of the occurrence of serendipitous learning increases exponentially as the speed of access to information increases, as Web users search more efficiently, and as the landscape of resources available becomes, for practical purposes, unlimited. This exponential increase in potential random encounters requires no additional effort on the part of the learner, but is often significant to the learning process. Our recent research demonstrated several points at which serendipitous finds resulted in significant conceptual breakthroughs and novel interconnections of information for the learner (DeSchryver & Spiro, in preparation).

Such serendipitous learning can happen at any time on the Web, including moving "backwards" in the search process. For instance, in our recent study, it was common for the learner to open several new browser tabs during LICRA searches to accommodate multilevel iterative search phrase development. Several times, after a "stopping-point" in this process was reached, and while returning to his original search results (e.g., "backing out" of the iterative searches), the subject noticed information that did not seem relevant when it was first viewed but that was particularly useful in the context of the new information he had since encountered. This new relevance may simply have occurred as a result of the accumulated context. However, the benefits of "considering," or reviewing, information multiple times within a relatively short period, need to be explored.

Resource Evaluation

The evaluation of a Web resource for authority and accuracy typically has extensive extraneous cognitive load implications. Learners are often provided with rubrics containing several questions to answer for each Web resource they encounter. For example: Is it a personal page? What type of domain is it from? Who wrote the page? Is it dated? Is it current enough? What are the author's credentials? Are sources documented? Are there links to other sources? (see Finding, 2008, for a full and much more exhaustive list). The use of such guidelines requires very high extraneous cognitive loads. And, while such rubrics may be valuable for young or introductory learners, our recent study provided evidence that this process may become fully integrated into the learning process for advanced learners.

One reason for such integration is that the Web is full of "trusted aggregators" that reduce the evaluation necessary. For instance, the New York Times has a list of articles related to climate change. At the same time, especially in extended learning on the Web, learners will visit the same sites over and over. Subsequent visits to a trusted site (including what are often visits to topically disparate sub-nodes in the site) require no additional evaluation.

However, more significantly, we submit that the integration of resource evaluation results primarily when the learner becomes adept at seeking out multiple representations of one concept while learning about it. Through recognizing the similarities and differences among the multiplicity of resources, it quickly becomes apparent which resources are trusted and which are not, without working through a guideline of static questions about each resource. In this way, resource evaluation is *fully integrated into the learning process* for the advanced learner by an often unwitting method of *triangulation*. Each site visited not only provides new information, new contexts, new representations, and new perspectives that are germane to learning, but also tightens the criteria for filtering out resources that do not reflect the quality, accuracy or credibility of acceptable sites. From a cognitive load perspective, an extraneous load requirement has been transferred to a germane learning activity.

Blogs

The nature of the information found during deep Web learning has significant implications for the learner's cognitive load. For instance, blogs now increasingly find their way into Web search results. The casual form of information typically encountered in blogs has largely been considered useless for serious knowledge inquiry (e.g., see Head, 2007). However, we found that several categories of blogs were examined during our recent study, running the gamut from factual to opinionated. Analysis of the various types of blogs that were used, as well as their interactions with the differing stages of meaningful learning (Sheull, 1990), indicated that blogs can be quite valuable to the learning process and that the related cognitive load requirements vary.

For instance, the use of blogs in early stage learning (fact-based) does not seem beneficial. The above methods for resource evaluation require more authoritative resources for fact-finding. Though many excellent factual blogs exist and were encountered by the subject during early stage learning, most were filtered out by resource triangulation methods. It seems additional extraneous cognitive load would be required to determine the credibility of facts in a blog, whereas the triangulation techniques for resource evaluation for other more authoritative factual sites require less effort. Evaluating blogs for "facts" just does not seem worth the extra mental effort at this stage in learning.

However, during latter stages in the learning process (involving more problem-solving and abstract thinking) triangulation methods for resource evaluation valued the opportunity for "idea play" that less factual blogs provided. We submit that taking into account the speed with which ideas from different blog sites can be experienced on the Web can result in a dialogical interaction among the user and multiple blog sites. In this way, the learner can agree with some authors and disagree with others; accept some points, and discount others; all the while developing his or her own ideas through arguing, counter-arguing, and the related combinatorial idea play. In this way, the voices in blog entries are *virtually synchronized* and provide great potential for deep learning of ill-structured domains on the Web. At the same time, the extraneous load associated with early (fact-based) use of blogs is minimized, since the interactions with blogs become germane during creative knowledge assemblies. Extraneous cognitive load is also further reduced by the conversational text found in most blogs that is often easier to comprehend for the learner than other, more authoritative resources. Examples of this type of learning are provided in DeSchryver and Spiro (in preparation), where the use of blogs was directly responsible for several significant conceptual breakthroughs on the part of the learner.

Adjunct Online Tools

Cognitive support tools for deep learning on the Web are widely available. From Web text highlighting to page specific sticky notes, *adjunct online tools* abound, and new ones are released regularly. The number of modern tools to support deep Web learning dwarfs those available for traditional text or lecture-based environments. The unprecedented availability of these adjunct supportive aids also has significant cognitive load implications.

External memory aids like Clipmarks (http://www.clipmarks.com) make it effortless to save whole pages as well as small segments of pages (including images and video) in a personalized database. The information contained therein can be accessed by keyword, Boolean search, or chronologically, and increases the ease with which learners can criss-cross and revisit the information. This facilitates the recognition of interconnections, multiplicities, and differing contexts that are critical to the acquisition of flexible knowledge (Spiro & Jehng, 1990). At the same time, tools such as Google History and Trailfire (or even just concerted use of the tabs available in modern browsers for individual sessions) allow users to document and easily trace their paths through the web. While keeping extraneous cognitive load low relative to navigation (e.g., you can never really get lost), these services also have the potential to dramatically support the additional (germane) meta-cognitive activities that deep Web learning affords, by making explicit the paths of inquiry for later review, with little or no additional extraneous load.

Though not a specific tool, the functionality of *tagging* (assigning keywords to content of interest) also increases the potential benefits of deep Web learning. Most contemporary Web-based content management systems (e.g., Clipmarks), utilize this feature extensively. Tagging is a categorization activity that requires the assignment of meaning, a task that therefore represents germane cognitive load. Tagging information can represent an elaboration that fosters the encoding of long-term memories (Budiu, Pirolli, & Hong, 2007; van Merrienboer & Sweller, 2005). However, more important to learning in ill-structured domains, the application of multiple tags facilitates learner recognition of interconnections that may exist among the newly discovered information and previous information that has been similarly tagged. And, as Sinha (2005) noted, the practice of applying multiple tags to information may actually require a lower cognitive load than assigning information to a single category. She argued that upon encountering worthwhile information, multiple semantic concepts are typically activated in a learner. When forced to assign the information to one category, significant cognitive load is spent determining the *best* category, often resulting in what she called "post activation analysis paralysis" (para 16). However, the multiplicity of tagging allows learners to freedom to assign information to all related semantic concepts that are activated. She claims that it "taps into an existing cognitive process without adding much cognitive cost" (para 19).

There are also broader benefits to be accrued from tagging. For instance, Weinberger (2007) asserted that tagging *en masse*, combined with sophisticated data mining and search technologies, will facilitate better reconstructions of the implicit meaning of information on the Web. When available as part of the search process, this "meaning" will greatly focus searches for everyone, thus lowering any extraneous cognitive load related to filtering results.

The Question of Motivation

Recent developments in CLT have begun to integrate the importance of motivation, especially when considering "real-life," extended time tasks

(Pass, Tuovinen, van Merrienboer & Darabi, 2005; van Merrienboer & Sweller, 2005). And, while Astleitner and Wiesner (2004) raised concern for how a high density of motivational strategies explicitly designed for instructional settings may lead to increased cognitive load, we submit that motivation in deep Web learning is largely inherent, not an additional learning component that vies for space in working memory. Clearly, the specific cognitive load implications of motivation are just beginning to emerge. However, as Pass et al. (2005) noted "meaningful learning can only commence if training experience is coupled with the motivation to achieve well" (p. 26). More specifically, higher levels of learner motivation have the potential to increase the use of mental efforts that are germane to the learning outcomes.

Deep Web learning affords such increased motivation, and it was manifested in surprising ways in our recent study. Several times the subject noted his excitement about a particular path of inquiry, new and promising search phrases based on ideas from the current page, or a significant conceptual breakthrough. In a few cases a sort of learning *momentum* affected his motivations. Why? Although the exact processes are speculative at this point, we offer three ways deep Web learning appears to increase motivation, through attention, choice, and speed.

In Keller's (1987a; 1987b; 1999) Model of Motivation Design, he noted that *attention* and subsequent motivation could be gained through novel, surprising, incongruous, and uncertain events. We argue that novel, surprising, incongruous, and uncertain *ideas* discovered in the course of deep Web learning have a similarly enhancing effect on learner attention. And, since the Web interactions are not "designed" elements, they do not increase extraneous cognitive load, so that any increase in attention and effort likely serves to optimize germane cognitive load.

In addition, as Brophy has noted, much of what is known about optimal conditions for motivation cannot be easily applied in classrooms (in Gaedke & Shaughnessy, 2003). However, these very same ideas can and often do apply in deep Web learning, such as *learner control*. In typical classrooms, learner choice is minimized; however, considerable choice is afforded to Post-Gutenberg learners. They learn whenever and wherever they want. They choose what search phrases to use. They choose which results to visit. The assemblage of knowledge is personal in every way. The impact of this type of learning on personal attributions, expectations, and self-efficacy needs to be explored in more detail. However, we propose that once learners begin to feel empowered (which does not take long), they begin to *believe* that they can achieve their learning goals.

Finally, *speed* is essential to learner motivation on the Web. Everything happens fast. Resources are available in an instant. However, even more significant, our recent study of deep Web learning demonstrated a *quick transition* from early "fact-finding" to the *advanced stages of meaningful learning* that involved rapid interconnectedness and abstract/problem based thinking (Sheull, 1990). Accordingly, the subject indicated more excitement about these latter stages. The speed with which new information is available also impacts motivation. The learner can read reaction to yesterday's Senate hearings on climate change today, or in some cases can evaluate information in real-time. Access to *information this timely* empowers the learner, and may even make them feel like they know something others do not (including teachers).

Together, the impact of these motivational considerations on cognitive load needs further exploration, particularly for deep Web learning. However, we envision increased motivation, inherent in the learning process itself, providing substantial benefits to the process through increased germane mental efforts.

CONCLUSION

The chapter discusses how deep learning on the Web provides affordances that are well matched to the learning goals and requirements for flexible knowledge construction in ill-structured domains. In so doing, we have outlined a reconceptualization of germane cognitive load that is more appropriate for learning in ill-structured domains than that conceived by traditional Cognitive Load Theory. This new perspective makes it clear that mental efforts that are germane to learning in ill-structured domains are often extraneous in well-structured domains, and vice-versa. We have also discussed how specific aspects of deep Web learning in ill-structured domains help to optimize the ratio of germane cognitive load to extraneous cognitive load, including LICRA searching, serendipitous learning, resource evaluation, blog use, the availability of adjunct cognitive tools, tagging, and learner motivation.

Not everyone is prepared to take advantage of the benefits we have outlined. Nor do all domains of knowledge benefit from unfettered exploration in online multimedia environments. Additionally, cognitive load may be increased at first for novice Web learners. However, for ill-structured domains, deep Web learning holds great promise. The automatization of the basic skills required to successfully learn in depth on the Web happens quickly. The learner's skills improve with practice, and soon these skills are incorporated effortlessly within higher-level aspects of learning (e.g., LICRA searching and integrated resource evaluation), similar to how an advanced driver uses a steering wheel while navigating during rush hour. Expert drivers steer unconsciously while they attend to the landscape around them. So, too, will expert searchers direct themselves around the vast knowledge landscape of the Web without concerted effort. The automaticity of simple search skills leads to decreased extraneous cognitive load, freeing up resources for activities that are germane to flexible knowledge acquisition in ill-structured domains.

We are confident that the learning we describe above can become a large-scale reality. However, while we have concentrated on the potential benefits that deep Web learning will provide, obstacles to its success remain. For instance, the sheer quantity of resources available on the Web imposes a challenge. Learners may, even with the assistance of adjunct cognitive tools, drift away from relevant search results to information that is, in effect, extraneous to their efforts. In order to address these issues, we anticipate the need to develop loose meta-data structures similar to those employed in Cognitive Flexibility Hypertext systems (e.g., see Spiro, Collins & Ramchandran, 2007).

In addition, managing the *personal knowledge landscape* constructed by individuals immersed in deep Web learning may require support beyond what current adjunct online tools provide. Such support may take the form of more integrated tools capable of reducing the associated extraneous cognitive load. One such possibility would be the development of representative dynamic visual displays. We see the need for evolving network mind-maps, developed in an ongoing fashion with input from both the learner and the software in use. The information could be tagged by context so that the visual display highlights different information in different contexts, demonstrates the interconnectedness among ideas, facilitates revisitation, and provides significant meta-cognitive learning benefits. For example, this system could support three-dimensional mind-maps, each node mashed-up with the information from a Clipmarks-like database and relevant Google History details, and include visual representations of the interconnections that have formed.

FUTURE RESEARCH

To conclude, we offer the following: The Web will change the way that we think and learn, and these changes will be dramatic. It is inevitable. The timing of this revolution is fortuitous, since we are faced with increasingly complex and ill-structured educational and societal issues, both local and global in scale. Consider the difficulty we face assuming, as our forefathers did, that we have an "informed constituency" to support democracy. For example, climate change, health care, and globalization are "grand social challenges" that an informed electorate should understand to a much greater extent than is currently the case. These issues and others like them demonstrate ill-structuredness and they therefore require high levels of intrinsic cognitive load to understand. We believe that with concerted research efforts, ways for the Web to make learning about such issues cognitively manageable will result.

As a field, we need to be proactive in thinking about how to ensure that this happens. Our chapter is provided to encourage more researchers to examine this phenomenon in earnest. Myriad research opportunities and lines of inquiry exist and beg examination under both well-controlled and "real-life" conditions to determine the circumstances, knowledge domains, and learner characteristics for which new ways of learning will apply most broadly. What are the best ways to ensure that learners approach deep Web learning with an opening mindset? How will we best use tagging, LICRA searches and blogs to develop flexible knowledge appropriate to the challenges we face? Is the cognitive load for LICRA searching less than that for embedded links? If not, do the benefits outweigh the costs? Do triangulation methods of resource evaluation ensure an acceptable level of credibility? Do learners need to develop a certain level of expertise in the basic skills of deep Web learning in order for the motivational benefits we have offered to emerge? These questions, and many more, require the attention of researchers in order to better understand ongoing changes in the way we think and learn about critically important issues.

REFERENCES

Astleitner, H., & Wiesner, C. (2004). An integrated model of multimedia learning and motivation. *Journal of Educational Multimedia and Hypermedia*, *13*, 3–21.

Bernstein, M., Bolter, J. D., Joyce, M., & Mylonas, E. (1991) Architectures for volatile hypertext. *Proceedings of Hypertext' 91*, ACM Press, San Antonio, Texas, December, 241-260.

Budiu, R., Pirolli, P., & Hong, L. (2007). Remembrance of things tagged: How tagging affects human information processing. *Palo Alto Research Center.* Retrieved January 17, 2008 from http://asc.parc.googlepages.com/2007-09-12-tagmemory-12-blogged-vers.pdf

DeSchryver, M. & Spiro, R.J. (in preparation). Nonlinear knowledge acquisition in a post-Gutenberg world: An examination of emerging constructs for deep learning on the Web.

DeStefano, D., & LeFevre, J. (2007). Cognitive load in hypertext reading: A review. *Computers in Human Behavior*, *23*, 1616–1641. doi:10.1016/j.chb.2005.08.012

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–245. doi:10.1037/0033-295X.102.2.211

Feltovich, P. J., Coulson, R. L., & Spiro, R. J. (2001). Learners' understanding of important and difficult concepts: A challenge to smart machines in education. In P.J. Feltovich & K. Forbus (Eds.). *Smart Machines in Education.* Cambridge, MA: MIT Press.

Feltovich, P. J., Spiro, R. J., & Coulson, R. L. (1989). The nature of conceptual understanding in biomedicine: The deep structure of complex ideas and the development of misconceptions. In D. Evans & V. Patel (Eds.), *The Cognitive Sciences in Medicine* (pp. 113-172). Cambridge, MA: M.I.T. Press.

Feltovich, P. J., Spiro, R. J., & Coulson, R. L. (1997). Issues of expert flexibility in contexts characterized by complexity and change. In P.J. Feltovich, K.M. Ford, & R.R. Hoffman (Eds.). *Expertise in context: Human and machine*. Cambridge, MA: MIT Press.

Finding information on the Internet: A tutorial, (2008). Retrieved February 10, 2008 from http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Evaluate.html

Gaedke, B., & Shaughnessy, M. F. (2003). An interview with Jere Brophy. *Educational Psychology Review*, *15*(2), 199–211. doi:10.1023/A:1023481231990

Gerjets, P., & Scheiter, K. (2003). Goals and strategies as moderators between instructional design and cognitive load: Evidence from hypertext-based instruction. *Educational Psychologist*, *38*, 33–41. doi:10.1207/S15326985EP3801_5

Graff, M. (2003). Learning from Web-based instructional systems and cognitive style. *British Journal of Educational Technology*, *34*(4), 407–418. doi:10.1111/1467-8535.00338

Griffiths, J., & Brophy, P. (2005). Student searching behavior and the Web: Use of academic resources and Google . *Library Trends*, *53*(4), 539–554.

Head, A. J. (2007). Beyond Google: How do students conducts academic research? *First Monday*, *12*(8). Retrieved January 15, 2008 from http://firstmonday.org/issues/issue12_8/head/index.htm

Jones, S. (2002). The Internet goes to college: How students are living in the future with today's technology. *Pew Internet and American Life Project*. Retrieved January 20, 2008 from http://www.pewinternet.org/PPF/r/71/report_display.asp

Keller, J. M. (1987a). Development and use of the ARCS model of motivational design. *Journal of Instructional Development*, *10*(3), 2–10. doi:10.1007/BF02905780

Keller, J. M. (1987b). Strategies for stimulating the motivation to learn. *Performance & Instruction*, *26*(8), 1–7. doi:10.1002/pfi.4160260802

Keller, J. M. (1999). Motivation in cyber learning environments. *Educational Technology International*, *1*(1), 7–30.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86. doi:10.1207/s15326985ep4102_1

Kuiper, E., Volman, M., & Terwel, J. (2005). The Web as an information resources in K-12 education: Strategies for supporting students in searching and processing information. *Review of Educational Research*, *75*(3), 285–328. doi:10.3102/00346543075003285

Morville, P. (2005). *Ambient findability: What we find changes who we become*. Cambridge, MA: O'Reilly Media.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4. doi:10.1207/S15326985EP3801_1

Paas, F., Tuovinen, J. E., van Merriënboer, J. J. G., & Darabi, A. (2005). A motivational perspective on the relation between mental effort and performance: optimizing learners' involvement in instructional conditions. *Educational Technology Research and Development*, *53*, 25–34. doi:10.1007/BF02504795

Sheull, T. J. (1990). Toward a unified approach to learning as a multisource phenomenon. *Review of Educational Research*, *60*(4), 531–547.

Sinha, R. (2005). A cognitive analysis of tagging (or how the lower cognitive cost of tagging makes it popular). Thoughts About Cognition, Design, and Technology: Rashmi Sinha's weblog. Retreived December 12, 2007 from http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html

Spiro, R. J. (2006a). The "New Gutenberg Revolution": Radical new learning, thinking, teaching, and training with technology…bringing the future near. *Educational Technology*, *46*(1), 3–4.

Spiro, R. J. (2006b). The Post-Gutenberg world of the mind: The shape of the new learning. *Educational Technology*, *46*(2), 3–4.

Spiro, R. J. (2006c). Old ways die hard. *Educational Technology*, *46*(3), 3–4.

Spiro, R. J. (2006d). Approaching the post-Gutenberg mind: The revolution is in progress. *Educational Technology*, *46*(4), 3–4.

Spiro, R. J. (2006e). The "New Gutenberg Revolution": Radical new learning, thinking, teaching, and training with technology…bringing the future near. *Educational Technology*, *46*(6), 3–5.

Spiro, R. J., Collins, B. P., & Ramchandran, A. R. (2007). Reflections on a Post-Gutenberg epistemology for video use in ill-structured domains: Things you can do with video to foster complex learning and cognitive flexibility. *Video Research in the Learning Sciences*. Goldman, R., Pea R., Barron B., and Derry, S. (Eds.). Mahwah, NJ: Lawrence Erlbaum Associates.

Spiro, R. J., Coulson, R. L., Feltovich, P. J., & Anderson, D. (1988). Cognitive flexibility theory: Advanced knowledge acquisition in ill-structured domains. *Tenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, [Reprinted in In R. B. Ruddell (Ed.), *Theoretical models and processes of reading* (5th ed.). Newark, DE: International Reading Association, pp. 602-616.]

Spiro, R. J., & DeSchryver, M. (in press). Constructivism: When it's the wrong idea and when it's the only idea. In T. Duffy & S. Tobais (Eds.). Mahwah, NJ: Lawrence Erlbaum.

Spiro, R.J., & DeSchryver, M. (in preparation). The new nonlinear reading comprehension: Deep learning on the Web and the post-Gutenberg mind.

Spiro, R. J., Feltovich, P. J., & Coulson, R. L. (1996). Two epistemic world-views: Prefigurative schemas and learning in complex domains. *Applied Cognitive Psychology*, *10*, 52–61. doi:10.1002/(SICI)1099-0720(199611)10:7<51::AID-ACP437>3.0.CO;2-F

Spiro, R. J., Feltovich, P. J., Coulson, R. L., & Anderson, D. (1989). Multiple analogies for complex concepts: Antidotes for analogy-induced misconception in advanced knowledge acquisition. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 498-531). Cambridge, MA: Cambridge University Press.

Spiro, R. J., Feltovich, P. J., Jacobson, M. J., & Coulson, R. L. (1991). Knowledge representation, content specification, and the development of skill in situation-specific knowledge assembly: Some constructivist issues as they relate to cognitive flexibility theory and hypertext. *Educational Technology*, *31*(9), 22–25.

Spiro, R. J., & Jehng, J. (1990). Cognitive flexibility and hypertext: Theory and technology for the nonlinear and multidimensional traversal of complex subject matter. In D. Nix & R. J. Spiro (Eds.), *Cognition, education, and multimedia* (pp. 163-205). Hillsdale: Lawrence Erlbaum Associates.

Spiro, R. J., Vispoel, W. L., Schmitz, J., Samarapungavan, A., & Boerger, A. (1987). Knowledge acquisition for application: Cognitive flexibility and transfer in complex content domains. In B. C. Britton & S. Glynn (Eds.), *Executive control processes*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning . *Cognitive Science*, *12*, 157–285.

Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design . *Learning and Instruction*, *4*, 295–312. doi:10.1016/0959-4752(94)90003-5

Sweller, J., van Merrienboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. doi:10.1023/A:1022193728205

Thompson, C. (2003). Information illiterate or lazy: How college students use the Web for research. *Libraries and the Academy*, *3*(2), 259–268. doi:10.1353/pla.2003.0047

Tremayne, M., & Dunwoody, S. (2001). Interactivity, information processing, and learning on the World Wide Web. *Science Communication*, *23*(2), 111–134. doi:10.1177/1075547001023002003

van Merrienboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*(2), 147–177. doi:10.1007/s10648-005-3951-0

Van Scoyoc, A. M. (2006). The electronic library: Undergraduate research behavior in a library without books. *Libraries and the Academy*, *6*(1), 47–58. doi:10.1353/pla.2006.0012

Weinberger, D. (2007). *Everything Is Miscellaneous*, New York: Henry Holt.

This work was previously published in Cognitive Effects of Multimedia Learning, edited by R. Zheng, pp. 134-152, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Chapter 8.17
General Strategy for Querying Web Sources in a Data Federation Environment

Aykut Firat
Northeastern University, USA

Lynn Wu
Massachusetts Institute of Technology, USA

Stuart Madnick
Massachusetts Institute of Technology, USA

ABSTRACT

Modern database management systems are supporting the inclusion and querying of non-relational sources within a data federation environment via wrappers. Wrapper development for Web sources, however, is a convolution of code with extraction and query planning knowledge and becomes a daunting task. We use IBM DB2 federation engine to demonstrate the challenges of incorporating Web sources into a data federation. We, then, present a practical and general strategy for the inclusion and querying of Web sources without requiring any changes in the underlying data federation technology. This strategy separates the code and knowledge in wrapper development by introducing a general-purpose capabilities-aware mini query-planner and a data extraction engine. As a result, Web sources can be included in a data federation system faster, and maintained easier.

INTRODUCTION

Federated databases offer information integration on demand in dynamic environments, where data warehousing approaches are not feasible (Sheth & Larson, 1990; Geer, 2003). In modern relational database management systems, even non-relational sources can be included in a data federation via "wrappers" so that they can be queried as if they are part of a single large database (Somani, Choy, & Kleewein, 2002; Thiran, Hainaut, Houben, & Benslimane, 2006). Wrappers are mechanisms by which the federated server interacts with non-relational data sources by performing operations such as connecting to a data source and retrieving data from it iteratively.

Retrieving data from Web sources, however, is complicated because data is semistructured and Web sources may have requirements (e.g., they may require forms to be filled before returning data); thus general-purpose wrappers for arbitrary Web pages are not provided in data federation systems. Instead the user needs to implement a custom wrapper for each Web source by coding data extraction patterns and parts of the federated query planning protocol in a low-level programming language such as C. This convolution of code with the data extraction and planning knowledge turns wrapper development into a daunting task, results in code duplication, and slows down the data federation process.

Within the last decade or so, many research projects (Papakonstantinou, Gupta, & Haas, 1998; Levy, Rajaraman, & Ordille, 1996; Li & Chang, 2000; Zadorozhny, Bright, Vidal, Raschid, & Urhan, 2002; Li, 2003; Pentaris & Ioannidis, 2006) offered algorithmic solutions to "query planning with source restrictions." The goal of these studies was to offer an expressive language to specify source restrictions, and let the federated query planner come up with an optimal plan using this knowledge. These approaches do not need any cooperation from the individual data sources other than knowing about their limitations. Had they found their way into commercial systems, they would eliminate part of the code and knowledge convolution problem: the wrapper developer would only need to code the data extraction knowledge and not worry about the query planning aspects. Yet the separation of code and knowledge would still not be satisfactorily achieved in non-cooperative federated query planners. For this study, we have chosen to work with IBM DB2's cooperative federated query planner, which poses more challenges than the non-cooperative ones. Our focus is on improving the usability and maintenance aspects of the wrapper development process without requiring any changes in its underlying data federation technology. We do not offer yet another proposal to rewrite a state-of-the-art distributed query planner (Kossmann, 2000), or create an independent infrastructure for querying Internet data sources (Braumandl et al., 2001; Suciu, 2002), but provide a non-intrusive approach that works with what is available today with minimal effort.

We have tested our prototype implementation with numerous Web sites. A moderate user with no programming experience can include a typical Web site into a data federation in less than an hour. The process often takes much longer when the existing procedural coding approach is used by an experienced programmer. Furthermore, explaining, learning, and tutoring wrapper development becomes much easier, as the task changes from writing and debugging a *program* to specifying and debugging *knowledge*.

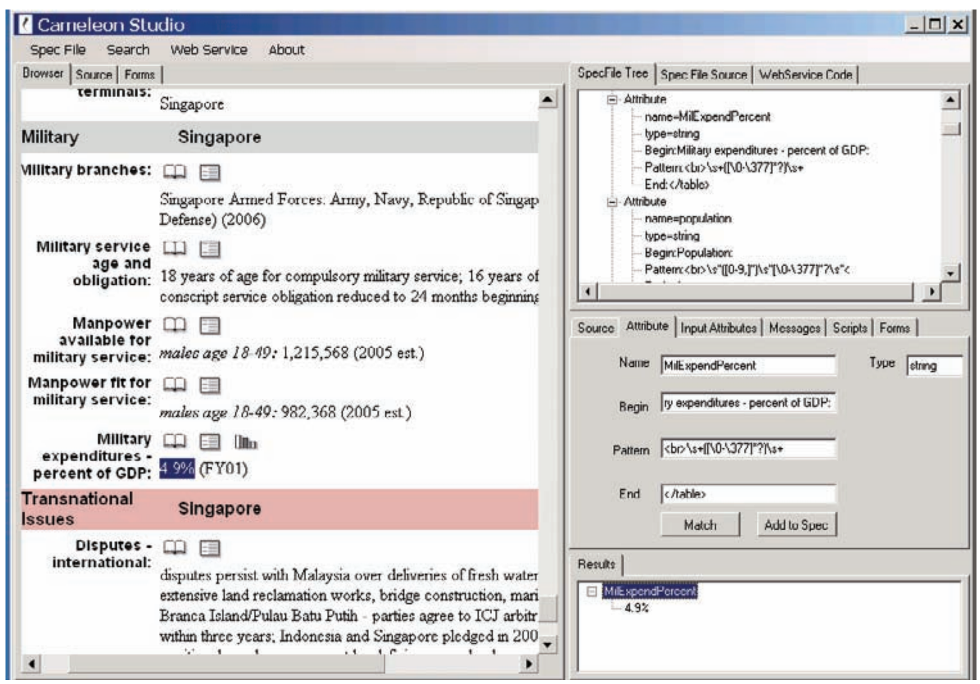
In the rest of this paper, we start with a motivational example that illustrates the need for data federation involving Web sources. We then provide some background on data federation with non-relational data sources and describe the current architectural difficulties of incorporating a Web source. Next, we describe our approach to wrapper development, and the algorithms used to perform planning and optimization for Web sources with capability restrictions. We end with an overview of related work and future research issues.

MOTIVATIONAL EXAMPLE

Consider, first, finding the *military expenditure per capita* of countries in the world using the CIA world fact book Web site. This information is scattered inside the world fact book (see Figure 1), and first needs to be located and extracted.

Figure 1. Available data in CIA World Fact Book site

Figure 2. CIA World Fact Book site is visually wrapped with Cameleon# Wrapper Engine



By using the Web wrapper, Cameleon# (Firat, Madnick, Yahaya, Kuan, & Bressan, 2005; Firat, Madnick, & Siegel, 2000) and its accompanied visual helper, Cameleon# Studio, we can wrap the CIA world fact book site using simple regular expressions and treat it as a very simple relational table as illustrated in Figures 2 and 3.

The Cameleon# wrapper engine's main functionality is, however, extraction and thus is only able to answer SQL queries involving a single source, with required inputs bound to a single set of values at a time. For that reason, we decided to use the powerful query planning, optimizing, and execution capabilities of a commercial data federation engine to handle more complex query situations. Using the extended architecture to be described later on, we define a nickname for our Web source in DB2 as shown below:

```
CREATE NICKNAME CIA (
country char(20),
population dec(10,1),
```

GDP dec(10, 2),
GDP_unit char(20),
MilExpendPercent dec(9,4) for server Cameleon#_server
options(SERVER_NAME 'http://interchange.mit.edu/Cameleon_sharp/camserv.aspx?',
PREDICATES 'country')

The options in the above description indicate the location of Cameleon# information extraction server, and the required input column *country*. We are then able to treat the CIA fact book like a relational table and issue the following query using DB2:

```
Q1: SELECT country, population, GDP,
    gdp _ unit, MilExpendPercent
       FROM cia
       WHERE country IN
   ("Singapore", "Israel", "United States",
"United Kingdom", "Malaysia")
```

(see Table 1)

Since we want to calculate the *military expenditure per capita*, we need to perform the appropriate calculation with a mathematical expression. In addition, we must perform unit conversions (e.g., adjust for the fact that some GDP values are in billions and some in trillions) with the auxiliary database table scalefactor:

TEXT	SCALE
Billion	1000000000
Trillion	1000000000000

This is achieved by joining the non-relational CIA Web source with the relational scalefactor table using the following query:

```
Q2: SELECT country, (MilExpendPercent * GDP
       * scalefactor.scale / population)
AS MilExpPerCapita
   FROM cia, scalefactor
   WHERE scalefactor.text=cia.gdp_unit AND
country IN
("Singapore", "Israel", "United States", "United
Kingdom", "Malaysia")
```

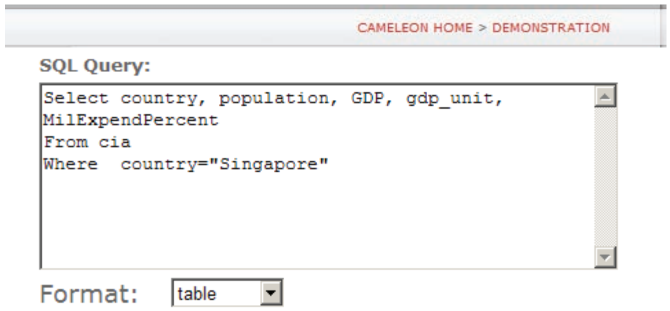
COUNTRY	MilExpPerCapita
Singapore	1379.85
Israel	1901.93
United States	1674.64
United Kingdom	716.72
Malaysia	238.91

Finally, we would like to obtain the *military expenditure per soldier* by creating another NICKNAME for a Wikipedia Web source that has the sizes of armed forces and formulating a federated query joining multiple Web sources, as shown in Figure 4.

Table 1.

COUNTRY	POPULATION	GDP	GDP_UNIT	MILEXPEND-PERCENT
Singapore	4492150	126.5	billion	4.90
Israel	6352117	156.9	billion	7.70
United States	298444215	12.31	trillion	4.06
United Kingdom	60609153	1.81	trillion	2.40
Malaysia	24385858	287	billion	2.03

Figure 3. Simple SQL Query against the wrapped CIA World Fact Book



country	population	gdp	gdp_unit	milexpendpercent
Singapore	4492150	126.5	billion	4.90

Figure 4. Available data in Wikipedia


List of countries by size of arme
From Wikipedia, the free encyclopedia
This list of countries by size of armed forces displays n
and aircrafts. This list is indicative only, as strict comparis
forces might include administative or paramilitary functions
the below figures.

Rank	-	Country	Active troops ('000s)	Reserve troops ('000s)
1.		People's Republic of China **	2255	0800
2.		United States **	1426	0858
3.		India **	1325	1155
4.		North Korea **	1106	4700
5.		Russia **	1037	2400
6.		South Korea	0687	4500
7.		Pakistan **	0619	0528
8.		Iran	0545	0350
9.		Turkey	0514	0378
10.		Vietnam	0484	3000
11.		Egypt	0450	0254

```
CREATE NICKNAME ARMFORCES (
country char(15),
armed_forces integer)
for server Cameleon#_server
options(SERVER_NAME
'http://interchange.mit.edu/Cameleon#_sharp
/camserv.aspx?',
PREDICATES 'country')
```

```
Q3: SELECT cia.country, armed_forces,
      (MilExpendPercent*GDP*scalefactor.uni
      t)/ (armed_forces*1000) AS
      milpersoldier
FROM cia, armforces, scalefactor
WHERE cia.country IN ('Singapore',
'Israel',
```

COUNTRY	ARMED_FORCES	MILPERSOLDIER
Singapore	60	103308.33
Israel	168	71912.50
United States	1426	350481.07
United Kingdom	190	228631.58
Malaysia	110	52964.54

As this simple example shows, querying Web sources using a data federation offers many operational benefits. One can take advantage of the relational database technology in processing semistructured Web data. For example, Web sources can be joined with each other and with other sources, calculations and set operations can be performed, and queries can be optimized. Currently, however, even setting up this motivational example is extremely difficult, if not impossible using one of the data federation engines. The most direct solution offered by DB2 requires coding a custom wrapper for each Web source, but even then those Web sources cannot be joined with each other on the required input attributes (IBM, 2006).

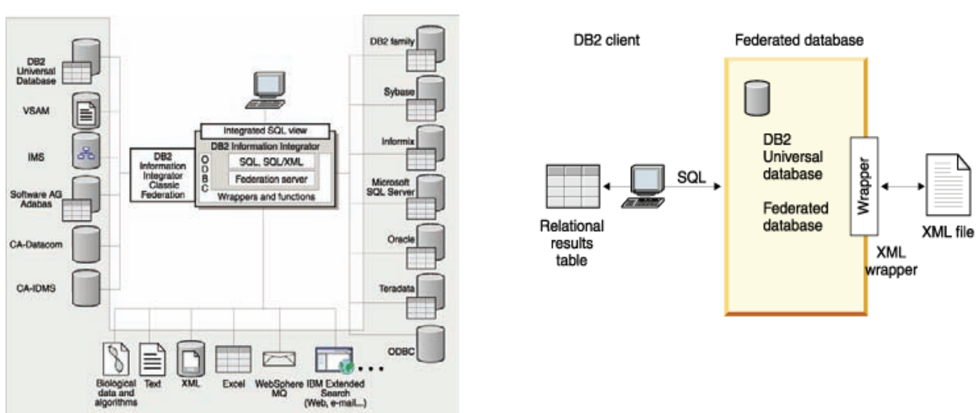
We designed and implemented a new architecture that drastically accelerates the inclusion and querying of Web sources in a data federation. The motivational example, for instance, can be set up in less than an hour without any low-level programming. Users only need to locate and specify the information they want to use on the Web with Cameleon# Studio--a point and click helper tool--and define the Web sources with data definition statements similar to classical "CREATE TABLE" statements. Before explaining the details of our extended architecture, we provide background on the typical operation of data federation systems by using DB2 and its Request-Reply-Compensate protocol as an example.

QUERYING NON-RELATIONAL SOURCES IN A DATA FEDERATION

The goal of a data federation system is to allow clients to access diverse and distributed data sources, regardless of location, format, or access language, from a single interface. While data federation may have a slower access performance compared to data consolidation (as in data warehousing), it has the benefits of (i) *reduced implementation and maintenance costs*, (ii) *access to current data from the source of record*, and (iii) *combining traditional data with mixed format data* (IBM, 2006; Haas, Lin, & Roth, 2002).

As shown in Figure 5, a data federation system uses wrappers to access non-relational data sources such as flat files, XML pages, and Web services. After the user submits a query, the federated server collaborates with the wrapper for each data source to generate an optimized access plan for the query and then evaluates it. Such a plan might call for parts of the query to be processed by the wrappers, by the federated server, or partly by the wrappers and partly by the federated server. The federated server chooses among the plans primarily on the basis of cost.

Figure 5. IBM DB2 data federation architecture [Adopted from (DB2 Information Center 2006)]

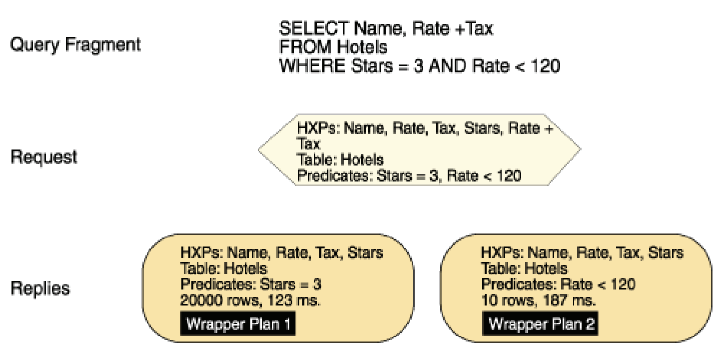


Upon receiving the *request*, the wrapper indicates which sub-pieces of the query fragment it can evaluate, and puts this information in the *reply* to the request. Request properties such as cost, cardinality and ordering properties can also be included. For a typical request, a wrapper could return zero or more reply objects. Each reply represents a different accepted fragment. By the end of query planning, the federated server will weigh all the cost estimations and determine a query execution plan incorporating some set of the accepted fragments offered up by the wrapper in response to requests. During query execution, the federated server will ask the wrapper to execute these query fragments. The federated server can also *compensate* for any query fragments that have not been accepted. Examples of this include a complex predicate or sorting that is beyond the capability of the data source in question. This protocol is therefore called a *request-reply-compensate protocol* in IBM DB2.

Consider Figure 6 as an example. The query fragment (`SELECT Name, Rate + Tax FROM Hotels WHERE Stars=3 AND Rate < 120`) is passed to the wrapper as a request by indicating the head expressions (HXPs), table name, and the predicates. In this case, we assume that the wrapper cannot handle the complete request as it cannot do the `Rate + Tax` calculation and it cannot do two predicates at a time, so replies with two separate parts, which when combined in the federated server answers the original query.

The request-reply-compensate protocol offers a generic framework allowing the federated server to communicate with non-relational data sources through a black box wrapper. Among the built in wrappers that comes with IBM DB2, there are two that are particularly relevant to querying Web sources: XML and Web services wrappers. These wrappers can be used if Web sources can be turned into XML format, or Web services. Neither of these, however, satisfies our desire to include an arbitrary Web source in a data federation and query them without artificial restrictions. The XML wrapper, for instance, does not have the concept of a required input attribute: the XML page should be accessible with a fixed address. Many Web sources are dynamically generated based on input attributes, which precludes the use of XML wrapper as it is. The Web services wrapper, on the other hand, has artificial query restrictions such as "*no IN or OR predicates are allowed for input columns*" (IBM, 2006). For instance, even our simplest query `Q1` cannot be handled by the Web services wrapper assuming our Web source was somehow turned into a Web service.

Figure 6. Request-Reply-Compensate protocol example [Adopted from (DB2 Information Center 2006)]

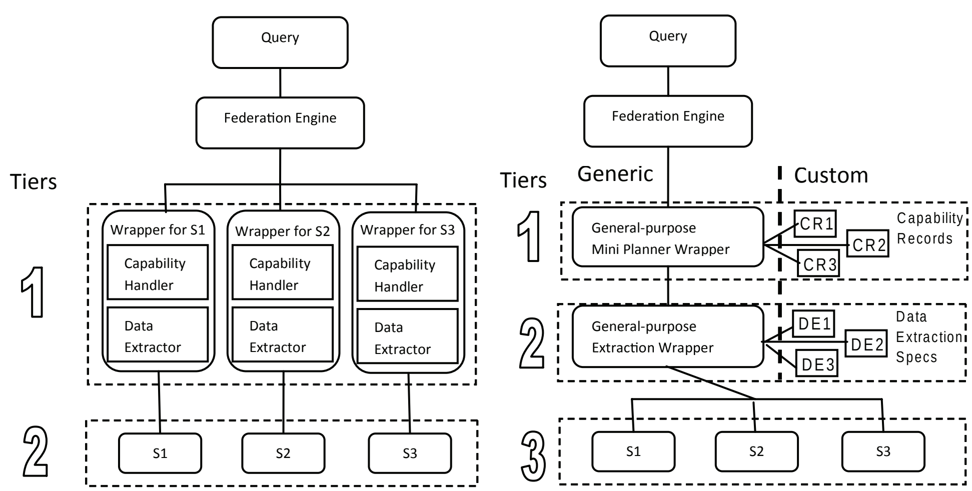


THREE-TIER ARCHITECTURE FOR QUERYING WEB SOURCES IN A DATA FEDERATION

The solution we offer for the inclusion and querying of Web sources in a data federation involves extending the existing two-tier custom wrapper architecture into a three-tier architecture while separating the generic and custom aspects of wrapper development as shown in Figure 7. This new architecture separates code and knowledge, minimizes redundancy, and complements the central query planner when incorporating web sources in a data federation by following the wrapper development protocol specified in DB2 Information Integrator Wrapper Developer's Guide (IBM, 2004).

In the first tier of our solution we have a general-purpose mini planner-wrapper responsible for planning queries involving Web sources. We call it a mini-planner because Web sources have characteristics that limit the query planning space; therefore we do not have to deal with the complexity of a traditional query planner. Our mini planner, in most cases, only needs to handle query planning for a single Web source, leaving complex planning involving multiple sources to the federated server. The mini planner can run Web source queries in parallel, and order them intelligently when they are joined, while respecting their capabilities. Web source capabilities are expressed using simple capability records, which indicate the required input attributes, and whether input attributes can be bound to more than one value at a time. For example, the capability record [b(1), f, f, f, f] for the CIA nickname means that the first attribute *country* needs to be bound with one value at a time (b(1)), and the rest of the attributes must be free (f). In general b(N) indicates that the attribute can be bound with up to N values; "f" indicates that the attribute must be free; and "?" indicates that the attribute can be either bound or free. These capability records are implemented for each source using the nickname definition, right after the predicate keyword. The second-tier is a general-purpose data extraction engine responsible for retrieving data from a Web source and presenting it in the format expected by the data federation engine. For this task, any capable general-purpose data extraction engine

Figure 7. Comparison of Architectures. The extended architecture separates data extraction and capability handling functionalities. Furthermore the primary wrapper is responsible for planning queries posed against web sources with capability restrictions



can be used. We used the data extraction engine, Cameleon#, which uses declarative rules based on regular expressions to extract data from Web pages. Cameleon# Studio can be used to help generate the necessary specification file. An example specification file is shown in Figure 8.

The second-tier extraction wrapper accepts these specification files as input to extract data from any Web source without any procedural coding. Next, we provide the details of the mini query planner for Web sources with capability restrictions.

Figure 8. Example Specification File For the CIA Web Source

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <RELATION name="cia">
  - <SOURCE URI="https://www.cia.gov/cia/publications/factbook/index.html">
    - <ATTRIBUTE name="Link" type="string">
      - <BEGIN>
          <![CDATA[ <body ]]>
        </BEGIN>
      - <PATTERN>
          <![CDATA[ <option value="([^"]*)"[^>]*>$Country$ ]]>
        </PATTERN>
      - <END>
          <![CDATA[ </[Bb][oO][dD][yY]> ]]>
        </END>
      </ATTRIBUTE>
    </SOURCE>
  - <SOURCE URI="https://www.cia.gov/cia/publications/factbook/#Link#">
    - <ATTRIBUTE name="MilExpendPercent" type="string">
      - <BEGIN>
          <![CDATA[ Military expenditures - percent of GDP: ]]>
        </BEGIN>
      - <PATTERN>
          <![CDATA[ <br>\s+([\0-\377]*?)\s+ ]]>
        </PATTERN>
      - <END>
          <![CDATA[ </table> ]]>
        </END>
      </ATTRIBUTE>
    - <ATTRIBUTE name="population" type="string">
      - <BEGIN>
          <![CDATA[ Population: ]]>
        </BEGIN>
      - <PATTERN>
          <![CDATA[ <br>\s*([0-9,]*)\s*[\0-\377]*?\s*< ]]>
        </PATTERN>
      - <END>
          <![CDATA[ </tr> ]]>
        </END>
      </ATTRIBUTE>
    - <ATTRIBUTE name="GDP" type="string">
      - <BEGIN>
          <![CDATA[ purchasing\s*power\s*parity ]]>
        </BEGIN>
      - <PATTERN>
          <![CDATA[ <br>\s+[$]*([,0-9\.]*)[\0-\377]*?\s*</td> ]]>
        </PATTERN>
      - <END>
          <![CDATA[ </tr> ]]>
        </END>
      </ATTRIBUTE>
    - <ATTRIBUTE name="GDP_unit" type="string">
      - <BEGIN>
          <![CDATA[ purchasing\s*power\s*parity ]]>
        </BEGIN>
      - <PATTERN>
          <![CDATA[ <br>\s+[$]*[,0-9\.]*([\0-\377]*?)\s*\( ]]>
        </PATTERN>
      - <END>
          <![CDATA[ </tr> ]]>
        </END>
      </ATTRIBUTE>
    </SOURCE>
  </RELATION>
```

MINI QUERY PLANNER FOR WEB SOURCES

The mini query planner creates a plan that can efficiently retrieve remote data while satisfying query restrictions. Generally, a query planning engine needs to decompose the original query into component subqueries (CSQ), such that each CSQ can be answered using a single data source (Alatovic, 2001; Fynn, 1997). Our mini query planning engine does not need to perform the decomposition since the federated database engine already divides the original query into CSQs, known as requests, where each request can be processed by a single data source. In addition to query decomposition, a query planning engine also needs to maintain the CSQ execution order. Typically, independent CSQs are executed first, followed by dependent CSQs that can be answered using prior results. Thus, detecting the dependencies among the CSQs is crucial to successful planning. Our query planning engine uses both the federated engine and capability records to analyze CSQ dependencies. When the CSQ dependency can be determined using query semantics, our query planning engine uses the federated database engine. When a CSQ does not meet all the capability restrictions of a source, however, the query planning engine will determine if information from other parts of the query can be used to satisfy the capability restrictions. If the restrictions can be satisfied, the CSQ will be modified with the required information so that it can be answered by the native data source.

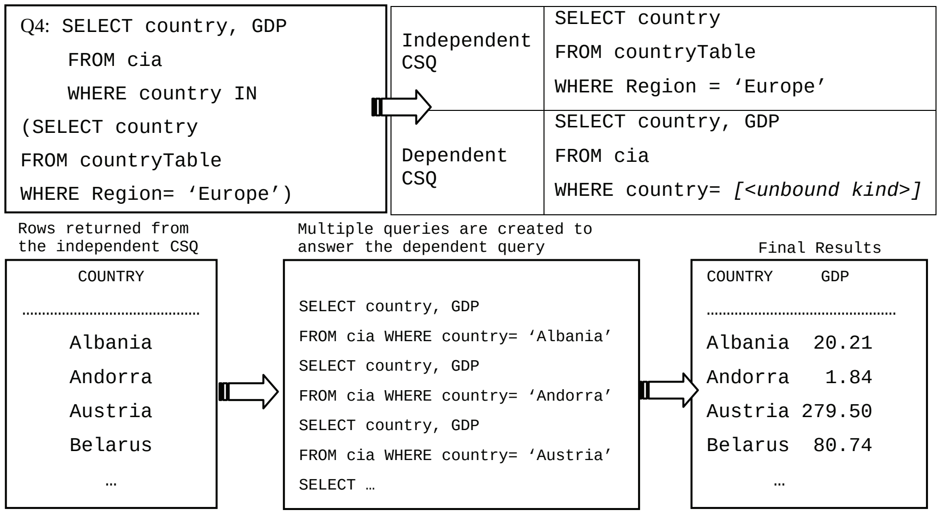
The simplest case for a query execution plan (QEP) is when all CSQs meet the capability restrictions imposed by their native data sources, and they can be executed independently and in parallel. In this case, the federated engine simply decomposes the original query into CSQs and sends them to the native sources through wrappers. After receiving all processed row sets from the native sources, the federated engine aggregates the data and returns the final result.

When a CSQ cannot be executed by itself, however, it is necessary to determine if the CSQ can still be processed using results from other CSQs. Two procedures are used to determine the dependencies: the first method relies on detecting dependencies using query semantics; the second method employs the capability records to meet any unsatisfied restrictions using information from other processed CSQs. The next two sections describe in detail how the two procedures work and how they compensate for each other.

Dependencies Detected via Query Semantics

In Figure 9, we show an example dependency between CSQs that can be detected using the query semantics. In this example, the original query is decomposed into an inner-select CSQ, which can be executed independently, and an outer-select CSQ, which depends on the data returned by the inner-select CSQ. The federated engine facilitates the detection of this dependency by tagging country attribute with a type called *"unbound kind"* to signal to the wrapper that the binding values would be available after the inner-select CSQ is executed. Once the result from the inner-select CSQ is returned, the wrapper needs to create a new set of CSQs by replacing the *"unbound kind"* tag in the original CSQ with the returned value(s). In this example, as illustrated in Figure 9, since country names are returned from the inner-select CSQ; (e.g., Albania, Andorra, Austria, Belarus, etc.), new CSQs are formed after binding each country name to the country attribute. The wrapper then needs to send this new set of CSQs to the native data source. Once the native source processes the CSQs, the wrapper needs to assemble the results and return them to the federation engine. In this example, the wrapper sends the queries to the CIA Web source, retrieves the GDP values and returns them to the federated engine.

Figure 9. An example query dependency that can be detected by query semantics



Dependencies Implied by Capability Restrictions

Some CSQ dependencies may be not be detected via query semantics, but are implied by capability restrictions. Consider for example query Q5, which asks for the GDP and armed-force size of countries that are ranked in the top 10 both in terms of highest GDP and largest armed-force size. Like in the previous example, *countryTable* is a relational source that has the list of countries and their regions.

```
Q5: SELECT cia.country, armed_forces, GDP
FROM countryTable,
(SELECT GDP
 FROM cia
 ORDER BY GDP DESC FETCH FIRST 10 ROWS ONLY) cia,
(SELECT armed_forces
 FROM armforces
 ORDER BY armed_forces DESC FETCH FIRST 10 ROWS ONLY) armforces
WHERE cia.country = countryTable.country AND
armforces.country = countryTable.country
```

To process this query, the query planning engine needs to invoke the countryTable relation to retrieve the list of all countries, and then pass them to the cia and armed_forces relations to obtain the requested data. In order to answer this query, however, the federated engine creates the following two CSQs on Web sources:

```
CSQ1: SELECT GDP
 FROM cia

CSQ2: SELECT armed_forces
 FROM armforces
```

Since none of the CSQs has unbound parameters, the federated engine assumes that they can be executed independently by using the native data sources. Both Web sources, however, require that *country* must be bound before they can return any results. Thus, we cannot produce an answer to the query by only using query semantics. If we consider the capability information, however, it is possible to process both CSQs by finding the missing information from other parts of the query. Using the join conditions “cia.country = countryTable.country” and “armforces.country = countryTable.country” we can rewrite CSQ1 and CSQ2 into CSQ3 and CSQ4 by providing the values for the country attribute from the countryTable relation:

```
CSQ3:SELECT GDP
  FROM cia
  WHERE country IN
  (SELECT country FROM coun-
tryTable)
```

```
CSQ4:SELECT armed_forces
  FROM armforces
  WHERE country IN
  (SELECT country FROM coun-
tryTable)
```

With this added condition, CSQ3 and CSQ4 satisfy the capability restrictions and thus can be processed by the native sources. Although CSQ3 depends on the result from countryTable, this dependency can now be resolved via query semantics with the help of the federation engine as in Figure 10.

The query execution plan (QEP) algorithm, which uses capability records to process CSQs, is presented in Figure 11. The algorithm is based on finding independently executable CSQs in the query and processing them before any dependent CSQs. In most cases, the CSQs that cannot be executed independently lack at least one binding restriction. Once such a CSQ is detected, the algorithm determines if the CSQ can still be executed by searching for the missing binding from other CSQs. If the algorithm finds the missing binding, it is incorporated into the CSQ so that it can be processed by the native source.

Figure 10. An example query dependency implied by capability restrictions

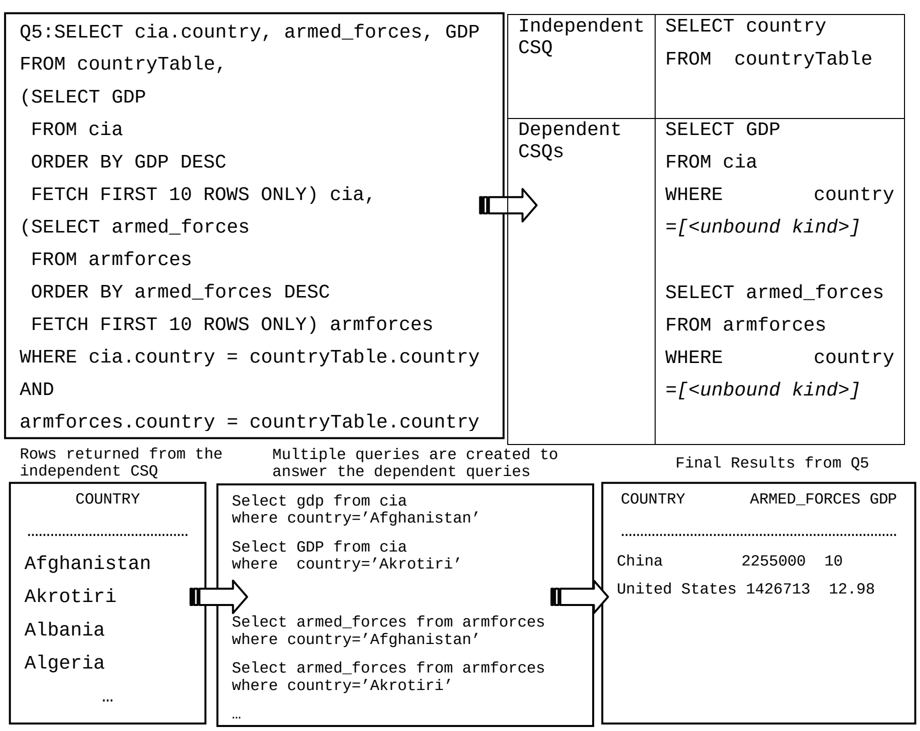


Figure 11. QEP generation algorithm supporting binding query restrictions

```
Input: Single Query q
Output: Query Execution Plan (QEP)

QEP Generation Algorithm:
1.   initialize set S to an empty set
2.   for all CSQs c in S
3.        if c is independently executable
4.        add c to set S
5.        add entry 0:c to QEP
6.   repeat until no more CSQs are added to S
7.        for all CSQs c outside of S
8.             if CSQ c can be executed using bindings from CSQs in S
9.               add an entry for c to QEP including all join bindings of c
10.              add CSQ c to set S
11.   if S does not contain all CSQs in a query
12.        throw exception "query cannot be executed"
13. return QEP
```

There are two non-trivial steps in this algorithm: a) determining if a CSQ can be independently executed (step 3), and b) deciding whether a CSQ can be processed using join bindings from a set of executed CSQs (step 8). The details of these two steps are illustrated in the following sections.

Determining Independently Executable CSQs

Figure 12 shows the algorithm for determining whether a CSQ is independently executable. The algorithm uses the capability restrictions to detect any missing binding in the CSQ, and if they exist, the algorithm determines if these binding conditions can still be satisfied.

Determining Whether a CSQ is Executable Given a set of Executed CSQs

The algorithm for determining whether a CSQ is executable, given a set of CSQs that have already been executed, is depicted in Figure 13. Consider the earlier example in Figure 10 once more. Although the cia and armforces CSQs cannot be executed independently, they can still be processed by finding the missing binding through the use of join conditions in the query. This algorithm detects this class of CSQs that are missing bindings, but can still be executed using information made available through executing other parts of the query. For the specific example of Figure 10, upon finding the attribute country to be unbound, the

Figure 12. Algorithm for determining whether a CSQ is independently executable

```
Independently Executable CSQ:
1. for all binding specifiers bs of c's underlying relation r
2.     for all attribute specifiers as of bs
3.     if as is of type bound and there is no binding in CSQ c
for corresponding attribute
4.                 continue 1
5.            else
6.                 continue 2
7.      end for
8.   return true
9.   end for
```

Figure 13. Algorithm for determining whether a CSQ is executable given a set of executed CSQs

```
Input:    set of executed CSQs S, new CSQ n
Output:  if n cannot be executed given join bindings from CSQs in S
                 returns null
             else
                  returns list of join bindings for CSQ n
CSQ Executable:
1.   for all binding specifiers bs of CSQ n
2.        initialize list of join bindings to an empty list jbl
3.        for all attribute specifiers as of bs
4.             if as is of type bound and CSQ n does not contain binding
                  for attribute matching as
5.                  if there is a join binding jb from n's attribute
                        matching as to one of CSQs in S
6.                  add jb to jbl
7.                  continue 3
8.             else
9.                  continue 1
10.        return jbl
11. return null
```

algorithm discovers a joint binding, countryTable.country=cia.country, that can provide the missing values to the attribute *country*. After modifying the cia CSQ with the new joint binding, the cia CSQ can be executed. Similarly, the binding for armforces CSQ is discovered from the countryTable.country=armforces.country predicate; the CSQ is modified and executed.

Handling Key-at-a-Time Query Restriction

Many Web sources require a single key value to be provided at a time. Consider for example the query Q1 again. (Cia web source has b(1) – one binding at a time – restriction on the attribute country):

```
Q1: SELECT country, population, GDP,
gdp _ unit, MilExpendPercent
    FROM cia
    WHERE country IN
```
("Singapore", "Israel", "United States", "United Kingdom", "Malaysia")

In order to answer this query, the mini-planner needs to change the query into a union of four one-key-at-a-time queries, and perform the union operations locally in parallel. In general, Web sources may have b(N) – N binding at a time – restriction. The short algorithm, shown in Figure 14, handles the general case by recursively rewriting the original query into subqueries. Finally, the algorithm returns the result by performing the union operator on the results of all the subqueries.

Cost Statistics Generation

Cost statistics are especially important for federated queries (Kache, Han, Markl, Raman, & Ewen, 2006). The mini planner wrapper can also return cost statistics for Web sources to the federated engine to aid in query optimization. These cost statistics, as described in DB2 Information Center, are:

1. The cardinality of a nickname. This is defined as the number of rows contained in the nickname (default 1000 rows).
2. The setup cost for a nickname. Setup cost represents the typical time, in milliseconds, that it takes a wrapper to get a query fragment ready to submit to the remote source (default 25 milliseconds).
3. The submission cost for a nickname. Submission cost represents the typical time, in milliseconds, that it takes a wrapper to submit a query fragment to the remote source (default 2000 milliseconds).
4. The advance cost for a nickname. This is the typical time, in milliseconds, that it takes to fetch a single row for the nickname (default 50 milliseconds).

Among these cost statistics, the set up and submission cost can be easily figured out, but the cardinality and the advance cost for a nickname are not easy to calculate for dynamic Web sources. We can, however, estimate the cardinality and advance cost for a nickname by keeping time statistics and cardinality information of previously executed CSQs on the same underlying relation. The estimation process can be initiated by starting with a conservative default time estimate and then improving on it using time statistics on recently executed CSQs on the same underlying relation.

RELATED WORK AND DISCUSSION

Our general strategy for querying Web sources in a data federation system fundamentally differs from other studies (see Florescu, Levy, & Mendelzon, 1998, for a review) in the same area for two reasons:

1. We clearly separate knowledge from code in wrapper development, and improve wrapper development speed and ease of maintenance.
2. We do not assume that we have the liberty to recode the existing federated database systems; thus we focus on improving the process of including and querying Web

sources in cooperation with the existing data federation planners.

The majority of the studies in the area are concerned with query planning under source capability restrictions, and we find two types of approaches in the existing literature: 1) the black-box approach of pushing the capability handling to the wrapper level, and 2) the central planning approach by using a complex declarative language to describe capability restrictions. The IBM DB2 follows the first approach: handling capability restrictions is pushed down to the wrapper layer and it relies on Request-Reply-Compensate protocol to communicate with the wrappers. Although this is a generic framework to incorporate many different sources, coding a different wrapper every time for a Web site with different capability restrictions can be extremely wasteful and error-prone, since most of the code between these wrappers will be common.

There are projects that follow the second approach by describing capability restrictions with a declarative yet complex language. Examples of research projects, which more or less take this route, are Garlic Project at IBM (Roth & Schwarz, 1997; Papakonstantinou, Gupta, & Hass, 1998; Hass, Kossman, Wimmers, & Yang, 1997), TSIMMIS Project at Stanford (Chawathe, Garcia-Molina, Hammer, Ireland, Papakonstantinou, Ullman, & Widom, 1994), Information Manifold (Levy, Rajaraman, & Ordille, 1996), and DISCO (Tomasic, Raschid, & Valduriez, 1998). While this approach is more generic, it has not found its way into existing data federation technologies – perhaps due to its complexity.

The approach we take is a hybrid of these two. As in the black box approach, we push the capability handling to the wrapper level, and like the central planning approach we use declarative capability records. Yet these capability records are designed only to handle Web source access limitations and are not as general as the approaches found in the literature. This restriction simplifies the development of the query planner. Furthermore, our mini query planner creates query plans in cooperation with the central federated query planner, and thus differs from the central planning approach, which does not cooperate with the individual sources.

Another major difference we present is the clear separation of extraction and planning knowledge from the code. This is summarized in Table 2. The wrapper developer only deals with the task of specifying extraction and capability knowledge, and is not involved with low level coding as in other approaches.

Do not Web Services Solve the Problem?

It may be mistakenly thought that the solution offered here would not be needed if the Web sources were Web services returning XML. In fact, we are able to create virtual Web services from any semistructured Web source by using a version of the Cameleon# Web wrapping tool. The capability restrictions, however, are still valid problems for Web services, which often require input attributes before returning any results (Petropoulos, Deutsch, Papakonstantinou, & Katsis, 2007). There is an extra benefit of using

Table 2. Code and knowledge separation in Web wrapper development

	Extraction knowledge	Planning knowledge
Cooperative Planning Approach	Embedded in code	Embedded in code
Central Planning Approach	Embedded in code	Declarative
Our approach	Declarative	Declarative

Web services, as the capability restrictions could be automatically deduced from the Web service description language (WSDL) document instead of declaring them in the nickname statements. All the query dependency issues for arbitrary Web sources, however, equally apply to Web services as well. In fact, the built-in IBM wrapper for Web services prohibits the formulation of queries where dependencies create problems. Our solution is more general and can be used for Web services without artificial restrictions.

CONCLUSION

The Web is undoubtedly the largest and most diverse repository of data; unfortunately it was not designed to offer the capabilities of traditional database management systems. Modern databases promise to include Web sources in a data federation via "wrappers" so that they can be queried as if they are part of a single large database. There are still, however, significant hurdles to fulfilling this promise. With this study we introduced an improved way of dealing with Web source wrappers in federated database applications. With this new general strategy not only do we accelerate the inclusion of Web sources in federated databases, but also we are able to eliminate unnecessary query restrictions. Our contribution is not only at a conceptual level, but also has been implemented using IBM's commercial database engine DB2. Most importantly, all of this has been achieved via extensions allowed by the federation engine, and without requiring any implementation changes in the existing data federation technology.

REFERENCES

Alatovic, T. (2001). *Capabilities aware, planner, optimizer, executioner for Context Interchange project.* Unpublished master's thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

Braumandl, R., Keidl, M., Kemper, A., Kossmann, D., Kreutz, A., Seltzsam, S., & Stocker, K. (2001). ObjectGlobe: Ubiquitous query processing on the Internet. *The VLDB Journal, 10*(1), 48-71.

Chawathe, S.S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J.D., & Widom, J. (1994). The TSIMMIS project: Integration of heterogeneous information sources. In *Proceedings of the 16th Meeting of the Information Processing Society of Japan*, Tokyo, Japan (pp. 7–18).

Firat, A., Madnick, S., Yahaya, N., Kuan., C., & Bressan, S. (2005). *Information aggregation using the Caméléon# Web wrapper* (LNCS 3590, pp. 76-86).

Firat, A., Madnick, S., & Siegel, M. (2000). The Caméléon Web wrapper engine. In *Proceedings of the VLDB2000 Workshop on Technologies for E-Services* (pp. 1-9).

Florescu, D., Levy, A., & Mendelzon, A. (1998). Database techniques for the World Wide Web: A survey. *SIGMOD Record*, *27*(3), 59–74.

Fynn, K. (1997). *A planner/optimizer/executioner for context mediated queries.* Unpublished master's thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

Geer, D. (2003). Federated approach expands database-access technology. *Computer, 36*(5), 18-20.

Haas, L.M., Kossmann, D., Wimmers, E.L., & Yang, J. (1997). Optimizing queries across diverse data sources. In M. Jarke, M.J. Carey, K.R. Dittrich, F.H. Lochovsky, P. Loucopoulos, & M.A. Jeusfeld (Eds.), *Proceedings of the 23rd International Conference on Very Large Data Bases* (pp. 276-285). San Francisco: Morgan Kaufmann Publishers,.

Haas, L.M., Lin, E.T., & Roth, M.A. (2002). Data integration through database federation. *IBM Systems Journal*, *41*(4), 578-596.

IBM. (2006, March 14). *DB2 information center.* Retrieved January, 31, 2007, from http://publib.boulder.ibm.com/infocenter/db2luw/v8//index.jsp

IBM. (2004, September 8). *DB2 information integrator wrapper developer's guide*. Retrieved January, 31, 2007, from http://publibfp.boulder.ibm.com/epubs/pdf/c1891740.pdf

Kache, H., Han, W., Markl, V., Raman, V., & Ewen, S. (2006). POP/FED: Progressive query optimization for federated queries in DB2. In U. Dayal, K. Whang, D. Lomet, G. Alonso, G. Lohman, M. Kersten, S.K. Cha, & Y. Kim (Eds.), *Proceedings of the 32nd International Conference on Very Large Data Bases* (pp. 1175-1178). San Francisco: Morgan Kaufmann Publishers

Kossmann, D. (2000). The state of the art in distributed query processing. *ACM Computing Surveys, 32*(4), 422-469.

Levy, A.Y., Rajaraman, A., & Ordille, J.J. (1996). Querying heterogeneous information sources using source descriptions. In T.M. Vijayaraman, A.P. Buchmann, C. Mohan, & N. L. Sarda (Eds.), *Proceedings of the 22th International Conference on Very Large Data Bases* (pp. 251-262). San Francisco: Morgan Kaufmann Publishers.

Li, C., & Chang, E. (2000). Query planning with limited source capabilities. In *Proceedings of the 16th International Conference on Data Engineering* (p. 401). Washington, DC: ICDE, IEEE Computer Society.

Li, C. (2003). Computing complete answers to queries in the presence of limited access patterns. *The VLDB Journal*, *12*(3), 211–227.

Papakonstantinou, Y., Gupta, A., & Haas, L. (1998). Capabilities-based query rewriting in mediator systems. *Distributed Parallel Databases, 6*(1), 73-110.

Pentaris, F., & Ioannidis, Y. (2006). Query optimization in distributed networks of autonomous database systems. *ACM Transactions on Database Systems, 31*(2), 537-583.

Petropoulos, M., Deutsch, A., Papakonstantinou, Y., & Katsis, Y. (2007). Exporting and interactively querying Web service-accessed sources: The CLIDE system. *ACM Transactions on Database Systems, 32*(4), Article 22.

Roth, M.T., & Schwarz, P.M. (1997). Don't scrap it, wrap it! A wrapper architecture for legacy data sources. In M. Jarke, M.J. Carey, K.R. Dittrich, F.H. Lochovsky, P. Loucopoulos, & M.A. Jeusfeld (Eds.), *Proceedings of the 23rd International Conference on Very Large Data Bases* (pp. 266-275). San Francisco: Morgan Kaufmann Publishers.

Sheth, A.P., & Larson, J.A. (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, *22*(3), 183-236.

Suciu, D. (2002). Distributed query evaluation on semistructured data. *ACM Transactions on Database Systems*, *27*(1), 1-62.

Somani, A., Choy, D., & Kleewein, J.C. (2002). Bringing together content and data management systems: Challenges and opportunities. *IBM System Journal*, *41*(4), 686-696.

Thiran, P., Hainaut, J., Houben, G., & Benslimane, D. (2006). Wrapper-based evolution of legacy information systems. *ACM Transactions on Software Engineering Methodologies*, *15*(4), 329-359.

Tomasic, A., Raschid, L., & Valduriez, P. (1998). Scaling access to heterogeneous data sources with DISCO. *IEEE Transactions on Knowledge and Data Engineering*, *10*(5), 808–823..

Zadorozhny, V., Raschid, L., Vidal, M.E., Urhan, T., & Bright, L. (2002). Efficient evaluation of queries in a mediator for Web sources. In *Proceedings of the 2002 ACM SIGMOD international Conference on Management of Data* (pp. 85-96). New York: ACM.

This work was previously published in Journal of Database Management, Vol. 20, Issue 2, edited by K. Siau, pp. 1-18, copyright 2009 by IGI Publishing (an imprint of IGI Global).

Chapter 8.18
Empirical Studies for Web Effort Estimation

Sergio Di Martino
Università di Salerno, Italy & Università degli Studi di Napoli "Federico II", Italy

Filomena Ferrucci
Università di Salerno, Italy

Carmine Gravino
Università di Salerno, Italy

ABSTRACT

Web technologies are being even more adopted for the development of public and private applications, due to the many intrinsic advantages. Due to this diffusion, estimating the effort required to develop Web applications represents an emerging issue in the field of Web engineering since it can deeply affect the competitiveness of a software company. To this aim, in the last years, several estimation techniques have been proposed. Moreover, many empirical studies have been carried out so far to assess their effectiveness in predicting Web application development effort. In the chapter, we report on and discuss the results of the most significant empirical studies undertaken in this field.

DOI: 10.4018/978-1-60566-040-0.ch017

INTRODUCTION

The availability of powerful server-side, Web-oriented component technologies, such as J2EE, ASP.NET, and so forth, has led to profound changes in the scenario of software systems, allowing developers to create "Web applications," that is, highly-dynamic systems able to deliver a complex amount of functionalities, while running in a Web browser. The approach provides many advantages. In particular, it permits deployment of applications without caring of the client platform, it fully exploits the MVC architecture, and it allows different applications to easily interoperate, by using standard communication protocols and languages, such as XML. As a result, Web applications are becoming an essential support for the every-day activities of both public and private organizations. For instance, to date, most intranet applications, such as document management systems, workflow, and business organization, and B2B solutions are developed with this approach.

On the other hand, the development of these applications has introduced a set of unique features and characteristics, quite different from traditional software construction (Deshpande, 2002; Ginige & Murugesan, 2001). The main issues can be summarized as follows: the requirements are instable, their development is usually characterized by pressure time and compressed schedule; the employed technologies rapidly changes (technology instability), they are usually developed by a small team including young developers, with different backgrounds and knowledge, compared to a traditional software development team. So, a lot of research is needed to provide software engineers with tools and methodologies able to ensure a cost-effective development of this kind of systems. In particular, the traditional approaches for software cost estimation need to be adequately modified to take into account the specific characteristics of these applications. To this aim, currently, many researchers are addressing the crucial problem of estimating the effort required to develop Web applications. Indeed, development effort, meant as the work carried out by software engineers, is the dominant project cost, being also the most difficult to estimate and control, with significant effects on the overall costs. So, effort estimation is a critical activity for planning and monitoring software project development and for delivering the product on time and within budget. Significant over or under-estimates can be very expensive and deleterious for a company. Thus, it is paramount for the competitiveness of a company to be able to effectively predict in advance the effort required to develop a Web-based project (Baresi, Morasca, & Paolini, 2003; Costagliola, Di Martino, Ferrucci, Gravino, Tortora, & Vitiello, 2006a; Mendes, Counsell, & Mosley, 2003b; Reifer, 2000; Ruhe, Jeffery, & Wieczorek, 2003b), and Web effort estimation is an important topic in the field of Web engineering. In this context, special attention is devoted to identifying suitable tools and approaches and to proving by empirical studies that the proposals can be effectively and affordably used in the industrial context.

Goal of the Chapter

The objective of this chapter is to report on the most significant empirical studies undertaken so far and aimed at assessing the effectiveness of measures and techniques for estimating Web application development effort.

BACKGROUND

In the literature, a lot of different methods to estimate software development effort have been proposed. A widely accepted taxonomy of estimation methods classified them in *Non-Model Based* and *Model Methods* (Briand & Wieczorek, 2002).

While *Non-Model Based Methods* mainly take into account expert judgments (thus with highly subjective factors), *Model Based Methods* involve the application of some algorithms to a number of inputs to produce an effort estimation. The inputs for these algorithms are the factors that heavily influence the resulting development effort of a software project. Among these, *Software Size* is accepted as a key cost driver, since it deeply affects total development effort, and thus total project cost (Bohem et al., 2000). Consequently, being able to obtain an early *size measure* for a project can provide a significant estimation of the overall development cost.

In the context of Software Engineering, widely employed *Model Based* estimation methods are Linear Regression (LR), Case-Based Reasoning (CBR), and Regression Tree (RT) (see, for example, Briand, El-Emam, Surmann, Wiekzorek, & Maxwell, 1999b; Briand, El-Emam, & Wiekzorek, 2000; Shepperd & Schofield, 2000).

These approaches use data from past projects, characterized by attributes that are related to effort (e.g., the size), and the actual effort to develop the projects, to estimate effort for a new project under development. In the last years, many researchers and practitioners tried to apply these methods to Web applications. In particular, in the literature there are several works that address the problem of estimating the effort required to develop Web applications (Baresi et al., 2003; Costagliola et al., 2006a, 2006b; Mendes, Counsell, & Mosley, 2002a, 2003a, 2003b; Mendes, Counsell, Mosley, Triggs, & Watson, 2002, 2003; Reifer, 2000; Ruhe, Jeffery, & Wieczorek, 2003a) by identifying measures related to effort and validating them by empirical studies. Indeed, it is largely acknowledged that a software measure and/or estimation method or technique can be acceptable and effectively usable only if it has been validated through several empirical studies proving its usefulness (Basili, Briand, & Melo, 1996; Basili, Shull, & Lanubile, 1999; Briand et al., 1999b; Genero & Piattini, 2001; Mendes, Counsell, & Mosley, 2003b; Myrtveit, Shepperd, & Stensrud, 2005; Schneidewind, 1992).

A crucial aspect in performing empirical studies concerns the availability of a data set of past projects. Indeed, in order to build a model to estimate the effort of a new project, *Model Based Methods* exploit information of past projects on the actual development effort and on the factors that can influence the effort (e.g., the software size). Recently, some researchers have addressed the issue to understand whether a cross-company data set (i.e., a data set containing project data from several software companies) can provide estimates for the new projects comparable to the one obtained from single-company data sets (i.e., data sets containing project data from a single company). Several empirical studies have addressed this issue (see, for example, Briand, El-Eman, Maxwell, Surmann, & Wieczorek, 1999; Jeffery, Ruhe, & Wieczorek, 2000, 2001; Lefley & Shepperd, 2003; Mendes, Lokan, Harrison, & Triggs, 2005; Ruhe & Wieczorek, 2002).

Estimation Techniques

Linear Regression (LR) is a statistical technique that explores the relationship between a dependent variable and one or more independent variables, providing a prediction model described by an equation$y = b_1x_1 + b_2x_2 + ... + b_nx_n + c$where y is the dependent variable, $x_1, x_2, ..., x_n$ are the independent variables, for i=1,..,n, b_i is the coefficient that represents the amount the variable y changes when the variables x changes 1 unit, and c is the intercept (Montgomery, Peck, & Vining, 2001).

In effort estimation, LR is usually exploited to obtain linear regression models that use the variable representing the effort as dependent and the variables denoting the employed size measures (e.g., number of Web pages, number of *Web Objects*, number of *COSMIC-FFP*, etc.) as independent. Once the prediction model has been constructed (i.e., an instance of the previous equation), the effort estimation for a new project can be obtained by substituting in this equation the project size, expressed in terms of the employed measure.

The idea behind the use of the Case-Based Reasoning (CBR) technique is to predict the effort of a new project by considering data on similar projects previously developed. In particular, the completed projects are characterized in terms of a set of *p* features, forming the *case base*. The new project is also characterized in terms of the same *p* attributes and it is referred as the *target case*. Then, the similarity between the target case and the other cases in the *p*-dimensional feature space is measured, and the most similar cases (or projects) are used, possibly with adaptations to obtain a prediction for the target case (Aamodt & Plaza, 1994; Myrtveit & Stensrud, 1999; Shepperd & Schofield, 2000). To apply the CBR technique, a Measurer has to select: the relevant project features, the appropriate similarity function, the number of analogies to select the similar projects to consider for estimation, and the analogy adaptation strategy for generating the estimation. In the effort estimation, usually the size measures have been exploited as the set of features characterizing the projects. It is worthwhile to point out that the selection of the similarity function and the number of analogies is a crucial decision. Many researchers suggested use of Euclidean distance as similarity measure and 1, 2, and 3 analogies to identify similar projects (Briand et al., 2000; Briand, El-Emam, Surmann, et al., 1999; Mendes, Counsell, & Mosley, 2002a; Mendes, Counsell, Mosley, Triggs et al., 2003). The analogy adaptation step allows for deciding how to obtain estimation after the most similar cases have been determined. Widely used adaptation techniques are the *nearest neighbour* (Briand, El-Emam, Surmann, et al., 1999; Mendes, Counsell, & Mosley, 2002a) the *mean of the closest analogies* (Mendes, Counsell, & Mosley, 2002a; Mendes, Counsell, Mosley, Triggs, et al., 2003; Myrtveit & Stensrud, 1999), the *inverse distance weighted mean,* and *inverse rank weighted mean* (Shepperd & Schofield, 2000).

Regression Tree (RT) is a variant of decision trees that can be used to approximate real-valued functions (Briand et al., 2000; Briand, El-Emam, Surmann, et al., 1999b; Mendes, Counsell, Mosley, Triggs et al., 2003). This technique takes as input a set of numerical variables (usually the size measures) and generates a binary tree predicting the value of the target variable (the development effort). In particular, the leaves of the binary tree suggest the values for the target variable on the base of the values of the predicting variables. The binary tree is built by recursively splitting the input data (i.e., the values of the predicting variables) into partitions. At the beginning, all data are associated to the root. Then, they are split in two parts, minimizing the sum of the squared deviations from the mean in the separated parts. At each split, the process determines the input variable to be used for splitting, and its values to associate to the left and right child nodes, respectively. Let us observe that each node has associated the mean value of the target variable. The process ends when, for each node, a minimum size, specified for the node by the user, is obtained. Subsequently, to determine the predicted value for the target variable, we start from the root node and then follow the right or left branch, based on the value of the splitting variable. We continue until a leaf node is reached, which contains the predicted value.

Validation Method and Evaluation Criteria

Validation is a crucial step in the empirical studies in order to verify whether or not the predicted efforts are useful estimations of the actual development efforts. A *cross validation* is widely employed to address it. In particular, this method partitions the data set into two randomly selected sets: the *training set* for model building and the *test set* for model evaluation.

To assess the acceptability of the derived estimations, summary measures, like *MMRE* and *Pred*(0.25) (Conte, Dunsmore, & Shen, 1986), together with *boxplots of (absolute) residuals* (Kitchenham, Pickard, MacDonnell, & Shepperd, 2001; Mendes, Counsell, & Mosley, 2005b). *MMRE* and *Pred*(0.25) are widely employed in the literature to assess the accuracy of effort estimation (see, for example, Briand, El-Emam, Surmann, et al., 1999; Briand, El-Emam, & Wieczorek, 1999, 2000; Briand & Wieczorek, 2002; Costagliola et al., 2006a; Costagliola, Martino, Ferrucci, Gravino, Tortora, & Vitiello, 2006b; Mendes, Counsell, & Mosley, 2003a, 2003b, 2003c; Mendes, Counsell, Mosley, Triggs, et al., 2003; Mendes & Kitchenham, 2004a, 2004b; Myrtveit. & Stensrud, 1999; Ruhe et al., 2003a, 2003b). In the following, we will briefly recall the main concepts underlying *MMRE* and *Pred*(0.25), and boxplots.

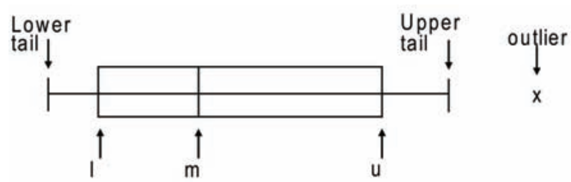
The *Magnitude of Relative Error* (Conte et al., 1986) is defined as $MRE = |EFHreal - EFHpred| / EFHreal$ where *EFHreal* and *EFHpred* are the actual and the predicted efforts, respectively. MRE has to be calculated for each observation in the data set. To have a cumulative measure of the error, all the MRE values are aggregated across all the observations using the mean and the median, thus obtaining two measures of the central tendency, known as Mean of *MRE* (*MMRE*), and Median MRE (*MdMRE*), where the latter is less sensitive to extreme values (Mendes, Counsell, Mosley, Triggs, et al., 2003).

The Prediction at level 0.25 (Conte et al., 1986) is defined as*Pred* (0.25) = k /Nwhere k is the number of observations whose *MRE* is less than or equal to 0.25, and N is the total number of observations. In other words, *Pred*(0.25) is a quantification of the predictions whose error is less than 25%. According to Conte et al. (1986), a good effort prediction model should have a *MMRE*≤0.25 and *Pred*(0.25)≥0.75, meaning that at least 75% of the predicted values should fall within 25% of their actual values.

Boxplots are widely employed in exploratory data analysis since they provide a quick visual representation to summarize the data using five numbers: the median, upper and lower quartiles, minimum and maximum values, and outliers (Kitchenham et al., 2001). The box of the plot is a rectangle with an end at each quartile and a line is drawn across the box at the sample median (*m* in Figure 1). The lower quartile (*l* in Figure 1) is determined considering the bottom half of the data, below the median, for example, by finding the median of this bottom data, while the upper quartile (*u* in Figure 1) is the median of the upper half of the data, above the median. The length of the box *d* is the interquartile range of the statistical sample. Lower tail is *u*+1.5**d* while *u*-1.5**d* is the Upper tail. Points at a distance from the median greater than 1.5 times the interquartile range represent potential outliers and are plotted individually.

In recent studies (see, for example, Costagliola et al., 2006b; Kitchenham et al., 2001; Mendes, Counsell, Mosley, Triggs, et al., 2003; Mendes, Counsell, & Mosley, 2005a, 2005b; Mendes & Kitchenham, 2004b; Ruhe et al., 2003b), the effectiveness of the obtained estimations of effort have been assessed by taking into account first the summary statistics and then the boxplots of absolute residuals, where residuals are calculated as (*EFHreal — EFHpred*).

Figure 1. The boxplot



In the following, we recall the main concepts of LR, CBR, and RT, and the methods widely adopted in the literature to validate the obtained estimations. Then, we will report on and discuss the most important empirical studies dealing with the estimation of Web development effort.

EMPIRICAL STUDIES

In the context of Web effort estimation, many empirical studies have been carried out aiming to address one of the following research issues:

- Verify the usefulness of size measures and cost factors for estimating Web application development effort.
- Verify the appropriateness of methods and techniques to provide effective effort estimations for Web applications.
- Verify the usefulness of cross-company data set for estimating Web application development effort.

In the following, we report on and discuss the empirical studies provided in the literature, taking into account the above research directions.

Empirical Studies Focused on Size Measures

In the last years some measures have been proposed to size Web Applications, such as *Web Objects* defined by Reifer (2000), *COSMIC-FFP* (COSMIC, 2005), the measures defined in the *Tukutuku* database by Mendes, Counsell, and Mosley (2003c), and *Length* and *Functional* measures investigated by Costagliola et al. (2006b)

Web Objects represent an extension of Function Points (FPs) (Albrecht, 1979), which are briefly recalled in the following. The current standard definition and counting procedure of the *FP* approach is reported in the IFPUG Counting Practices Manual (IFPUG, 2004). The measurement of the system size starts with the identification of all the functions, which can be of the following types: *external input*, *external output*, *external inquiry*, *internal logical file* and *external interface file*. The first three classes are considered transaction function types while the last two are considered data function types. Then, these identified functions are weighted in agreement with standard values specified in the Counting Practices Manual, by using their types and the level of their complexity. The *Web Objects* approach extends FPs by introducing four new Web-related components (*multimedia files*, *Web building blocks*, *scripts* and *links*), used as predictors together with the five traditional function types of FPs.

Ruhe et al. (2003a) were the first to investigate *Web Objects* by performing an empirical study based on a data set of 12 Web projects from industrial world, in order to establish whether or not this measure can be used to predict the development effort of Web applications in terms of person-hours, gathering interesting prediction results. In particular, these authors compared *Web Objects* with FPs and the results of the empirical analysis revealed that the model based on *Web Objects* presented significantly better prediction accuracy. In particular, they used LR as estimation technique and obtained an MMRE=0.24 and Pred(0.25)=0.67 for *Web Objects*, while the FPs produced a MMRE=0.33 and Pred(0.25)=0.42. These authors exploited 12 Web projects developed by a small-sized Australian software development company, which were very typical application in Web domain, for example, one for analysing and managing stock or transactions, several Web-based content management systems.

Successively, *Web Objects* measure has been investigated in other two studies (Di Martino, Ferrucci, Gravino, & Mendes, 2007; Ruhe et al., 2003b). In particular, Ruhe et al. (2003b) used *Web Objects* as size measure in the application of *WebCOBRA*, which is an adaptation of the *COBRA* method (Briand, El-Emam, & Bomarius, 1998) for the Web. They obtained better results than with LR, in particular MMRE=0.17, MdMRE=0.15, and Pred(0.25)=0.75. More details on *WebCOBRA* are reported in the next section. The empirical study performed by Di Martino et al. (2007) was aimed at comparing several size measures. By employing LR technique they obtained good results with *Web Objects*, in particular MMRE=0.17, MdMRE=0.11, and Pred(0.25)=0.80.

Common Software Metrics Consortium - Full Function Points (COSMIC-FFP) is a widely adopted method for sizing software, approved as an International Standard (ISO/IEC 19761:2003). It turns out to be particularly suited for real-time and/or multilayered software, whose complexity is mostly dominated by the need to manage large amounts of data (COSMIC, 2005). The basic idea underlying this approach is that, for many kinds of software, the biggest programming efforts are devoted to handle data movements, and thus their number can provide a meaningful sight of the system size. With the COSMIC-FFP method, a set of models, rules, and procedures have to be applied to the Functional User requirements (FUR) to obtain a numerical value, which represents the functional size of the software, expressed in terms of *Cosmic Functional Size Unit* (CFSU) (COSMIC, 2005). To apply the method "the COSMIC-FFP model […] requires FUR to be broken down into functional processes, each consisting of *data movements*, where a data movement moves a data group containing attributes of a single *object of interest*" (COSMIC, 2005).

Mendes, Counsell, and Mosley (2002a) applied the *COSMIC-FFP* measurement to Web hypermedia systems. Using data about 37 Web projects developed by academic students, they constructed an effort prediction model by applying LR. However, the derived model did not present reasonable prediction accuracy (MMRE < 0.50), and replications of the empirical study were highly recommended by authors to find possible biases in the collection of the data and/or in the application of the *COSMIC-FFP* method. Successively, Costagliola et al. (2006a) applied *COSMIC-FFP* taking into account dynamic aspects of Web applications. The *COSMIC-FFP* method was evaluated by using 44 Web projects (mainly Web portals, e-commerce sites, etc.), developed by academic students. The prediction models obtained by applying LR provided a good level of accuracy (MMRE=0.15, MdMRE=0.10, and Pred(0.25)= 0.75).

Tukutuku database is part of the Tukutuku project, which aims to collect data about Web applications to be used to develop Web effort estimation models and to benchmark productivity across and within Web Companies (Mendes, Counsell, & Mosley, 2003c). Each Web project is characterized by 25 variables related to the application and its development process, such as *Web pages*, *Images*, *Features reused without any adaptation*, and so forth. The complete list can be found in Mendes, Counsell, and Mosley (2005a). These size measures have been obtained from several empirical investigations performed by Mendes, Counsell, and Mosley (2001, 2002a, 2003a, 2003b, 2003c) and Mendes, Counsell, Mosley, Triggs et al. (2002, 2003). These studies highlighted that the employed measures can be profitably exploited to predict effort since the best estimations (obtained with LR and CBR) were characterized by MMRE values less than 0.25 and Pred(0.25) values greater than 0.75. In particular, the final set of size measures and cost drivers included in the Tukutuku database have been defined from the results of a survey investigation (Mendes, Counsell, & Mosley, 2005a), using data from 133 online Web forms aimed at giving quotes on Web development projects. In addition, these measures and cost drivers have also been confirmed by an established Web company and a second survey involving 33 Web companies in New Zealand. Several empirical studies have been carried out to assess the effectiveness of these size measures (Mendes & Kitchenham, 2004a; Mendes, Counsell, & Mosley, 2005a; Mendes, Di Martino, Ferrucci, & Gravino, 2007). Mendes Counsell, and Mosley (2005c) also presented a survey literature of hypermedia and Web size measures proposed in recent papers and classifies the analyzed studies according to a proposed taxonomy. The aim was to provide a mean for classifying and understanding the body of knowledge about Web size measures and their employing for effort estimation.

Costagliola et al. (2006b) investigated two sets of size measures: *Length* and *Functional* measures. *Length* measures were derived from both previous research (Mendes, Counsell, & Mosley, 2001, 2002a) and interviews with the company's project managers. The set of *Length* measures includes *Web Pages*, *Media*, *Scripts and Application*, and so forth (the complete list can be found in Costagliola et al., 2006b). As for the *Functional* measures, Costagliola et al. (2006b) used the nine components that are part of *Web Objects*. These size measures were compared by exploiting data from 15 Web projects provided by an Italian small-medium sized company whose core business is the development of enterprise information systems, mainly for local and central government. The company is highly specialized in the design, development and management of solutions for Web portals, enterprise intranet/extranet applications (such as content-ware, e-commerce, work-flow managers, etc.), and Geographical

Information Systems. The 15 Web applications employed in that case study are in several domains, such as e-government, e-banking, Web portals, and intranet applications. The results of the empirical study revealed that *Length* measures provided better estimates when using CBR (i.e., *MMRE*=0.15 and *Pred*(0.25)=0.73), while *Functional* measures provided better estimates when using LR (*MMRE*=0.21 and *Pred*(0.25)=0.73). However, their analysis suggested that there were no significant differences in the estimations and the residuals obtained with *Length* measures and *Functional* measures. Successively, Di Martino et al. (2007) compared the following size measures: *Web Objects* proposed by Reifer (2000), the *Length* and *Functional* measures used by Costagliola et al. (2006b), the *Tukutuku* measures proposed by Mendes, Counsell, and Mosley (2003c). As a data set, they exploited the Web projects employed in Costagliola et al. (2006b). The empirical results showed that all the measures provided good predictions in terms of *MMRE*, *MdMRE*, and *Pred*(0.25) and the study largely confirmed the results of previous work (Costagliola et al., 2006b). Moreover, the analysis of residuals suggested that *Length* measures and *Web Objects* presented significantly superior predictions than *Functional measures* when estimates are obtained using LR; however all presented similar predictions to the *Tukutuku* measures.

Finally, we would recall other two works. Abrahão and Pastor (2003) proposed the *OOmFPWeb* method, which maps the *FP* concepts into the primitives used in the conceptual modeling phase of *OOWS*, a method for producing software for the Web (Abrahão, Fons, & Pastor, 2001). In a recent work (Abrahão, Pastor, & Poels, 2004), a preliminary evaluation of the *OOmFPWeb* has been provided.

Baresi and Morasca (2003) focused their attention on Web applications automatically generated from a design model expressed in terms of W2000. They defined several measures on the basis of attributes obtained from design artifacts, and conducted empirical studies aiming to identify the attributes that may be related to the effort required for designing Web applications. This empirical study involved the students of an advanced university class on modeling Web applications, enrolled in engineering curricula, and employed Linear Regression. The empirical results suggested that the size of the information model, as well as the reuse level and the characteristics of the navigation model, influence the total design effort. It is worth noting that these studies differ from the ones presented above since Baresi and Morasca (2003) focused on design effort and not on the total effort.

Empirical Studies Focused on Estimation Techniques

Several empirical studies have been carried out to analyze the effectiveness of LR, CBR, and RT in the case of desktop applications. In particular, Briand, El-Emam, Surmann, et al. (1999) and Briand, El-Emam, and Wieczorek (2000) applied LR, RT, and CBR, using 1 and 2 analogies, and combinations of these techniques. Their results pointed out that LR is better than CBR in predicting efforts. To date, numerous studies have investigated these estimation techniques for Web effort estimation (e.g., Costagliola et al., 2006a, 2006b; Di Martino et al., 2007; Mendes, Counsell, & Mosley, 2001, 2002a, 2003a, 2003c, 2005b; Mendes, Counsell, Mosley, Triggs et al., 2002, 2003; Mendes & Kitchenham, 2004a, 2004b, Mendes & Mosley, 2002; Ruhe et al., 2003a;).

In the following, we focus our attention on the studies that have employed LR, CBR, and RT (Costagliola et al., 2006b; Di Martino et al., 2007; Mendes, Counsell, Mosley, Triggs et al., 2002, 2003; Mendes & Mosley, 2002;).

Costagliola et al. (2006b) performed an empirical analysis employing LR, CBR, and RT and exploiting *Length* and *Functional* measures (described in the previous section). They found that none of the effort estimation techniques used in their study was statistically significantly superior to others; however, the three accuracy measures used (i.e., MMRE, MdMRE, and Pred(0.25)) suggested that LR presented overall the best and the worst accuracy using elements of *Functional* and *Length* size measures, respectively. The study conducted by Di Martino et al. (2007) largely confirmed the above results, revealing that both LR and CBR can be profitably exploited to predict Web application development effort. Mendes, Counsell, Mosley, Triggs et al. (2002, 2003) exploited a data set of 37 Web projects developed by academic students and highlighted that LR models generally gave statistically significant better results than CBR. On the contrary, the case study carried out by Di Martino et al. (2007) revealed that there is no significant difference between estimations obtained with CBR and those obtained with the LR. As for CBR Mendes, Counsell, Mosley, Triggs et al. (2003) considered three choices for number of analogies (1, 2, and 3) and three choices for the analogy adaptation (mean of k analogies, inverse rank weighted mean, and median of k analogies). The best result was achieved by using weighted distance and 1 analogy. Moreover, in another work on the analysis of the CBR, Mendes and Mosley (2002) proved that the use of adaptation rules presented statistically better prediction accuracy than their counterparts than did not use adaptation. Costagliola et al. (2006b) obtained the best results with CBR using 2 analogies and inverse distance weighted mean as adaptation strategy (employing *Functional* measures), while Di Martino et al. (2007) obtained the best results with CBR using 2 analogies and the mean of 2 analogies as adaptation strategy (with *Web Objects* as size measure) and 2 analogies and the inverse distance weighted mean as adaptation strategy (for *COSMIC-FFP* as size measure).

Morever, we want to recall two works that have proposed cost estimation models specific for the Web, namely *WebMO* (Reifer, 2000) and *WebCOBRA* (Ruhe et al., 2003b).

WebMO is a direct extension of the COCOMO II Early Design model (Bohem, 1981). The idea underlying the approach is to adapt the existing effort estimation methods in order to use them in the context of Web applications. As an extension of the Early Design of the COCOMO II method, *WebMO* does not require a deep knowledge about the influence of cost drivers on the development process, and generic cost drivers can be defined in order to estimate in the early phase of the development process those features which will be more influent on the final development effort (Reifer, 2000). Except the empirical analysis performed by Reifer in order to calibrate the *WebMO*, to the best of our knowledge no empirical studies have been carried out employing this method.

WebCOBRA is an adaptation to the Web domain of the COBRA (COst estimation, Benchmarking and Risk Analysis) method (Briand et al., 1998). The key issue of this method is to use both expert knowledge and few past project data to obtain a COBRA instance, named COBRA model, which results specifically tailored for the intended software development context. In order to obtain a COBRA model, information about Project Characteristics, such as project type, application domain, and so forth, the Size Measure, calculated in a consistent way among all projects, and some Cost Drivers, specifying the resources expected to influence the development effort have to be determined. These three factors form the causal model, which is intended to describe all the cost factors involved in the project development. The causal model, together with the past project data, define a resulting COBRA model. The past project data are used to define the relationships between cost overhead and cost, for the considered software development context. Ruhe et al. (2003b) derived the *WebCOBRA* approach by modifying the causal model, and by adopting the *Web Objects* method to measure the size of the software. Furthermore, to evaluate the prediction accuracy of the *WebCOBRA* model, they performed an empirical study by exploiting the 12

Web projects developed by an Australian software development company and obtained *MMRE*=0.17, *MdMRE*=0.15, and *Pred*(0.25)=0.75. These results are better than those they obtained in Ruhe et al. (2003a) by employing as estimation technique LR (see the previous section). In the future, it could be interesting to exploit both *COSMIC-FFP* and *Web Objects* measures in the application of the *WebCOBRA* method using a larger industrial data set and compare the obtained results.

Empirical Studies Focused on Cross-Company Data Set

Recently, in the context of the Web, few studies have investigated whether a cross-company data set (i.e., a data set containing project data from several software companies) can provide estimates for the new projects comparable to the one obtained from single-company data sets (i.e., data sets containing project data from a single company), which deals with the generalization of the results achieved in a given context (Mendes et al., 2007; Mendes & Kitchenham, 2004a, 2004b).

The use of a cross-company data set seems particularly useful for companies that do not have their own data on past projects from which to obtain their estimates, or that have data on projects developed in different application domains and/or technologies (Mendes et al., 2007; Mendes & Kitchenham, 2004a, 2004b). The goal of these studies is to investigate how successful is a cross-company effort model to estimate effort for Web projects that belong to a single company and were not used to build the cross-company model, and compared to a single-company effort model. The data sets employed in these three studies were obtained from the Tukutuku database, and the effort estimates were obtained by means of LR and CBR. About the results, these empirical studies have revealed that the predictions based on the single-company model were significantly more accurate than those based on the cross-company ones. In particular, in the study presented by Mendes et al. (2007), the accuracy of estimates obtained using the cross-company model obtained with LR shows low prediction accuracy, with a quite high MMRE (85.86%), and an extremely poor Pred(25) (6.67%), far from the thresholds suggested by Conte et al. (1986). The same pattern was present for predictions obtained using CBR: MMRE is 92.54% and Pred(25) is 0%, both really poor. These results corroborate those obtained by Mendes and Kitchenham (2004b).

It is worth noting that the majority of the empirical studies provided in the literature, and reported in the previous sections, exploited data coming from a single company. On one hand, this can lead to more accurate results, since the data collection task can be carried out in a more controlled fashion than in a cross-company scenario, as recognized by Kitchenham, Mendes, Travassos, and Guilherme (2007). But, on the other hand, the obtained results might not hold in other industrial contexts. In particular, the models obtained with data coming from a given company/domain might not supply accurate predictions when used in different companies/domains, where some recalibrations could be necessary as suggested by Mendes and Kitchenham (2004b).

CONCLUSION

In this chapter, we have reported on and discussed the most significant empirical studies undertaken so far and aimed at verifying:

- The usefulness of size measures and cost factors for estimating Web application development effort.
- The appropriateness of methods and techniques to provide effective effort estimations for Web applications.
- The usefulness of cross-company data set for estimating Web application development effort.

The analysis has revealed that several size measures have been successfully employed to estimate the size and the effort of Web applications. However, further empirical investigations are needed to achieve widely accepted size measures for the Web. Regarding the employed estimation techniques, LR and CBR, widely adopted in Software Engineering to estimate effort of software projects, have also obtained interesting results in the context of Web.

FUTURE RESEARCH DIRECTIONS

It is worth noting that only few empirical studies in the field of Web effort estimation exploited industrial data sets (Costagliola et al., 2006b; Di Martino et al., 2007; Mendes, Counsell, & Mosley, 2005a, 2005b, 2005c; Mendes & Kitchenham, 2004a, 2004b; Ruhe et al., 2003a, 2003b). This is partly due to the lack of "meaningful" data sets publicly available that allow researchers to perform investigations. It is worth noting that the "quality" of the data set used can deeply influence the results of the study. Indeed, if data is not properly collected, or if it is not representative of a wider context, the effectiveness of the empirical study is heavily reduced. In the field of software engineering, these recommendations are usually fitted by using data coming from the industrial world, collected in a correct and consistent way (see, for example, Briand et al., 2000; Briand, El-Emam, & Surmann, et al., 1999; Greves & Schreiber, 1996; Jeffery et al., 2000; Mendes, Lokan et al., 2005). However, in the Web domain, it may be difficult to obtain large and significant data sets from the industrial world, since the time required to accumulate enough data on past projects may be very high, as well as the fact that technologies evolve very quickly, making old data useless (Mendes & Kitchenham, 2004b). Thus, in the future, Web companies should pay more attention to cope with the problem of collecting data from past projects in order to exploit them for effort estimation in several and different contexts.

Moreover, the research in the future should also focus on designing and developing tools able to automatically obtain information on cost factors and size measures. In particular, these tools should allow researchers to collect past projects data from companies developing Web applications employing different methods, techniques, and technologies with the aim of supporting managers in the identification of the most effective measures and estimation techniques to be applied in several and different contexts. Other techniques should also be investigated for effort estimation, such as fuzzy strategies, and the achieved results should be compared with those obtained by employing LR and CBR. Finally, future work should also cover the identification and investigation of size measures to be used not only in the early phase of development process but also in other phases, such as testing and maintenance.

REFERENCES

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. [IOS Press.]. *AI Communications*, *7*(1), 39–59.

Abrahão, S. M., Fons, J. J., & Pastor, O. (2001). Object-oriented approach to automate Web applications development. In *Proceedings of International Conference on Electronic Commerce and Web Technologies* (EC-Web '01) (pp. 16-28). Springer Verlag.

Abrahão, S. M., & Pastor, O. (2003). Measuring the functional size of Web applications. *International Journal of Web Engineering and Technology*, *1*(1), 5–16. doi:10.1504/IJWET.2003.003265

Abrahão, S. M., Pastor, O., & Poels, G. (2004). Evaluating a functional size measurement method for Web applications: An empirical analysis. In *Proceedings of International Software Metrics Symposium* (METRICS'04) (pp. 358-369).

Albrecht, A. J. (1979). Measuring application development productivity. In *Proceedings of the Joint SHARE/GUIDE/IBM Application Development Symposium* (pp. 83-92).

Baresi, L., & Morasca, S. (in press). Three empirical studies on estimating the design effort of Web applications. *ACM Transactions on Software Engineering and Methodology*.

Baresi, L., Morasca, S., & Paolini, P. (2003). Estimating the design effort of Web applications. In *Proceedings of the International Software Metrics Symposium* (METRICS'03) (pp. 62-72).

Basili, V. R., Briand, L. C., & Melo, W. L. (1996). A validation of object-oriented design metrics as quality indicators. *IEEE Transactions on Software Engineering*, *22*(10), 751–761. doi:10.1109/32.544352

Basili, V. R., Shull, F., & Lanubile, F. (1999). Building knowledge through families of experiments. *IEEE Transactions on Software Engineering*, *25*(4), 456–473. doi:10.1109/32.799939

Bohem, B. (1981). *Software engineering economics*. Englewood Cliffs, NJ: Prentice-Hall.

Bohem, B. W., Abts, C., Brown, A. W., Chulani, S., Clark, B. K., Horowitz, W., et al. (2000). *Software cost estimation with COCOMO II*. Englewood Cliffs, NJ: Prentice Hall.

Briand, L. El. Emam, K., Surmann, D., Wiekzorek, I., & Maxwell, K. (1999). An assessment and comparison of common software cost estimation modeling techniques. In *Proceedings of International Conference on Software Engineering, (ICSE'99),* (pp. 313-322).

Briand, L. El. Emam, K., & Wiekzorek, I. (2000). A replicated assessment and comparison of common software cost modeling techniques. In *Proceedings of the International Conference on Software Engineering, (ICSE'00),* (pp. 377-386).

Briand, L., El Emam, K., & Bomarius, F. (1998). COBRA: A hybrid method for software cost estimation, benchmarking, and risk assessment. In *Proceedings of International Conference on Software Engineering (ICSE'98),* (pp. 390-399).

Briand, L., El Emam, K., & Wieczorek, I. (1999). Explaining the cost of European space and military projects. In *Proceedings of International Conference on Software Engineering (ICSE'99),* (pp. 303-312).

Briand, L., & Wieczorek, I. (2002). Software resource estimation. In J.J. Marciniak (Ed.), *Encyclopedia of software engineering* (Vol. 2, pp. 1160-1196) (2nd ed.). New York: John Wiley & Sons.

Briand, L. C., El-Emam, K., Maxwell, K., Surmann, D., & Wieczorek, I. (1999). An assessment and comparison of common cost estimation models. In *Proceedings of the International Conference on Software Engineering (ICSE'99),* (pp. 313-322).

Conte, D., Dunsmore, H. E., & Shen, V. Y. (1986). *Software engineering metrics and models*. The Benjamin/Cummings Publishing Company, Inc.

COSMIC. (2005, December). *Guideline for sizing business application software using COSMIC-FFP, Version 1.0.* Retrieved June 13, 2008, from http://www.lrgl.uqam.ca/cosmic-ffp/manual.html

Costagliola, G., Di Martino, S., Ferrucci, F., Gravino, C., Tortora, G., & Vitiello, G. (2006a). A COSMIC-FFP: Approach to predict Web application development effort. *Journal of Web Engineering*, *5*(2), 93–120.

Costagliola, G., Di Martino, S., Ferrucci, F., Gravino, C., Tortora, G., & Vitiello, G. (2006b). Effort estimation modeling techniques: A case study for Web applications. In *ACM Proceedings of the 6th International Conference on Web Engineering* (ICWE 2006), (pp. 9-16).

Deshpande, Y. (2002). Consolidating Web engineering as a discipline. *Software Engineering Australia*, 31-34.

Di Martino, S., Ferrucci, F., Gravino, C., & Mendes, E. (2007). Comparing size measures for predicting Web application development effort: A case study. In *Proceedings of Empirical Software Engineering and Measurement* (ESEM'07).

Genero, M., & Piattini, M. (2001). Empirical validation of measures for class diagram structural complexity through controlled experiments. In *Proceedings of International ECOOP Workshop on Quantitative Approaches in Object-Oriented Software Engineering* (QAOOSE 2001).

Ginige, A., & Murugesan, S. (2001). Web engineering: An introduction. *IEEE MultiMedia*, *8*(2), 14–18. doi:10.1109/93.923949

Greves, D., & Schreiber, B. (1996). The ESA initiative for software productivity benchmarking and effort estimation. ESA Bulletin, 87. Retrieved June 13, 2008, from <http://esapub.esrin.esa.it/bulletin/bulett87/greves87.htm>

International Function Point Users Group (IFPUG). (2004). *Function point counting practices manual*. Release 4.2.1, 2004.

Jeffery, R. Ruhe. M., &Wieczorek, I., (2001). Using public domain metrics to estimate software development effort. In *Proceedings of International Software Metrics Symposium (METRICS'01),* (pp. 16-27).

Jeffery, R., Ruhe, M., & Wieczorek, I. (2000). A comparative study of two software development cost modeling techniques using multi-organizational and company-specific data. *Information and Software Technology*, *42*, 1009–1016. doi:10.1016/S0950-5849(00)00153-1

Kitchenham, B., Mendes, E., Travassos, G. H., & Guilherme, H. (2007). Cross versus within-company cost estimation studies: A systematic review. *IEEE Transactions on Software Engineering*, *33*(5), 316–329. doi:10.1109/TSE.2007.1001

Kitchenham, B., Pickard, L. M., MacDonell, S. G., & Shepperd, M. J. (2001). What accuracy statistics really measure. *IEE Proceedings. Software*, *148*(3), 81–85. doi:10.1049/ip-sen:20010506

Lefley, M., & Shepperd, M. J. (2003). Using genetic programming to improve software effort estimation based on general data sets. In *Proceedings of GECCO'03*, (LNCS 2724, pp. 2477-2487). Springer-Verlag.

Mendes, E., Counsell, S., & Mosley, N. (2001). Web metrics – estimating design and authoring effort [Special issue on Web engineering]. *IEEE MultiMedia*, 50–57. doi:10.1109/93.923953

Mendes, E., Counsell, S., & Mosley, N. (2002a). Comparison of Web size measures for predicting Web design and authoring effort. *IEE Proceedings. Software*, *149*(3), 86–92. doi:10.1049/ip-sen:20020337

Mendes, E., Counsell, S., & Mosley, N. (2003a). Early Web size measures and effort prediction for Web costimation. In *Proceedings of International Software Metrics Symposium (METRICS'03),* (pp. 18-39).

Mendes, E., Counsell, S., & Mosley, N. (2003b). A replicated assessment of the use of adaptation rules to improve Web cost estimation. In *Proceedings of International Symposium on Empirical Software Engineering (ISESE'03),* (pp. 100-109).

Mendes, E., Counsell, S., & Mosley, N. (2003c). Investigating early Web size measures for Web cost estimation. In *Proceedings of (EASE '03),* (pp. 1-22).

Mendes, E., Counsell, S., & Mosley, N. (2005a). Investigating Web size metrics for early Web cost estimation. *Journal of Systems and Software*, *77*(2), 157–172. doi:10.1016/j.jss.2004.08.034

Mendes, E., Counsell, S., & Mosley, N. (2005b). Exploring case-based reasoning for Web hypermedia project cost estimation. *International Journal of Web Engineering and Technology*, *2*(1), 117–143. doi:10.1504/IJWET.2005.007467

Mendes, E., Counsell, S., & Mosley, N. (2005c). Towards a taxonomy of hypermedia and Web application size metrics. In *Proceedings of the International Conference on Web Engineering (ICWE'05)*, (LNCS 3579, pp. 110-123). Springer.

Mendes, E., Counsell, S., Mosley, N., Triggs, C., & Watson, I. (2002). A comparison of development effort estimation techniques for Web hypermedia applications. In *Proceedings of International Software Metrics Symposium (METRICS'02)*, (pp. 131-140).

Mendes, E., Counsell, S., Mosley, N., Triggs, C., & Watson, I. (2003). A comparative study of cost estimation models for Web hypermedia applications. *Empirical Software Engineering*, *8*(2), 163–196. doi:10.1023/A:1023062629183

Mendes, E., Di Martino, S., Ferrucci, F., & Gravino, C. (2007). Effort estimation: How valuable is it for a Web company to use a cross-company data set, compared to using its own single-company data set?" In *ACM Proceedings of the International World Wide Web Conference (WWW2007)*, (pp. 963-972).

Mendes, E., & Kitchenham, B. (2004a). A comparison of cross-company and single-company effort estimation models for Web applications. In *Proceedings EASE 2004*, (pp. 47-55).

Mendes, E., & Kitchenham, B. (2004b). Further comparison of cross-company and within-company effort estimation models for Web applications. In *Proceedings of International Software Metrics Symposium (METRICS'04)*, (pp. 348-357).

Mendes, E., Lokan, C., Harrison, R., & Triggs, C. (2005). A replicated comparison of cross-company and within-company effort estimation models using the ISBSG database. In *Proceedings of International Software Metrics Symposium (METRICS'05)*, (p. 36).

Mendes, E., & Mosley, N. (2002). Further investigation into the use of CBR and stepwise regression to predict development effort for Web hypermedia applications. In *Proceedings of International Symposium on Empirical Software Engineering (IESE'02)*, (pp. 79-90).

Montgomery, D., Peck, E., & Vining, G. (2001). Introduction to linear regression analysis (3rd ed.). John Wiley & Sons, Inc.

Myrtveit, I., Shepperd, M., & Stensrud, E. (2005). Reliability and validity in comparative studies of software prediction models. *IEEE Transactions on Software Engineering*, *31*(5), 380–391. doi:10.1109/TSE.2005.58

Myrtveit, I., & Stensrud, E. (1999). A controlled experiment to assess the benefits of estimating with analogy and regression models. *IEEE Transactions on Software Engineering*, *25*(4), 510–525. doi:10.1109/32.799947

Reifer, D. (2000). Web-development: Estimating quick-time-to-market software. *IEEE Software*, *17*(8), 57–64. doi:10.1109/52.895169

Ruhe, M., Jeffery, R., & Wieczorek, I. (2003a). Using Web objects for estimating software development effort for Web applications. In *Proceedings of the IEEE International Software Metrics Symposium (METRICS'03)*, (pp. 30-37).

Ruhe, M., Jeffery, R., & Wieczorek, I. (2003b). Cost estimation for Web applications. In *Proceedings of International Conference on Software Engineering (ICSE'03),* (pp. 285-294).

Ruhe, M., & Wieczorek, I. (2002). How valuable is company-specific data compared to multi-company data for software cost estimation? In *Proceedings of the IEEE Symposium on Software Metrics (METRICS'02),* (pp. 237-246).

Schneidewind, N. (1992). Methodology for validating software metrics. *IEEE Transactions on Software Engineering*, *18*(5), 410–422. doi:10.1109/32.135774

Shepperd, M., & Schofield, C. (2000). Estimating software project effort using analogies. *IEEE Transactions on Software Engineering*, *23*(11), 736–743. doi:10.1109/32.637387

ADDITIONAL READING

Albrecht, A. J. (1979). Measuring application development productivity. In *Proceedings of the Joint SHARE/GUIDE/IBM Application Development Symposium* (pp. 83-92).

Boehm-Davis, D. A., & Ross, L. S. (1992). Program design methodologies and the software development process. *International Journal of Man-Machine Studies*, *36*(1), 1–19. doi:10.1016/0020-7373(92)90050-U

Carver, J., Jaccheri, L., Morasca, S., & Shull, F. (2003). Issues in using students in empirical studies in software engineering education. In *Proceedings of International Software Metrics Symposium (METRICS'03),* (pp. 239-249).

Costagliola, G., Ferrucci, F., Gravino, C., Tortora, G., & Vitiello, G. (2004). A COSMIC-FFP based method to estimate Web application development effort. In *Proceedings of International Conference on Web Engineering (ICWE'04),* (LNCS 3140, pp. 161-165). Springer Verlag.

Costagliola, G., Ferrucci, F., Tortora, G., & Vitiello, G. (2005). Class point: An approach for the size estimation of object-oriented systems. *IEEE Transactions on Software Engineering*, *31*(1), 52–74. doi:10.1109/TSE.2005.5

Dawson, R., Bones, P., Oates, B. J., Brereton, P., Azuma, M., & Jackson, M. L. (2003). Empirical methodologies in software engineering. In *Annual International Workshop on Software Technology and Engineering Practice (STEP'03),* (pp. 52-58).

Dolado, J. J. (2000). A validation of component-based method for software size estimation. *IEEE Transactions on Software Engineering*, *26*(10), 61–72. doi:10.1109/32.879821

El Emam, K. (2001). A primer on object-oriented measurement. In *Proceedings of IEEE International Software Metrics Symposium (METRICS'01),* (pp. 185-188).

Finnie, G. R., Wittig, G. E., & Desharnais, J. M. (1997). A comparison of software effort estimation techniques: Using function points with neural networks, case-based reasoning and regression models. *Journal of Systems and Software*, *39*, 281–289. doi:10.1016/S0164-1212(97)00055-1

Jørgensen, M., & Sjøberg, D. I. K. (2001). Impact of effort estimation on software project work. *Information and Software Technology*, *43*, 939–948. doi:10.1016/S0950-5849(01)00203-8

Kadoba, G., Cartwright, M., Chen, L., & Shepperd, M. (2000). Experiences using case-based reasoning to predict software project effort. In *Proceedings of Evaluation and Assessment in Software Engineering (EASE'00)*.

Kadoba, G., & Shepperd, M. (2001). Using simulation to evaluate predictions systems. In *Proceedings of International Software Metrics Symposium (METRICS'01)*, (pp. 349-358).

Kemerer, C. F. (1987). An empirical validation of software cost estimation models. *Communications of the ACM*, *30*(5). doi:10.1145/22899.22906

Kitchenham, B. (1998). A procedure for analyzing unbalanced datasets. *IEEE Transactions on Software Engineering*, *24*(4), 278–301. doi:10.1109/32.677185

Kitchenham, B., Foss, T., Stensrud, E., & Myrtveit, I. (2003). A simulation study of the model evaluation criterion MMRE. *IEEE Transactions on Software Engineering*, *29*(11), 985–995. doi:10.1109/TSE.2003.1245300

Kitchenham, B., Pfleeger, S. L., Pickard, L. M., Jones, P. W., Hoaglin, D. C., & El Emam, K. (2002). Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering*, *28*(8), 721–734. doi:10.1109/TSE.2002.1027796

Kitchenham, B., & Taylor, N.R. (1984). Software cost models. *ICL Technical Journal*, 73-102.

Lokan, C., & Mendes, E. (2006). Cross-company and single-company effort models using the ISBSG database: A further replicated study. In *Proceedings of International Symposium on Empirical Software Engineering (ISESE'06)*, (pp. 75-84).

Maxwell, K. (2002). *Applied statistics for software managers*. Prentice Hall, NJ: Software Quality Institute Series.

Maxwell, K., Wassenhove, L. V., & Dutta, S. (1999). Performance evaluation of general and company specific models in software development effort estimation. *Management Science*, *45*(6), 787–803. doi:10.1287/mnsc.45.6.787

Mendes, E., & Mosley, N. (2006). *Web engineering 2006, XIX 438* (p. 143) (ISBN: 978-3-540-28196-2).

Morisio, M., Stamelos, I., Spahos, V., & Romano, D. (1999). Measuring functionality and productivity in Web-based applications: A case study. In *Proceedings of the International Software Metrics Symposium* (pp. 111-118).

Musilek, P., Pedrycz, W., Sun, N., & Succi, G. (2002). On the sensitivity of COCOMO II software cost estimation model. In *Proceedings of International Software Metrics Symposium (METRICS'02)*, (pp. 13-20).

Pfleeger, S. (1991). Model of software effort and productivity. *Information and Software Technology*, *33*(3), 224–231. doi:10.1016/0950-5849(91)90137-Z

Putnam, L. H. (1978). A general empirical solution to the macro software sizing and estimating problem. *IEEE Transactions on Software Engineering*, *4*, 345–361. doi:10.1109/TSE.1978.231521

Shepperd, M., & Cartwright, M. (2001). Predicting with sparse data. *IEEE Transactions on Software Engineering*, *27*(11), 987–998. doi:10.1109/32.965339

Shepperd, M., Schofield, C., & Kitchenham, B. (1996). Effort estimation using analogy. In *Proceedings of International Conference on Software Engineering (ICSE'96)*, (pp. 170-178).

This work was previously published in Information Systems Research Methods, Epistemology, and Applications, edited by A. Cater-Steel & L. Al-Hakim, pp. 311-326, copyright 2009 by Information Science Reference (an imprint of IGI Global).

Index

D

I

J

K

L

M

N

O

P

T

U

V

W

X

Y